# INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

**The quality of this reproduction is dependent upon the quality of the copy submitted.** Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA

UMI®
800-521-0600

# Theoretical and Empirical Considerations in Investigating Washback: A Study of ESL/EFL Learners

by
Shahrzad Saif

B.A., Allameh Tabatabai University, 1984
M.A., Shiraz University, 1987

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Linguistics

We accept this dissertation as conforming
to the required standard

Dr. John H. Esling, Supervisor (Department of Linguistics)

Dr. Joseph F. Kess, Departmental Member (Department of Linguistics)

Dr. Barbara P. Harris, Departmental Member (Department of Linguistics)

Dr. John O. Anderson, Outside Member (Department of Educational Psychology and Leadership Studies)

Dr. Jared Bernstein, External Examiner (Department of Linguistics, Stanford University)

Supervisor: Dr. John H. Esling

# ABSTRACT

Researchers' and educators' recognition of the positive/negative effects of tests on teaching and learning activities goes back at least four decades. However, this phenomenon, referred to as "washback" in the applied linguistic literature, has been examined empirically by only a few studies in the field of language testing. Even fewer have based their investigation into washback on an a priori theory outlining the scope and design of the study.

This study examines washback as a phenomenon relating those factors that directly affect the test to those areas most likely to be affected by the test. The goals of the study are: (i) to investigate the existence and nature of the washback phenomenon, (ii) to identify the areas directly/indirectly affected by washback, and (iii) to examine the role of test context, construct, task, and status in promoting beneficial washback.

Theoretically, this study conceptualizes washback based on the current theory of validity proposed by Messick (1989, 1996). It is defined as a phenomenon related to the consequential aspect of the test's construct validity and thus achievable, to a large extent, through the test's design and administration. Given this assumption, a conceptual and methodological framework is proposed that identifies "needs", "means", and "consequences" as the major focus areas in the study of washback. While the model recognizes tests of language abilities as instrumental in bringing about washback effects, it highlights an analysis of the needs and objectives of the learners (and of the educational system) and their relationship with the areas influenced by washback as the starting point

for any study of washback. Areas most likely to be affected by the test, as well as major variables that can potentially promote or hinder the occurrence of washback, are also delineated by the model.

This theoretical framework is examined empirically in this study through a long-term multi-phase investigation conducted in different educational contexts (EFL/ESL), at different levels of proficiency (advanced/intermediate), with different tasks (oral/written) and different groups of subjects. The stages in the experimental part of the study correspond to the different phases of the theoretical framework underlying the investigation. The approach to data collection is both quantitative and qualitative.

The results of the study indicate that positive washback can in fact occur if test constructs and tasks are informed by the needs of both the learners and the educational context for which they are intended. The extent, directness, and depth of washback, however, are found to vary in different areas likely to be influenced by washback. The areas most influenced by washback are found to be those related to immediate classroom contexts: (i) teachers' choice of materials, (ii) teaching activities, (iii) learners' strategies, and (iv) learning outcomes. The study also reveals that non-test-related forces and factors operative in a given educational system might prevent or delay beneficial washback from happening. Based on the theoretical assumption underlying the definition of washback adopted in this study, such consequences which cannot be traced back to the construct of the test are outside the limits of a washback study.

Examiners:

Dr. John H. Esling, Supervisor (Department of Linguistics)

Dr. Joseph F. Kess, Departmental Member (Department of Linguistics)

Dr. Barbara P. Harris, Departmental Member (Department of Linguistics)

Dr. John O. Anderson, Outside Member (Department of Educational Psychology and Leadership Studies)

Dr. Jared Bernstein, External Examiner (Department of Linguistics, Stanford University)

# TABLE OF CONTENTS

# TABLES

# FIGURES

# ACKNOWLEDGEMENTS

I would like to extend my deepest gratitude and appreciation to all those individuals without whose help and support this work would not exist. Many thanks to my supervisor Dr. John Esling who patiently guided me through my Ph.D. years. He has provided me with valuable materials, ideas, comments and moral support for which I am most grateful. I am also deeply indebted to Dr. Gordana Lazarevich, Dean of Graduate Studies, without whose financial and administrative support the experimental part of the project would never have taken place. She supported this study in every way possible and stood by it all the way to its completion.

I am grateful to the members of my supervisory committee: Dr. Kess for his prompt and careful reading of the manuscript and his very useful comments, Dr. Harris and Dr. Anderson for their help and insights, and Dr. Bernstein from Stanford University for agreeing to be my external examiner and going to the trouble of attending my defense in person. I would also like to express my deepest appreciation to all faculty and staff at the Department of Linguistics from whose classes, support and friendship I benefited immensely: Dr. Czaykowska-Higgins, Dr. Saxon, Dr. Hukari, Dr. Carlson, Dr. Hess, Dr. Warbey, Dr. Miyamoto, Dr. Lin, Dr. Nylvek, Darlene Wallace, Gretchen Moyer and Jocelyn Clayards. Special thanks to Drs. Czaykowska-Higgins and Saxon for kindly supervising my candidacy papers and to the graduate secretary, Ms. Wallace, for bringing to my attention deadlines, rules and regulations I would have otherwise missed.

Thank you also to the teaching and administrative staff of the English Language Centre at the University of Victoria, especially Dr. Wes Koczka, Michelle Cox, Maxine

Macgillivray, and Veronica Armstrong for helping me carry out the project at the University of Victoria. I am grateful to Dr. Parviz Maftoon, Dr. Fahime Marefat, Dr. Hamide Marefat, and Homa Khalaf, the faculty members at Allameh Tabatabai University and Free University, for implementing the experiment in their institutions and for the innumerable email messages providing me with information I needed. Special thanks to all EFL/ESL graduate and undergraduate students whose cooperation and participation in the experimental part of the study made this work possible.

I would also like to acknowledge the scholarship from the Ministry of Culture and Higher Education, Iran, which made my studies in Canada possible, the Graduate Teaching Fellowship from the University of Victoria which subsidized my graduate years, and the award from the Grants and Awards Committee of the TOEFL Policy Council which assisted me in the timely completion of my dissertation.

My thanks to my past and present fellow graduate students in the department, especially to Sandra Kirkham, Lili Ma, Vicky Man, Manami Iwata, Mavis Smith, Karen Topp, Suzanne Cook, Violet Bianco, Bill Lewis, Marie Louise Willet, and Melanie Sauer for their friendship and support. I also appreciate the statistical help of Eugene Deen from Computer User Services.

Finally, my heartfelt thanks to the members of my family: my father, for his unconditional love and never-ending support, my brother, for always being there when I need help, my husband, for his love, patience, and understanding, and my little boy, Hourmazd, for putting up with a part-time mother for all these years.

To all of them go many thanks again for their assistance and encouragement.

***To the memory of my very best friend,***
***my mother***

# *PART ONE*

# *PRELIMINARY REMARKS*

# CHAPTER ONE
# INTRODUCTION

## *1.1 Teaching and testing*

The proper relationship between teaching and testing has long been a matter of interest in both the educational and the applied linguistics literature. The fact that tests attract classroom teaching and the syllabus to themselves by requiring that the teachers teach and the students practice the activities which are necessary to pass the examination is a commonplace. Tests are generally used to make, among other things, inferences about test takers' abilities, predictions about their future success, and decisions (such as employment, placement, selection, etc.) about the examinee. The uses made of test results thus imply values and goals and have an impact on, or consequences for, the society and the educational system in general and individuals in particular (Bachman & Palmer, 1996).

In applied linguistics research, this influence of testing on teaching and learning has been referred to as *washback* [1] (see Hughes, 1989; Khaniya, 1990a; Alderson, 1991 among others), a phenomenon that depending on circumstances, can be both beneficial and harmful. As for researchers, depending on how enthusiastic they are about the role of testing in relation to teaching, they take different, sometimes contradictory, positions with respect to this matter. Some consider testing as detrimental to teaching by driving teachers and students away from the syllabus and towards the skills and activities required for passing the exams (Wiseman, 1961; Vernon,1956). Others (Davies, 1968, 1985) account for most testing as an obedient follower of its leader, teaching, while at the same time, if

---

[1] Another term referring to the same phenomenon is *backwash* (Heaton, 1988; Hughes, 1989). I will, however, use *washback* throughout this dissertation since it is the more commonly used term in applied linguistics research.

creative and innovative, capable of effectively changing the syllabus. A larger group of writers (Swain, 1985; Alderson, 1986; Pearson, 1988; Hughes, 1989), however, argue that efficient testing can create change by promoting effective teaching and learning. Still others (Morrow, 1986; Frederiksen & Collins, 1989) see the implications of test scores as so fundamentally important that they actually consider them as a validity requirement for the test. To them, a test is considered as invalid if the inferences made from test scores do not induce desirable changes in the educational system. Frederiksen & Collins (1989), for example, consider tests as critical stimulants in the educational system with the potential to bring about radical changes in teaching and learning methods. They introduce the concept of "systemic validity" as a quality of the test which has to do with the curricular and instructional changes eventually responsible for the further development of the skills primarily measured by the test. Likewise, the notion of "washback validity" has been suggested by Morrow in an attempt to enhance the development of language tests that are more likely to bring about positive washback effects.

Practically, however, the nature and the presence of washback has thus far been little studied, and as Alderson (1991) points out, "what there is is largely anecdotal, and not the result of systematic empirical research." The gap is so large that some writers (Alderson & Wall, 1993) even question the existence of washback. This is mainly because most of the existing studies on washback (see Chapter 3) are either indirect measures of washback, or lack the appropriate theoretical basis and adequate empirical comprehensiveness needed for studying such a complex phenomenon and distinguishing it from other factors operative in the educational system. That might be why, for example, in some studies (Wesdorp, 1982; Khaniya, 1990b; both cited in Alderson & Wall, 1993), the test impact is found to be much less than expected while, on the other hand, in some

others (Hughes, 1989) the preparation of students for a new test results in a considerably desirable effect on their performance on a very different test.

What is needed in a study of washback then is a clear definition of the version of washback adopted by the study, specifying the limits and aspects of the phenomenon (Alderson & Wall, 1993). Besides, critical factors assumed by a study to theoretically affect the occurrence of washback, as well as areas affected by the phenomenon, have to be specified and accounted for.

## *1.2 The study*

Existing empirical studies on washback (mostly reviewed in 3.6) are sparse yet increasingly informative, shedding light on issues that might have been otherwise left unnoticed. They are, however, somewhat similar in nature in that the focus of the research is on washback effects of high stakes tests of English as a Foreign Language (EFL)[2] administered overseas. Besides, most of these studies do not take a clear stand with respect to the relationship between a test's validity and washback, neither do they adopt or introduce any theoretical framework underlying their conceptual or methodological approach to the examination of washback. A potential danger of this is the incorporation and consideration of too many or too few variables in the process of accumulation, analysis and interpretation of the data; this inevitably results in contradictory evidence that fails to properly determine whether the observed negative/positive effects are due to the test itself or some other factors (political, administrative, budgetary, etc.) inside and outside the educational system. As for methodology, most of the studies have adopted

---

[2] Shohamy et al.'s (1996) study, is an exception in that it also focuses on ASL (Arabic as a Second Language) tests in Israel.

either a quantitative or a qualitative approach to the study of washback, and a few have used both methods of data analysis at the same time. Moreover, having focussed primarily on teaching contents and methodology, these studies have rarely studied tests' influence on learning outcomes.

The purpose of this dissertation is to examine test effects focusing on those factors not previously examined. The study will adopt both quantitative and qualitative approaches to the study of washback in two different ESL (English as a Second Language) and EFL contexts of language learning. The primary research question addressed is the presence of the washback phenomenon, its nature, and the extent to which it occurs. Also, it is intended to see if the presence and intensity of washback are affected by such factors as the test context (EFL vs. ESL), construct (oral vs. written ability), task, or status.

The principle adopted here is that tests of language ability can positively influence teaching and learning activities provided that their constructs and tasks are informed by the language needs of the learners, and that such attributing factors as non-test language use context, learners' motivation and background knowledge are accounted for. The areas most likely to be influenced by such tests are expected to be: (1) the materials used in the classroom, (2) teachers' methodology and teaching activities, and (3) learners' strategies and learning outcomes.

Theoretically, the washback phenomenon for the sake of this study will be defined in the light of recent advances in the theory of test validity. Different variables, active before and after the administration of a test, that play a crucial role in bringing about what is known as positive/negative washback will be delineated in a conceptual framework that

will serve as the theoretical basis underlying different steps taken in the experimental part of the study.

Empirically, a longitudinal research project whose main concern is the examination of the above-mentioned framework will be conducted in several phases which basically correspond to the different levels of the theoretical model adopted in the study. Three groups of subjects with different needs, proficiency levels, and in different learning environments have been the basis for the generalizations arrived at in this study.

## *1.3 The organization of the text*

The dissertation is organized into parts and chapters as follows:

Part I, *Preliminary Remarks*, includes the first chapter whose purpose is to introduce the topic, summarize the relationship between teaching and testing, and introduce a study which demonstrates how washback principles can be evaluated. Part I also summarizes the topics introduced in later chapters.

Part II of the thesis, *Theoretical Considerations*, consists of three chapters theoretically explaining and justifying the approach adopted in this work. The chapters in this part, while self-contained, are tightly related in that each chapter builds on the concepts and information introduced in the previous one while at the same time it serves as a basis for the information presented in the following chapter.

In Chapter Two, the most important characteristic of tests, *validity,* is discussed. The evolution of the validity concept over time is traced and the major breakthroughs in the process are highlighted. The main focus in this chapter is on two basic issues with

respect to the current theory of validity, namely, validity as a unitary concept, and the inclusion of the consequences of test use as part of the theory of validity. Such a discussion is of great significance in designing an evaluation of washback since, as we will see later, the washback phenomenon is in essence an instance of the modern theory of validity (Messick, 1989, 1996). So, the general validity theory laid out in this chapter serves as a foundation for the theoretical stand adopted throughout the study.

In Chapter Three, the concept of washback is examined as a phenomenon whose great significance for language testing theory and practice stems from its relation to the test's construct validity on the one hand, and its implications for a shift of interest from indirect discrete testing of skills to direct performance assessment of abilities on the other. To clarify the concept of washback in this study, the phenomenon is defined and its characteristic features are delineated. Furthermore, the mechanism through which it works, as well as the processes and factors contributing to the occurrence of positive washback, are discussed. The existing published literature on almost all empirical studies already carried out in this area are also critically reviewed with respect to their assumptions and methods of measurement. Finally, in this chapter, a model reflecting the conceptual framework underlying the methodological procedure adopted in this study is proposed. It not only clarifies the limits of the study, but also systematically presents the areas and participants that are inevitably affected in the process and are thus considered as reliable sources of evidence for any study of washback. The constituent parts of the model, therefore, underlie the major steps followed in the experimental study of washback, discussed in Part III of the dissertation.

Chapter Four of the thesis aims at theoretically justifying the move towards performance assessment in the field of language testing by emphasizing two characteristics

of performance testing, namely, authenticity and direct assessment of competence. They are also believed to be largely responsible for the beneficial washback effects of tests on teaching and learning. This and the fact that authenticity and directness are both aspects of a test's validity further support the idea, adopted in Chapter Two, that washback is in fact an instance or element of validity. Also discussed in this chapter is the framework proposed by Bachman & Palmer (1996) as a model reflecting the main characteristics of the test-taker's competence as well as the test tasks that link test and non-test domains of language use. The framework is to be used for the development of the testing instruments in this study since it potentially includes all conceptual notions applicable to the performance assessments of language abilities as described in this chapter.

Part III, *Empirical Considerations*, reports on a research project conducted at the institutional level to examine the above theory of washback. The experiment undergoes several phases that basically correspond to the different levels of the theoretical model introduced in Chapter Three. Three groups of subjects with different needs, proficiency levels, and in different learning environments participated in the study. Group one consists of international graduate students functioning as teaching assistants (ITAs henceforth) in an ESL environment at the University of Victoria, Canada. Groups two and three, on the other hand, consist respectively of Persian undergraduate and graduate EFL learners at Allameh Tabatabai University and Free University, Iran. While the members of the first group are primarily concerned with the development of their spoken-language ability, the subjects in the second and third groups are trying to increase their proficiency in writing.

Chapter Five concentrates on the first phase of the study: needs assessment. In order to identify the tasks that are of utmost importance to these learners, a systematic

assessment of the students' target language needs is conducted. The procedure adopted is similar to that of Munby (1981) in that it is, to a large extent, learner-based. However, information has also been gathered from stakeholders in the academic community directly or indirectly affected by the proficiency level of the subjects. In the case of ITAs, for example, native-speaking undergraduate students, supervisors, relevant departments, graduate advisors, university authorities and ESL teachers, are considered as such sources of information.

Once the objectives of our specific populations are set, in the second phase of the study, Chapter Six, a performance test whose major focus is the elicitation of the language behaviour illustrative of the needs of the population in question is developed for each group of subjects. The theory of language testing adopted in this phase is that of Bachman & Palmer (1996), presented in Chapter Four, since it addresses the question of the proper relationship between the test performance and non-test actual language use. Although, depending on the abilities and skills being tested, certain components of the model might be highlighted or left out, the communicative interactions between the components of the model are observed in the test design. The tests further include theoretically grounded rating instruments which enable the raters to assess the subjects' performance with respect to the rating categories that correspond to the components of the theoretical framework underlying the tests' tasks.

Having devised the tests, we then turn to the third phase of the study, the experiment. This is in fact, the final step in examining our theory of washback and describes an experiment conducted with the purpose of assessing the negative/positive washback effects of our testing instruments on every aspect of the training programs based on them. Homogeneous groups of learners were chosen on the basis of their English

language scores at the time of entry (i.e., TOEFL scores for ITAs and the English language score in the University Entrance Examination for Persian subjects). Before the start of the program, the subjects were required to take the performance tests developed for the purpose of this study for two main reasons: (i) to exclude from the program the candidates who already possessed language abilities measured by the test, and (ii) to have a set of scores for the candidates who were going through the program for the sake of comparison with their end-of-term scores on the same test. Teachers were provided with thorough information concerning the objectives, nature and the theoretical background of the tests. Raters were also given a detailed explanation of the performance categories used in the rating instruments so that they knew what to look for in the performance of the learners. The length of the training program was one semester, at the end of which the relevant performance tests were administered again. In the course of the program, the training sessions were observed so that in addition to what teachers and students reported in questionnaires or interviews regarding their motivation and reactions to the program and the test, the direct observations of the researcher or a third party can shed further light on other aspects of our washback theory: material development/choice, teaching methodology and learners' activities reflecting the learning brought about in learners. This qualitative approach to data gathering is of significance to a study of washback since it is an effective method for distinguishing test effects from those of other factors (such as good teaching or exceptionally high motivation on the part of the learners) present in an educational system.

Chapter Eight focuses on the analysis of the data with respect to the three major areas where washback effects are most likely to appear, i.e., the development/choice of the material, the teaching methodology, and the learning strategies. The results of the analysis

of the data, gathered through qualitative research methods, is used to describe: (i) whether or not the materials used illustrated, presented, and developed the skills assessed by the test, (ii) if the teaching activities and teachers' methodology were in the direction of the tests, and (iii) if the learning strategies adopted by the learners were affected in any way by the test. The results of the quantitative analysis of the data, on the other hand, are used to show whether the materials and the teaching and learning activities did in fact increase the achievement of the skills promoted and measured by the tests: if yes, to what extent, if not, why.

Part IV, *Concluding Remarks*, includes the final chapter, *Conclusions and Implications*. It concludes the dissertation with conclusions drawn on the basis of the quantitative and qualitative analysis of the data, summarizes the implications of the research on washback effects for a general theory of language teaching and testing, and makes some suggestions for further research.

# *PART TWO*

# *THEORETICAL CONSIDERATIONS*

# CHAPTER TWO
# VALIDITY

## *2.1 Introduction*

Validity, the foremost requirement in test evaluation, refers to the ability to make adequate, appropriate, and useful inferences from test scores[1], and the validation of a test is an ongoing process of accumulating evidence to support particular inferences made from the scores. A test is thus said to have validity based on the degree to which this evidence is consistent with the interpretations and actions determined on the basis of the test scores.

In examining the validity of a test, there are a few fundamental issues that have to be taken into consideration. First, what is to be validated is not the test or the assessment instrument itself but the inferences derived from the test responses. The fact that only test responses have validity is an established crucial point in test validation since responses result from an interaction between the test tasks and items, test takers, and the test context. Also, depending upon the specific uses to which we want to put these interpretations, we have to go beyond just the accuracy of the test scores and also consider the functional worth of scores in terms of social consequences of their use. Besides, although the evidence for validity can be obtained from a variety of sources, validity is a unitary concept encompassing a range of empirical and theoretical rationales behind the test score interpretations (American Psychological Association, 1985). This unified conception of validity has recently been the subject of heated debates in the field of

---

[1]The term *score* is used throughout this dissertation in its "most general sense of any coding or summarization of observed consistencies on a test, questionnaire, observation procedure, or other assessment device." (Messick 1989, p. 14)

testing in that there is a disagreement over how to define validity and what to subsume under construct validity, a topic to which we will turn in the subsequent sections.

The main purpose of this chapter then is to address two basic issues with respect to the current theory of validity, namely, validity as a unitary concept, and the inclusion of the consequences of test use as part of the theory of validity. Such a discussion is significant for the present study since the phenomenon under investigation here, i.e., washback effect, as a consequence of testing, is in essence an instance of validity. So, the general validity theory laid out in this chapter serves as a foundation for the theoretical stand adopted throughout the study. However, to clear the way for a discussion of validity in relation to the present trends in testing, it is necessary to briefly review the traditional approaches to validity first.

## *2.2 Traditional approaches to validity*

Validity has traditionally been conceived as comprising three different types, at least since the early 1950's as documented by the manuscripts[2] periodically issued by the three sponsoring organizations[3] guiding the development and use of tests. The *Standards for Educational and Psychological Tests and Manuals* (APA, 1966)[4], names the three aspects of validity as content validity, criterion-related validity, and construct validity[5].

---

[2] These documents are going to be referred to as *Standards* throughout this dissertation after this first citation.

[3] Namely, The American Educational Research Association (AERA), The American Psychological Association (APA), and The National Council on Measurement in Education (NCME).

[4] This is the third document published by APA replacing the earlier two, i.e., *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (APA, 1954), and *Technical Recommendations for Achievement Tests* (APA, 1955).

[5] Face validity, the term solely used to refer to what a test looks like, does not have a theoretical basis and cannot be considered as a basis for the interpretive inferences made from test scores. Face validity has thus never been seriously considered as an aspect of validity in the testing literature. For more discussion of this topic see Cronbach (1984) and Bachman (1990).

These concepts still persist in the current theory and practice of testing but have undergone major modifications and refinements. In the following two subsections, I will present validity types in their traditional versions and go over the concept of validity and its evolution over the years in an effort to provide a basis against which one can appreciate the current trends in testing.

### *2.2.1 Types of validity*

#### *2.2.1.1 Content validity*

Content validity is a validity concept primarily concerned with the content of a test and the degree to which it represents a particular ability or content area. The basic source of validity evidence for content validity, according to the 1966 *Standards* (APA, 1966), is the subjective evaluation of the test content in relation to the assumed universe of tasks, conditions, or processes. Major methods suggested by the *Standards for Educational and Psychological Tests* (APA, 1985) for demonstrating such evidence include: (i) expert judgments of the relationship between the parts of the test and the defined universe; (ii) empirical procedure of specifying the major facets of a domain of academic subject matter and allocating test items to the categories defined by those facets; and (iii) systematic observation of behaviour in a job with the intention of providing samples of the job domain that are representative of the critical aspects of the job performance while leaving out the relatively unimportant aspects. In his definition of content validity, Messick (1989) also acknowledges "professional judgment" as a basis for determining the "relevance" and "representativeness" of test content. Bachman (1990), on the other hand, clearly defines content validity as having two aspects: content relevance and content coverage. The two

categories roughly correspond to Messick's relevance and representativeness, but Bachman claims that content relevance involves not only the specification of the "ability" domain, but also that of the test method facets that define the measurement procedures. Examples of such measurement procedures include the specification of what the test measures, the attributes of the stimuli presented to the test taker, the nature of the expected responses (Popham, 1978; Hambleton, 1984), aspects of the setting in which the test is given (Cronbach, 1971) and so forth.

There are, however, some problems associated with considering content validity as such as the sole basis for validity. In tests of language proficiency, for instance, it is not always easy to specify the content evidence on the basis of which a particular test's content coverage and content relevance can be demonstrated (Bachman, 1990). In a test of speaking ability following a course of instruction, for example, the domain specification might range from the teaching points included in the curriculum to the actual content of the instruction including the grammatical forms, strategies, and illocutionary acts used in oral interactions during the course between the teacher and the students or between the students themselves. Added to these are non-linguistic factors, such as the physical conditions of the teaching environment, time, sex, age, and other characteristics of the participants.

Even if the content domains of language abilities could be determined clearly, a second problem of relying on the evidence of content relevance and content coverage alone, pointed out by Bachman, is the limitation imposed by the specified domain on the inferences made from test performance. In other words, the examiner can only make judgments about what the test taker is able to do with respect to the content area from which the test tasks are sampled. In language tests with the purpose of determining the

areas of *in*competency or *in*ability (Madaus, 1983; Linn, 1979, 1983), the limitation of the content-based interpretation is even more problematic since content relevance alone cannot be a basis for inferring *inability,* which involves a number of factors other than lack of ability (Messick, 1980).

However, the most important concern with content relevance as the only source of validity, is that content validity – as defined earlier in this section – has to do with the test instrument rather than the inferences made from the test responses/scores. This is a very important point because even though the information about the test content might be an accurate illustration of what tasks and abilities are included in the test, it does not give any clue to individual performances. That is why any differences observed between the performances of different groups of individuals on the same test are attributed to the test responses rather than the test instrument itself. It is, therefore, the test response that varies across individuals, not the test content. In other words, even though the issues of content relevance and representativeness are necessary requirements for score interpretation (Messick, 1980), the evidence of content validity, as a property of the testing instrument rather than the test scores is not sufficient for validity in general and has to be supplemented by other forms of evidence.

### *2.2.1.2 Criterion validity*

Unlike content validity which characterizes the test in relation to the specific domain for which it is intended, criterion validity demonstrates the relationship between the test scores and some other variable believed to be a criterion measure of the ability tested. The 1966 *Standards* defines the criterion measure as "a direct measure of the characteristic or behaviour in question" (APA, 1966, p. 13). However, the criterion can be the performance

on another measurement instrument or task – direct or indirect – that involves the same ability, or the "level of ability as defined by group membership" (Bachman, 1990).

Of prime importance in studies of criterion-related validity is the choice of the criterion measure upon whose relevance and appropriateness lies the value of the study. Not every testing instrument that measures the ability in question qualifies as a criterion measure since there are a number of factors that have to be taken into consideration before making such a choice. First and foremost, the criterion measure should possess validity itself, otherwise it is meaningless to validate a test against a criterion whose appropriateness and adequacy are not investigated. Studies based on criteria "chosen more for availability than for a place in a carefully reasoned hypothesis, are to be deplored" (APA, 1974, p. 27). Other factors interfering with the accuracy of criterion-related studies are the limitations imposed on the data due to the inadequate number of cases, the non-representativeness of the samples with respect to the population for which the resulting inferences are intended, access to meaningful criterion measures, and the changing conditions in the course of the study which make the accuracy of the predictive studies questionable.

Depending on whether the test is conducted for prediction of some future performance or for the assessment of the present status, criterion-related validity might be referred to as "predictive" or "concurrent" validity respectively. The most common use of a concurrent validity study concerns the examination of correlations among various measures of language ability in order to measure a specified construct. Nevertheless, as already mentioned above, a serious problem with this is whether or not the criterion measure itself is a valid test of the ability in question and correlates with other tests of the same ability. This problem, as Bachman (1990) rightly states, will lead to an "endless spiral

of concurrent relatedness"; however, one way to provide such evidence is the process of construct validation which will be discussed in some detail below. Still another limitation involving concurrent criterion-relatedness is that it is solely concerned with correlations between a specific test and the criterion, both of which measure the same ability. As such, concurrent validity ignores the extent to which these measures do not agree with measures of other abilities. At best it will tell us that a test of a given ability is related to other measures of the same ability but not that it is not related to measures of other abilities, a kind of evidence that leads us far beyond the limits of concurrent validity and once again into the process of construct validation.

The predictive use of criterion-related validity is more commonly used for selection purposes in educational and professional contexts. However, a potential problem with relying on predictive validity alone is that language tests designed for the purpose of prediction cannot be considered as valid measures of any particular ability due to the fact that the changing test properties as well as the conditions surrounding the test situation and the test-taker might affect the correlation between a test now and a future criterion (Cattell, 1964). Moreover, prediction, while being an important use of the language tests, is not the only thing we are interested in. Language tests are also used in a variety of educational settings for the purpose of determining the test takers' levels of ability to perform certain tasks. In this latter case, a clear theoretical conception of the behaviour in question is necessary because language proficiency is viewed as a theoretical construct and not as a pragmatic ascription which constitutes the thought underlying tests of prediction (Upshur, 1979). Thus, predictive criterion-relatedness cannot by itself constitute evidence for interpreting scores as indicators of abilities. Once again, it is the process of construct validation that provides us with enough evidence for making such an inference.

### *2.2.1.3 Construct validity*

While content and concurrent validities are mainly concerned with empirical relationships, construct validity accounts for the extent to which a test performance is compatible with the theory underlying the description of the behaviour being tested. It is especially relevant when the psychological characteristics of the ability or behaviour being measured are to be determined. Listening comprehension, sound/form correspondence, and reasoning ability are examples of constructs which can be measured by particular tests. The process of construct validation thus requires a conceptual network that specifies the characteristics/nature of the construct in question with such clarity that it can not only be distinguished from other non-related constructs but also from related but dissimilar ones.

The evidence for construct validity could come from a variety of sources ranging from intercorrelations between the test items, between different measures of the same construct and between different methods of measurement to experimental designs aimed at observing a certain type of behaviour in an attempt to examine a construct for which the researcher finds – or accepts – no existing criterion measure. The spectrum obviously includes the evidence for content relevance and representativeness as well as criterion relatedness since validation studies conducted for such purposes have implications for score interpretation and when supplemented by other evidence can contribute to the construct validation process. Also important is the information gathered qualitatively by questioning testers, test-takers and raters concerning their methods, performance and scoring strategies. Of course, not all these lines of evidence are necessary in a particular case of construct validation, and depending on the specific problem at hand, one can choose one or more approaches to gather evidence. In fact, the more the information

related to the score interpretation supports its underlying theoretical rationale, the stronger the evidential basis for the construct validity will be.

Thus, in spite of the unidimensionality of content and criterion-related validities, the three types of validity discussed thus far, taken together, include all kinds of validity information mainly because of the comprehensiveness of the reference of construct validity which has been considered as a unifying concept for test validity:

> Construct validity is indeed the unifying concept that integrates criterion and content considerations into a common framework for testing rational hypotheses about theoretically relevant relationships (Messick, 1980, p. 1015).

However, Messick also maintains that construct validity as such is meant not only for test score interpretation but also for "test use" justification, an issue to which we will return towards the end of the following section and on which the main focus of section 2.3 will rest.

### *2.2.2 The evolution of the validity concept: A historical approach*

An overview of the theoretical conceptions of validity over the past few decades (Anastasi, 1986; Angoff, 1988; Cronbach, 1989; Messick, 1989) reveals a shift of focus from the prediction of specific criteria (Guilford, 1946) to a limited number of validity types and finally to a unitary validity view (Messick, 1989; Cronbach, 1980 as cited in Shepard, 1993). An important aspect of this gradual transition, as illustrated by the following review of the historical trends in the field, is the move towards a style of validation which goes beyond the observable, specific, local concrete situation.

In his 1946 article, Guilford recognized two major types of validity: factorial and practical. While factorial validity stands for the reference factors responsible for ensuring that a test is measuring what it is supposed to measure and of appropriate dimentionality, he considers practical validity as a more comprehensive standard of test evaluation identified by a test's correlation with a practical criterion. In his view, therefore, validity is a matter of prediction, and "a test is valid for anything with which it correlates" (p. 429).

Almost a decade later, *Technical Recommendations* (APA, 1954) broke validity into four distinct types of content, predictive, concurrent and construct validities, a framework later adopted by Cronbach (1960) and Anastasi (1961). This list was then modified and reduced into three categories of content, criterion-related, and construct validities in the 1966 *Standards* by blending together the predictive and concurrent subtypes into one major category called criterion-related validity (also Cronbach, 1970; Anastasi, 1968). As previously indicated, these concepts have survived in the field of measurement but with many refinements initiated by this same edition. In this edition, as well as that of 1954, the three validity types have been associated with the three different aims of testing, namely, determining the achievement of certain educational objectives, predicting the individual's present or future performance with respect to an established variable used as a criterion, and inferring the degree to which the individual possesses some trait or quality. However, the 1966 *Standards* does not draw strict lines among the three types of validity by acknowledging that they are only conceptually separate and emphasizing that a particular test's validity requires information about all kinds of validity, not just one of them.

This move towards a unitary conception of validity is also followed in the 1970s, witnessing an increasing emphasis on construct validation and its comprehensiveness as an

all-embracing concept (Cronbach, 1970; Anastasi, 1976). The trend is clearly reflected in the 1974 *Standards* (APA, 1974) according to which validity aspects "are interrelated operationally and logically" and "only rarely is one of them alone important in a particular situation" (p. 26). Also, the reference in this edition to the notion of content validity as an indicator of the relevance and representativeness of the test "behaviours" – rather than "content" – in relation to those of the desired domain implies the need for construct-related evidence that the test behaviours are representative samples of the domain behaviours.

In his 1980 paper, however, Messick, while considering construct validity as a base upon which other approaches rest, argues in favour of a terminological reform in an attempt to recapitulate the specific validity procedures that might be singled out for answering specific practical questions. He reserves the term *construct validity* for referring to the procedures underlying the inferences made with respect to the meaningfulness of the test scores. On the other hand, he labels content validity as content relevance and content coverage to refer to procedures leading to domain specification and domain representativeness. Predictive and concurrent validation are also referred to as predictive and diagnostic utility, respectively.

Translating labels into procedures, Anastasi (1986) regards content and criterion validation as stages in the construct validation of the tests since such procedures eventually contribute to the construct validation. As already mentioned in the previous section, the fact that in a criterion-related validation, the criterion itself has to be investigated for validity brings construct validation into the picture. Similarly, construct validity is called for in content validation where the choice of the content to which the test conforms has to be theoretically and practically justified. The three types of validity are

thus "no more than spotlight aspects of the inquiry" (Cronbach, 1984), all contributing to the same validity goal that is "explanation" rather than mere prediction: "The end goal of validation is explanation and understanding. Therefore, the profession is coming around to the view that *all* validation is construct validation" (Cronbach, 1984, p. 126). The view is also maintained by the latest edition of *Standards* (APA, 1985) that refers to validity as a unified concept, emphasizing its preference for "obtaining a combination of evidence that optimally reflects the value of a test for an intended purpose" (p. 9).

Hence, construct validity, as reflected in the measurement textbooks and professional standards so far discussed, embraces all types of validity evidence. However, in the preface to the 1985 *Standards*, the "growing social concerns over the role of testing in achieving social goals," although not directly discussed in relation to the validity issue, has been mentioned as one of the reasons for the revision of the 1974 *Standards* . Earlier, Messick (1980) had suggested that the social values and social consequences of the test use have to be taken into consideration in a discussion of validity. This concern has also been echoed by Cronbach (1988) who believes that the validation process "must link concepts, evidence, social and personal consequences, and values" (p. 4). According to him:

> the bottom line is that validators have an obligation to review whether a practice has appropriate consequences for individuals and institutions, and especially to guard against adverse consequences. You ... may prefer to exclude reflection on consequences from meanings of the word *validation*, but you cannot deny the obligation (p. 6).

But Messick, in his chapter on validity (1989), goes even one step further, claiming not only that construct validity is the whole of validity, but also that we have to consider the social consequences of tests as part of the construct validity since the appropriateness, meaningfulness and usefulness of the inferences made on the basis of the test scores depends as well on the social values and social consequences of the test use.

To sum up the discussion so far, the transition in validity conception from three distinct traditional types to a unified concept has gradually taken place over years in the field of testing to the extent that it is now acknowledged by almost all prominent texts in this area. However, the nature of the concept of validity is still open to controversy, and the specific area of debate has to do with the concept of construct validity. The major question is what to include under construct validity other than the recognized content and criterion-related validation procedures, if any. And more specifically, should testing consequences be subsumed as an aspect of construct validity as Messick puts it?

It is basically the major disagreement over how to answer these questions which fuels the current controversy in the field of testing, a subject I am going to examine in some detail in the next section.

## *2.3 Current issues in validity*

As we have seen so far, validity has been traditionally viewed as consisting of three major categories of content-related, criterion-related and construct evidence. However, while both content and criterion-related evidence ultimately contribute to the meaningfulness of the test scores; i.e., the process of construct validation, neither of them can be solely responsible for a test's validity. This has gradually led the field of testing towards the acceptance of a unified view of validity overarched by construct validity.

During the recent years, nevertheless, an issue of considerable debate has been the role of consequences in the theory of validity first put forth and formulized by Messick. The topic attracted enormous attention from the scholars in the field of psychological and educational testing and soon became the subject of controversy in this area. To set the theoretical stage for the issue at hand in this dissertation, this section will concentrate on this "post-Messick" era in testing and review the literature on validity as the concept further evolves. To do so, I will first summarize the validity conception as proposed by Messick since almost all subsequent theoretical writings on this topic are triggered by his viewpoints as reflected in their attempts either to argue in favour of or against his position or to elaborate his relatively abstract concepts.

In 1980, Messick argued that for a fully unified view of validity, the evaluation of the intended and unintended social consequences of testing is necessary. Consequently in a 1989 attempt to formally incorporate the test consequences into a consideration of validity, he proposed a new way of categorizing validity evidence that not only emphasizes the centrality of construct validity but also accounts for the test's value implications and social consequences. His framework (as illustrated in Table 2.1) consists of two major facets of validity: (i) the source of justification of the testing being either an evidential basis or a consequential basis and (ii) the function or outcome of the testing being either test interpretation or test use.

**Table 2.1: Facets of Validity**
**(Messick, 1989, p. 20)**

| | TEST INTERPRETATION | TEST USE |
|---|---|---|
| EVIDENTIAL BASIS | Construct validity | Construct validity +Relevance/utility |
| CONSEQUENTIAL BASIS | Value implications | Social consequences |

As illustrated by the above framework, the evidential basis of test interpretation is construct validity. It provides evidence concerning the meaning of the item or test scores and their theoretical relationships to other constructs. The evidential basis of test use provides further theoretical evidence supporting construct validity in terms of the relevance of the test to the specific applied purpose and its utility in the applied setting. To justify the inclusion of construct validity in this cell, which has to do with test use rather than test interpretation, Messick argues that the empirical appraisal of the issues of relevance and utility depend on the meaning of the test score, i.e., the process of construct validation.

The lower cells of the table, on the other hand, reflect the two components of the consequential basis of validity that are more related to issues arising from social contexts and applied settings. The consequential basis of test interpretation, according to Messick (1989), is "the appraisal of the value implications of the construct label, of the theory underlying test interpretation, and of the ideologies in which the theory is embedded. A central issue is whether or not the theoretical implications and the value implications of the test interpretation are commensurate, because value implications are not ancillary but,

rather, integral to score meaning" (p. 20). So, the evidence related to both construct validity and its consequences are included in the consequential basis of test interpretation. The second component, the consequential basis of test use, accounts for both the potential and actual social consequences of the test when used. Assuming that social consequences require evidence of score interpretation/meaning and at the same time contribute to the evidence of the test's relevance and utility, the consequential basis of test use should, therefore, include all types of evidence included in other cells.

As such, the validity framework in Table 2.1 is designed to be a "progressive matrix" (Messick, 1989) with construct validity preserving its centrality by appearing in every cell and being enriched by the evidence of the relevance and utility of the test, value implications of test interpretation and social consequences of test use. More precisely, Messick's conception of validity implies that construct validity be taken as the whole of validity with the evidence coming before the consequences if they are ranked in the order of priority. We can summarize Messick's view of validity as (1) below:

(1) Evidence + Consequences —> Construct Validity —> Validity

Subsequent to the publication of Messick's influential chapter on validity in 1989, there appeared in the field of psychological and educational testing a number of articles providing pro and con arguments regarding the evaluation of test consequences as part of validity considerations. In the remaining part of this section, some of the most prominent of these discussions are reviewed.

In an attempt to simplify the concept of construct validity, Kane (1992) introduces an alternative argument-based approach to test validation based on what Cronbach (1989)

recommends as a strong program of hypothesis-dominated research, i.e., a strong program of construct validation.

> The argument-based approach to validation adopts the interpretive argument as the framework for collecting and presenting validity evidence and seeks to provide convincing evidence for its inferences and assumptions, especially its most questionable assumptions. One (a) decides on the statements and decisions to be based on test scores, (b) specifies the inferences and assumptions leading from the test scores to these statements and decisions, (c) identifies potential competing interpretations, and (d) seeks evidence supporting the inferences and assumptions in the proposed interpretive argument and refuting potential counterarguments (Kane, 1992, p. 527).

The quality and quantity of evidence in this approach is thus not necessarily dependent upon theory-based interpretations but on the inferences and assumptions in the interpretive argument. As such, it not only presupposes test consequences as part of test validity but also does not highlight any kind of validity evidence as being central or preferable to others. That is why Kane uses the term *argument-based approach,* rather than construct validity, to emphasize the applicability of this approach to theoretical constructs as well as to the applied setting. So, while Kane's approach does not basically leave anything out of construct validity, it adopts a less objective view of validity compared with that of Messick.

In the same vein as Kane, (1992) and borrowing closely from Cronbach (1988, 1989), Shepard (1993) suggests that the process of test validation be started with the question "What does the testing practice claim to do?" (p. 429), around which the

gathering of evidence would be organized. In this way, she not only contends that test consequences are a logical part of test validity – as Messick puts it – but also goes one step further by emphasizing that the focus should be more centrally placed on intended test use. While agreeing in essence with Messick's ideas on validity, she questions his *faceted* presentation (see Table 2.1) of this unified concept; and in spite of Messick's emphasis that none of the facets can be considered independently, her main concern is that placing construct validity in the upper-left cell, with the other cells contributing evidence to it, implies that the traditional "scientific" version of construct validity is being given priority over a consideration of value issues. Her view of validity can then be summarized as follows, with priority being given to test use and consequences rather than to scientific evidence:

(2) Consequences + Evidence —> Construct Validity —> Validity

Bachman & Palmer (1996), on the other hand, while acknowledging the significance of the value implications of test interpretation as well as the social consequences of test use for the development and use of language tests, prefer to keep the consideration of test use consequences out of a discussion of construct validity. They consider test use consequences – which they name test "impact" – along with authenticity and interactiveness as three test qualities which together with the reliability, construct validity, and practicality form the qualities of test usefulness. They further suggest that a discussion of test consequences is "important enough to the development and use of language tests to warrant separate consideration" (p. 42, n. 4). The authors, therefore, define construct validity in relation to two aspects of score interpretation: (1) the extent to which interpretations made on the basis of the test scores are indicative of the ability in

question and (2) the generalizability of score interpretations to other language use contexts.

Popham (1997), however, opposes Messick (1989) and Shepard (1993, 1997) by presenting an argument revolving around three points: the efficiency of the 1985 *Standards*, the confusion caused by cluttering the concept of validity with social consequences, and the test-use consequences being the business of test developers/users, not an aspect of validity. He specifically advocates the 1985 *Standards'* view of validity as referring to the accuracy of score-based inferences as opposed to that of Messick, asserting that "what needs to be valid is the meaning or interpretation of the scores as well as any implications for action that this meaning entails" (1995, p. 5). According to him, the concept of validity, if mixed with social consequences will unnecessarily confuse educational practitioners, who already have a problem digesting the fact that validity is indeed a property of test scores, not the test itself. Believing that one of the motives for adding these consequential *trappings* to the concept of validity is to draw the attention to the unsound uses of test results, Popham suggests that test consequences be left to be addressed by the test developers and test users. He, nevertheless, acknowledges that the concern about the consequences of test use is correct and of utmost importance which should be taken into consideration by every measurement person, but he does not want to include it as a part of a validity framework.

Arguing in favour of Popham (1997), Mehrens (1997) goes even beyond this by articulating what Popham implies; i.e., his preference for the traditional three-way distinction used for different kinds of validity evidence, and by questioning the legitimacy of construct validity as the whole validity by stating that "such reductionist labeling blurs distinctions among types of inferences" (p. 17). Mehren's main argument concerns the

point made by Shepard (1997) that the consequences of a test's use have to be taken into consideration in determining the construct validity of a test's interpretation if that test is intended for a particular use. The problem with this view, according to him, is that an inference based on a test score just informs one if the test measures the construct; the meaning and usefulness of the construct for any further action based on such an inference is a separate issue that has to be handled outside the realm of test validity. He, therefore, suggests that the concept of validity be *narrowed* rather than expanded and be reserved solely for "determining the accuracy of inferences about (and understanding of) the characteristic being assessed, not the efficacy of actions following assessment" (p. 18). Popham's (1997) and Mehrens' (1997) tendency for re-establishing multiple forms of validity can, therefore, be summarized in (3)[6]:

(3) Content Validity + Criterion Validity + Construct Validity —> Validity

However, as the vice-chair of the committee that developed the 1985 *Standards*, Linn (1997) argues that "the solution is not ... to artificially narrow the concept [of validity]" (p. 15). He maintains that even the 1985 *Standards*, which Popham (1997) praises and repeatedly recommends as a true picture of validity, does not characterize the concept as narrowly as that reflected in Popham's interpretation of the *Standards*. Consider the following opening statements of the *Standards*' chapter on validity:

> Validity is the most important consideration in test *evaluation* . The concept refers to the *appropriateness*, meaningfulness, and *usefulness* of

[6] The view is also shared by Zimiles (1996).

> the specific inferences made from the test scores (APA, 1985, p. 9, emphasis added).

All three words emphasized above, Linn (1997) argues, push beyond Popham's narrow characterization of validity as the "accuracy of score-based inferences." However, like Messick and Shepard, Linn believes that the Standards can still be improved since there are issues such as the inclusion and evaluation of the major intended and plausible unintended negative consequences of test uses (Shepard, 1993) that require further expansion and clarification.

Given this landscape, one can say that the concept of validity is still evolving in spite of the major breakthroughs in the process. However, there are certain facts that I would like to bring to the reader's attention in the next section in order to justify the conception of validity assumed in the present work.

## *2.4 General view of validity adopted in this study*

The existing discrepancy and confusion between authors with respect to the correct conceptualization of validity stems in large part from a concern about overburdening the validity concept. This complicates the already never-ending process of validation even further. However, two points are worth mentioning here: First, the term *consequential validity*, often used in the literature in conjunction with the current controversy over the role of test consequences has never been used by Messick himself. Instead, he refers to the term *consequential basis/aspect* of validity as only "a part" of the evidence required for test validation (1989, 1996). The term is misleading in that it is generally used to imply that a new kind of validity has been created, and more specifically, that a test is

automatically rendered invalid if there are negative consequences attached to its use. Messick clearly avoids both of these in his proposal.

Second, as Shepard (1997) convincingly argues, a concern for consequences of test use in relation to the validity is not a new phenomenon and has been part of the "underlying network of relationships that frame a validity investigation" (p. 5) for decades. The concern had already been voiced by Messick's predecessors, Cureton (1951) and Cronbach (1971), who respectively wrote the validity chapters in the first and second editions of *Educational Measurement*:

> The essential question of test validity is how well a test does the job it is employed to do. The same test may be used for several different purposes, and its validity may be high for one, moderate for another, and low for a third (Cureton, 1951, p. 621).

Cronbach, too, made a similar point and clearly related test validity to its uses by noting that validity is not a property of the test and has to be re-evaluated for each new application. He also included the *decisions* made on the basis of the test-scores as part of the evidence needed for validity considerations:

> A decision is a choice between courses of action. The college admits or rejects a prospective student. The high school allocates an algebra student to a fast, average, or slow section. The primary school decides that one child should be taught to read immediately, and another should first practice on auditory and visual discriminations. The justification for any such decision is a prediction that the outcome will be more satisfactory under one course of action than another. Testing is intended to reduce the

> number of incorrect predictions and hence the number of decisions that will be regretted later. When validating a decision-making process, the concern is with the question: What is the payoff when decisions are made in the proposed way, and how does this compare with the payoff resulting when decisions are made without these data? (p. 448)

So, for test-based decisions to be included as validity evidence, test use consequences have to be part of the validity picture, as an answer to questions about "worth" (Cronbach, 1988)[7]. Besides, earlier editions of the *Standards* (APA, 1954, 1955, 1966) all show concern for test uses by referring to "achieving aims" as part of the validity definition.

As it can be seen, then, there have been two different but related facets to the evolution of the concept of validity (Linn, 1997): one dealing with the accuracy of score-based inferences (as Popham, 1997 puts it) resulting in the unified theory of construct validity, and the other one concerned with the purposes to which the test scores are to be put leading to the recognition of the functional perspectives of the tests as validity evidence. As far as the former view is concerned, a meaningful construct should fit within a nomological network (or conceptual network) – traditionally associated with the validity studies – that serves as an organizing framework showing the construct to be measured and its hypothesized relationship(s) to other constructs and behaviours. Within such a framework, it is possible to evaluate the construct validity of a test without attending to the test uses. Examples of such tests are the cases in which the purpose of testing is merely to determine the presence or absence of some ability/behaviour apart from its

---

[7] See also the quotation in section 2.2.2, page 11.

significance for subsequent possible applications.

But how many testing situations are imaginable where the test developers/users aim for test results as tools for descriptive purposes not as a basis for some further action or decision? As Bachman (1990) points out, "tests are not developed and used in a value-free psychometric test-tube, they are virtually always intended to serve the needs of an educational system or of society at large" (p. 279). Consider the following hypothetical testing situations proposed by Kane (1992) and Popham (1997) supposedly to argue for and against the inclusion of consequential evidence in a validity framework: Popham introduces a testing situation in which valid inferences can be made concerning the achievement status of the students on the basis of the scores of an achievement test that is a representative sample of the intended content domain (p. 11, Figure 2). This is true, but the problem is that as soon as he decides to use those inferences for the purpose of making educational decisions (in this case placement in middle school), he will have to incorporate additional evidence to establish the validity of the score-based inferences for assigning the students to appropriate programs. Among such evidence is the predictive validity for future performance, a type of evidence that normally is not part of the picture unless decisions such as placement are involved. The same argument also holds true for the analysis of Kane's testing situation where the validity in question is that of an algebra test as a prerequisite for calculus. Here, too, the evidence supporting the interpretation of test scores will just inform us about the adequacy of the test as a valid measure of algebra skills, however, whether or not the test is suitable for making differential placement decisions crucially depends on how well it can predict the students' performance in the calculus course. "The relationship in the nomological net between prerequisite algebra

skills and later success in calculus is central to the meaning of the construct, not an addon social nicety" (Shepard, 1997, p. 7).

As such, the intended consequences of a test's use, or indeed the second facet referred to above, becomes part of the rational relationship represented in the conceptual framework and thus has to be taken into consideration in a validity investigation. What Messick introduced in 1989 was a framework that incorporated both of the above facets into one unified validity theory. He extended the notion of construct validity to include the evidence pertaining to the consequential aspect of the test's use, the evidence that for the reasons mentioned above cannot be ignored in an all-embracing theory of validity. Therefore, for a thorough analysis of the evidence supporting a certain conclusion, the validity concept should be *expanded*, Mehrens (1997) and Popham (1997) notwithstanding.

As for the unintended testing consequences, they are not automatically considered by Messick (1989) as evidence of invalidity unless they originate from construct over- and under-representation. So, plausible unintended side effects of a test, not the effects of test misuse, have to be taken into consideration in a validity investigation not only to minimize the adverse impact but also to guard against the two major threats to validity, namely, "construct-irrelevant variance" – i.e., the presence of irrelevant components in the construct definition that interfere with the assessment of the ability in question – and "construct under-representation" – i.e., the absence in the construct definition of those elements that are important to the assessment of the focal construct.

The general view of validity assumed in this study will thus consider the consequential basis of test use as part of the validity theory that, I believe, should be

comprehensive enough to serve as a unifying and, at the same time, analytical framework for all imaginable situations of testing. Against this background, I am going to turn to a particular instance of test consequences in the upcoming chapter, referred to as "washback" in the applied linguistic research literature.

# CHAPTER THREE
# WASHBACK

## *3.1 Introduction*

In applied linguistic research, the consequences of language test use are mostly sought after in the areas of language teaching and language learning, in that tests are believed to be in control of what happens in classrooms – a phenomenon referred to as the *washback* effect. References in the literature on language teaching and testing to the effects of tests on teaching and learning date back at least four decades; and the fact that tests affect teaching and classroom activities has always been acknowledged by both the theoreticians and practitioners in the field of applied linguistics. However, what has not yet been agreed upon is why and to what degree washback effect should be taken seriously in applied linguistics in general and language testing in particular. Unfortunately, very few empirical studies have been conducted to date with the purpose of systematically studying the nature, the type, and the extent to which washback effect takes place. Nor have there been detailed theoretically-based accounts of the tests that are expected to trigger such an influence. This is partly because there has been no conceptual framework which incorporates all the various key players in a theory of washback to serve as a working model for studying this complex phenomenon empirically.

This chapter examines the concept of washback as a phenomenon whose significance for language testing theory and practice stems from it being related to the test's construct validity, on the one hand, and its implications for a shift of interest in the field of testing from indirect discrete assessment of skills to performance assessments of abilities on the other. The main concern here is to come up with a conceptual framework

within which one can conduct research into a phenomenon "upon whose importance everybody seems to be agreed, but whose nature and presence have been little studied" (Alderson & Wall, 1993, p. 115). To that end, however, we will first review the existing literature on the notion of washback to see what proposals have been made thus far with respect to its definition, how it works, what elements contribute to its occurrence, and how it can be measured.

## *3.2 Defining washback*

A quick glance at the literature on language teaching and testing reveals that in spite of the terminological consistency, there is a considerable variation in authors' conceptions of washback. While some authors consider tests as having nothing but negative consequences for teaching methodology and syllabus content (Wiseman, 1961), others look at tests more positively as potential instruments for educational reform (Pearson, 1988). Extremist views have also been expressed with respect to the significance of a positive washback effect for test validity (Morrow, 1986; Frederiksen & Collins, 1989).

Originally, the influence of tests over curriculum and teaching was seen as a negative one, and the term *washback* was synonymous with the harmful effects of tests, the most prominent of which are the coaching classes where the main objective is practicing test techniques rather than language skills (Vernon, 1956; Wiseman, 1961). An implication of this is that a good test is the one that has no influence on class activities, what Davies calls "an obedient servant" (1968, p. 7) that follows its leader, namely, the syllabus and teaching (Davies, 1985). Davies, nevertheless, acknowledges that innovative testing can result in a syllabus change or a totally new syllabus (p. 8). So, if it is true that tests can, in effect, be used as a means towards the betterment of the syllabus, we have to

accept the reality of washback. Some other writers (Swain, 1985; Alderson, 1986; Pearson, 1988) see washback in this more positive way and consider it as having important implications for the curriculum. Pearson (1988), for example, defines the relationship between a good test and the classroom activities as mutual, in that the test serves as an instrument for teaching-learning activities while at the same time the teaching-learning tasks contribute to the testing purposes.

Considering positive washback effect as a starting point for the design of every public examination, Morrow (1986) argues that a measure of how far the intended washback effect was actually being met in practice is the primary validity criterion for these tests. He, therefore, suggests that a new kind of validity, namely "washback validity," has to be added to the list of validity types. He does not, however, specify how this new validity type relates to the issue of construct validity; in other words, if according to him "washback validity" is the primary validity criterion for a test, should a test be considered as not valid if the intended washback effects are not met? What if unintended effects also happen; can we still consider the test as valid? Moreover, Morrow states his uncertainty about how washback can be measured although he suggests class observations as an effective method for assessing the tests' effects. A similar concept, "systemic validity" has also been introduced into the literature by Frederiksen & Collins (1989). They define it as "...one that induces in the education system curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure" (p. 27). Yet, there are important pieces of evidence (as discussed in Chapter Two), other than the test's consequences, that determine the validity of a test. There are also other factors, other than the test itself, that might prevent the intended washback from happening, therefore, a lack of desired washback effect does not and cannot by itself

render a test as invalid, unless there is sufficient evidence that such a phenomenon (lack of washback) is a direct result of the test and its lack of construct validity (Messick, 1996).

However, the literature dealing exclusively with the concept and the issues related to it, consider the authority of tests and the significance of the decisions made on the basis of the tests' results for the future of both students and teachers as one of the most important factors affecting the relationship between testing and teaching. In their comprehensive treatment of the topic, Alderson & Wall (1993) define washback as the influence of testing on teaching and learning so that teachers and learners "do things *they would not necessarily otherwise do* because of the test" (p. 117). Also, Buck (1988), Hughes (1989), and Shohamy (1992) all refer to this element as potentially responsible for both positive and negative consequences attributed to the test influence:

> 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 washback effect can be either beneficial or harmful (Buck, 1988, p. 17).

> If a test is regarded as important, then preparation for it can come to dominate all teaching and learning activities. And if the test content and testing techniques are at variance with the objectives of the course, then there is likely to be harmful backwash [the term is synonymous to washback] (Hughes, 1989, p. 1).

> Central agencies and decision makers, aware of the authoritative power of external test, have often used (or abused) external tests to impose new curricula, new textbooks, and new teaching methods. ... The use of

> external tests as a device for creating impact on the educational process is often referred to as the 'washback effect' or 'measurement-driven instruction' (Shohamy, 1992, p. 513).

In other words, considering the students' unavoidable concern with the test content, if the activities required by the test task(s) are congruent with the course objectives, it is very likely that the test gives the students and teachers an appreciation of what they are expected to gain by the end of the year. As such, the test impact would be desirable in that it can "lead" the teaching and learning activities into the right direction. In the same way, it can be argued that tests can "mislead" class activities by failing to ensure that all course objectives are properly met by the test. Either way, the deterministic power of the tests and, therefore, the existence of washback effect is evident. The presence of washback, however, does not automatically imply an imminent change in teaching and learning since for such a change to happen, other educational and contextual factors should also be taken into consideration (this point will be further discussed later in this chapter).

On the basis of the discussion so far, I will thus conceptualize the washback phenomenon in terms of the following generalizations: (i) the term washback is used to refer to the influence of testing on teaching and learning, (ii) this influence can be negative and/or positive, (iii) whether or not a test is valid does not necessarily depend on its positive/negative washback effect (however, washback, as an instance of the consequential aspect of validity is related to the validity of a test), and (iv) the key issue in a washback hypothesis is the centrality of the test, its task(s) and objectives.

But if washback in this sense really exists, the question is how it operates and what the processes most likely to bring it about are. In the following section, I am going to discuss this matter in some detail.

### *3.3 The washback mechanism*

A very great importance is attached to the mechanisms through which washback operates because any research into this phenomenon has to take such processes into consideration. However, very few proposals have been made with respect to the type and organization of the factors interacting with the test, and each other, to bring about beneficial washback. Even those few are based on pure speculation and not backed by research findings.

In his unpublished paper (cited in Bailey, 1996), Hughes (1993) suggests that a basic model of washback can be constructed on the basis of a distinction between three different elements in teaching and learning. Referring to them as "participants" (including students, teachers, material developers and publishers), "process" (including actions taken by participants which may eventually lead to learning), and "product" (including the outcome and quality of learning), he maintains that by affecting the perceptions and attitudes of the participants, a test can in turn affect the process and product of learning thus promoting the desired effects. Alderson & Wall (1993, pp. 120-21), on the other hand, in an attempt to capture the essence of this phenomenon, suggest the following hypotheses, ranging from the most general to more specific ones:

(1) *A test will influence teaching.*
(2) *A test will influence learning.*
(3) *A test will influence **what** teachers teach.*

(4) *A test will influence **how** teachers teach.*

(5) *A test will influence **what** learners learn.*

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

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

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

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

(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.*

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

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

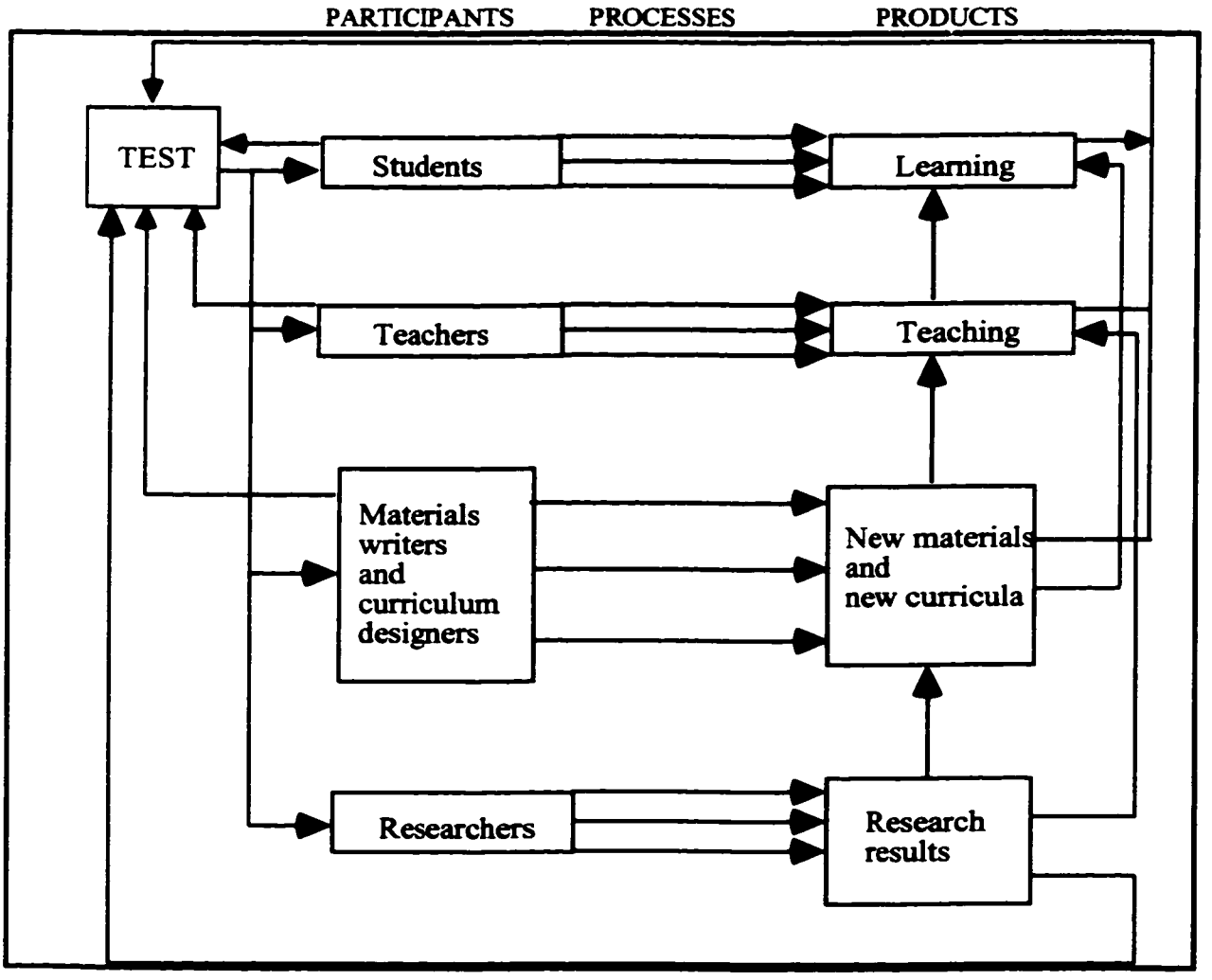
(15) *Tests will have washback effects for **some** learners and **some** teachers, but **not** for others.*

Given the scarcity of research in this area, Alderson & Wall do not reject the possibility of the "Washback Hypothesis" being even more complex than what they have put forward. They, therefore, suggest that any research into washback should first clarify the definition and the scope of the term, specify the nature and predicted effects of the test, and take into account the context in which the test is used as well as the decisions made on its basis. Besides, they emphasize that such research should also consider the research findings in the areas of performance and motivation as well as that of educational innovation and change.

Drawing upon the ideas proposed by Hughes (1993) and Alderson & Wall (1993), Bailey (1996) introduces a basic model of washback (Figure 3.1) representing Hughes' three major categories: participants, processes, and products. Her model illustrates the tests directly affecting the participants (i.e., students, teachers, materials writers and curriculum designers, and researchers) who, in turn, engage in the processes that will

eventually lead to the products (i.e., learning, teaching, new materials and new curricula, and research results). The model also allows the possible influences on the test from the participants:

**Figure 3.1: A Basic Model of Washback[1](Bailey, 1996, p. 264)**
Reprinted with permission




[1] Fine lines in this figure are in fact dotted in the original document.

Besides, Bailey distinguishes between what she calls "washback to the learners" and "washback to the programme" using the former to refer to the "effects of test-derived information provided to the test-takers and having a direct impact on them" (p. 263) and the latter to "the results of test-derived information provided to teachers, administrators, curriculum developers, counselors, etc." (p. 264). This distinction, according to her, is consistent with Alderson & Wall's hypotheses referred to above in that "washback to the learners" directly addresses hypotheses 2, 5, 6, 8 and 10 in their list, while "washback to the programme" refers to those of 1, 3, 4, 7, 9 and 11.

What is not clear, however, is what exactly the intermediate processes in Bailey's model are and how they lead to the corresponding products. She notes that the processes leading to washback vary widely, and provides general examples of the processes students might be involved in – such as studying vocabulary and grammar rules, practicing items similar in format to those of the test, reading widely in target language, ...etc. – when faced with an important test. She further suggests that depending on which processes are adopted, and whether or not they lead to "actual language development," beneficial or negative washback occurs. However, considering the fact that the function of her model is to illustrate *how* washback works, *processes* inevitably become the most important component of the model, and thus a specification and ordering of such processes is necessary.

A very important factor, not included in these models, is an assessment of the learners' needs as the main consideration in the development of the test tasks and as a factor that both reflects and contributes to the future decisions made on the basis of the test scores. On the assumption that learners enter a program and/or take a test due to an educational or vocational need, and that determination and specification of the learners'

communicative needs are a prerequisite to the contents of the test and syllabus, it is hard to imagine how washback can operate without accounting for such an important factor as learners' needs. Adopting Alderson & Wall's framework, one could, therefore, hypothesize that *tests that are based on an specification of the learners' needs will have washback;* conversely *tests that are not based on an specification of the learners' needs will not have washback.* Likewise, one could add considerations of the learners' needs as well as the future decisions made on the basis of the test scores to both Hughes' and Bailey's models. Such an incorporation would then exclude the "test" as the sole initiator of washback operation (as it is implied by the above models) while at the same time preserving its centrality as the main force behind washback effects.

And finally, although it is clear from the above discussion that similar washback mechanisms are assumed to be at work regardless of negative or positive outcomes, it is obviously the positive washback which is intended and promoted by educationalists. As a result, the use of tests that are more likely to produce beneficial washback is encouraged not only as a means for measuring language skills, but also as a stimulus for learning such abilities. Now the questions are how to put this into practice, what characteristics these tests should have and how they are different from those that do not promote positive washback. In the next section, we are going to see how these important issues have been addressed so far by researchers and what recommendations have been made.

### *3.4 Towards positive washback*

To date, several educators, applied linguists, and practitioners have offered suggestions concerning how to ensure that the effects of tests on teaching are beneficial. These suggestions (not all of which are backed by research) take the form of test design

specifications (Frederiksen & Collins, 1989), underlying principles (Messick, 1996), or guidelines (Hughes, 1989; Khaniya, 1990a; Shohamy, 1992; Bailey, 1996). They differ in scope and intensity depending on the stand that the author(s) take with respect to how much weight should be given to this phenomenon when considering the bigger picture of teaching and learning, that is, the extent to which tests can be used "as levers for change" (Pearson, 1988). Considering this, before going over the factors believed to promote positive washback, we would like to digress briefly to discuss the relationship between positive washback and educational change in some detail, as it helps to clarify some of the ambiguity and complexity surrounding the concept and its investigation.

As briefly discussed in section 3.2, for some researchers the validity of a test depends on whether or not it has positive teaching and learning consequences (Morrow, 1986; Frederiksen & Collins, 1989). This view considers beneficial washback as a factor in bringing about educational change (i.e., curricular, instructional changes as well as learning strategy changes in students). However, a very important point that needs to be taken into consideration is that although tests with positive washback effect have the potential to trigger change, they cannot by themselves be held accountable for global changes in the system since there are other factors such as teachers' knowledge and resourcefulness, students' motivation, availability of teaching aids, budget, etc. that might reinforce or interfere with such changes. Yet, because in the field of language testing, the evidence for washback is primarily drawn from the classroom behaviours of teachers and learners (Alderson & Wall, 1993), *tests* have been the most important area so far emphasized by the specialists in a move towards positive washback. However, as Wall (1996) correctly indicates, what are not properly brought into attention in language testing research on washback, if any, are "references to the settings in which tests are to be

introduced, the resources that are available to support their introduction, and the way that innovations should be managed" (p. 335). So, she turns to the research results in education (Heyneman & Ransom, 1990; Eisemon, 1990; Kellaghan & Greaney, 1992; all cited in Wall, 1996), in general and innovation theory (Fullan, 1991; Goodlad et al., 1970; both cited in Wall, 1996), and applied linguistics (Markee, 1993; Stroller, 1994; Smith, 1989; all cited in Wall, 1996) in particular, in an attempt to pinpoint the factors that might affect the type of influence that tests have on teaching and explain why such factors occur and what the relationship between them is. Among these factors that can constrain classroom innovation are cultural, political, administrative, educational and institutional factors, only some of which are directly related to education (Kennedy, 1988). These factors come into play especially when we have to deal with large scale changes at the regional or national level rather than on a relatively small scale. Still, the deterministic interpretation of beneficial/negative effects of tests is unrealistic and simplistic. In the same way, the validity of a test should not be judged solely on the basis of its positive/negative effects since good tests might have poor effects because of educational factors other than the quality of the test (Messick, 1996). Therefore, in this study, while supporting the existence and intuitive acceptability of tests' effects upon classroom activities and outcomes (the extent and degree of which is yet to be empirically examined), we will distinguish between the notions of washback and change beyond the classroom. While the former has the potential to bring about change, it does not guarantee it, due to other forces operative within and outside the educational realm.

Returning to the main point – how to work for positive washback – much of the current work in communicative language teaching and testing ties positive washback to communicative methods of language testing. Swain (1984) suggests "work for washback"

as one of the four general principles relevant to the design of communicative language tests of speaking and writing. In other words, test developers should make positive washback their primary goal by involving teachers in the design, implementation, administration and scoring of the tests. Frederiksen & Collins (1989), referring to the positive influence of tests on teaching as systemic validity, also analyze general characteristics of a testing system that can promote or impede such validity. According to them, such characteristics are the tests' directness and degree of subjectivity. They define "directness" as directly evaluating the cognitive skill in question "as it is expressed in the performance of some extended task" (p. 28). They argue that direct measures of skills will positively influence teaching and learning since the instruction and learning strategies that will lead to a higher score in such tests are basically the same as improved skill and performance. In indirect tests, where an abstract construct is measured using techniques that measure more concrete features of performance (e.g., measuring reading comprehension through vocabulary tests), class activities will be directed towards the skills represented on the tests rather than the construct itself.

The second characteristic Frederiksen & Collins talk about, "the degree of subjectivity," refers to the degree to which the scores are determined by judgment rather than simple algorithmic methods. Since scoring subjective tests requires judgment, analysis and reflection on the part of the raters, such tests do not restrict the examinee to choose or provide exact responses and allow more extended responses to the test items. This, in turn, will lead to the adoption of teaching and learning strategies – by teachers and learners – aimed at problematic contexts while objective tests "...emphasize low-level skills, factual knowledge, memorization of procedures, and isolated skills, and these are aspects of performance that correlate with but do not constitute the flexible, high level

skills needed for generating arguments and constructing solutions to problems" (p. 29). They, therefore, argue that the adoption of direct subjective assessment devices is an effective way of increasing the systemic validity of a testing system. Their plan for designing such a testing system has three major aspects: (a) *Components of the testing system* including a set of tasks, primary traits for each task and subprocess, a library of exemplars, and a training system for scoring tests; (b) *Standards* such as directness, scope, reliability, and transparency; and finally (c) *Methods for fostering improvement on the test* consisting of practice in self-assessment, repeated testing, feedback on test performance, and multiple levels of success. However, as one might have already noticed, some of the procedures listed above, such as direct testing of abilities, reliable scoring of test performance, training raters, etc. are costly procedures – in terms of both time and money – that might render a test as impractical. Frederiksen & Collins (1986), however, do not address this issue directly.

Hughes (1989), however, does take this factor of practicality into consideration and still votes in favour of tests with positive washback effects. He pleads for more consideration to be given to the consequences of *not* using such tests:

> Before we decide that we cannot afford to test in a way that will promote beneficial backwash, we have to ask ourselves a question: what will be the cost of *not* achieving beneficial backwash? When we compare the cost of the test with the waste of effort and time on the part of teachers and students in activities quite inappropriate to their true learning goals ... we are likely to decide that we cannot afford *not* to introduce a test with a powerful beneficial backwash effect (p. 47).

He encourages test-developers to work towards achieving such an effect by using the following guidelines (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 tests on objectives
(6) Ensure test is known and understood by students and teachers
(7) Where necessary, provide assistance to teachers

Similar recommendations are made by Shohamy (1992) in her proposed model for testing foreign language learning. Her model particularly emphasizes the need for providing diagnostic information addressing a variety of dimensions rather than reporting one general score. She also encourages positive washback by requiring test-developers to make their tests communicative by focussing on more direct and authentic language tasks. Similarly, Bailey (1996) notes that one main difference between traditional and communicative language tests is the latter's emphasis on promoting positive washback.

Messick (1996) also probes the characteristics of the communicative tests, namely directness and authenticity, as properties most likely to produce washback. Emphasizing the relationship between these properties and the validity of the test, Messick considers them as safeguards against two major threats to a test's validity: "construct under-representation" which jeopardizes authenticity by leaving out important features of the construct to be measured, and "construct-irrelevant variance" that endangers directness by the inclusion of features irrelevant to the construct being measured.

As it is now apparent from our review, almost every work on washback considers communicative testing and issues of authenticity, directness and detailed score-reporting as factors facilitating positive washback effects. However, the phenomenon has thus far only rarely been investigated with respect to these and other issues involved. This situation might have been caused partly by the multi-facetedness of the concept itself, and partly by the methodological complexity involved in such an investigation. For this reason, before reviewing the few studies on washback in the next section, we will turn to the question of how washback can be measured.

### *3.5 Measuring washback*

A problem researchers face when investigating washback is its apparent association with other educational and contextual factors that make it difficult to clearly identify, isolate and measure this phenomenon in terms of common experimental practices. Because washback is, by definition, intermingled with the activities going on in the language classroom, a study of it inevitably involves teachers and learners. Realizing this, Morrow (1986) – insisting on the extreme importance of washback effect as a validity factor – states that although he is not sure how washback can be measured, one thing he is clear about is that researchers have to go into the classroom and observe the influence of tests in practice. This position is also maintained by Alderson & Wall (1993, p. 127) who emphasize the "need to look closely at classroom events in particular, in order to see whether what teachers and learners say they do is reflected in their behaviour" (see also Messick, 1996). They also suggest that the above observation be triangulated by individuals' own perceptions of events going on inside and outside the classroom.

It thus appears that for studying the washback effects of tests, one has to adopt a qualitative approach in addition to or in lieu of the quantitative research methods in order to be able to control all other variables that might interfere or interact with the phenomenon under investigation. This can be accomplished through questionnaires or interviews held with teachers, learners, and/or parties involved in the design, development, and administration of the test.

## *3.6 Previous studies in washback*

There have so far been only a few projects in language education that have tried to empirically study washback and establish what it really is or how it works. Among them are: Wesdorp (1982), Hughes (1988), Khaniya (1990a,b), Wall & Alderson (1993), Alderson & Hamp-Lyons (1996), Shohamy et al. (1996), and Cheng (1997).

### *Wesdorp (1982)*

Alderson & Wall (1993) report on unpublished research done by Wesdorp (1982, cited in Alderson & Wall, 1993) in the Netherlands investigating the validity of objections to the use of multiple-choice tests for the assessment of both first and foreign language education. The results did not support the assumed negative washback effects. One of the assumptions, for example, was that the skills that could not be tested by multiple-choice questions would not be taught any more in primary schools. But a comparison of the essays written before the introduction of multiple-choice and twelve years after that showed no differences in quality. Differences between the teachers' activities in schools with and without a multiple-choice final test were also insignificant. The results did not show any changes in the students' study habits either. On the whole, the study revealed much less negative washback than had originally been assumed. However, it is not clear

what kind of tests had been in effect before the introduction of multiple-choice tests and how different the tests measuring first and second language education were. It could be that the old test methods (e.g., direct/indirectness, discrete-point/integrative approach) and content were so similar to those of multiple-choice tests that even after the introduction of the new technique teachers and learners didn't feel any need to change their attitudes towards the tests.

***Hughes (1988)***

Hughes (1988, 1989), on the other hand, describes a project conducted in a non-English speaking country, at a Turkish English-medium university. Before the study started, undergraduate students used to enter academic programs after spending a year of intensive English study, yet they demonstrated a very low level of English proficiency. As a result, the university had decided to establish a screening device to determine which students could continue with their studies and which students would have to leave the university. A new test was developed based on the English study skills needs of freshman students (e.g., reading, note-taking, etc.) which included tasks similar to those they would have to perform as undergraduates. Hughes (1989) reports that the introduction of this test in place of the old multiple choice test immediately affected teaching:

> ... the syllabus was redesigned, new books were chosen, classes were conducted differently. ... the students reached a much higher standard in English than had ever been achieved in the university's history. This is a case of *beneficial* backwash (p. 2).

According to Hughes (1988), the result of such a change was a rise in the standards of English in that university so that 83% of students achieved the minimum acceptable Michigan Score (compared with less than 50% in the past).

Nevertheless, the question which remains unanswered here is that if teachers' and learners' activities in classrooms changed as a result of the introduction of the new test, whose content was totally different from the Michigan test, how could it still result in a better performance on the Michigan Test? If the increased efficiency in the Michigan test was not originally a goal of the new proficiency test, this outcome can be attributed to construct irrelevant variance which is a source of invalidity. Besides, we are not given any details concerning what exactly changed in the teachers' and students' activities which led to an increase in English proficiency. In other words, how did the test produce the desirable washback effect?

***Khaniya (1990a,b)***

In a study in Nepal, Khaniya (1990a,b) attempted to study washback by designing a new communicative English language proficiency test and comparing it with the traditional SLC (School Leaving Certificate). According to Khaniya, the SLC has important consequences for the future of the students since it is a factor in the selection of university and job candidates. Consequently, students, teachers and parents are very much concerned with its results. As Khaniya describes, SLC requires students to memorize texts and answers to questions since many of the test questions and texts are taken directly from the textbooks. In such a situation, the exam would definitely have some sort of control over the course, but he does not explain how teachers actually teach to the exam, what and how students learn, and so on. He gave his new test to three different groups of students at the beginning and at the end of grade 10 when students are preparing for the

SLC. Two of these groups were doing their studies in English-medium schools, one emphasizing skills, the other the SLC exam. The third group, however, was enrolled in a Nepalese-medium school. Based on final results, Khaniya reports that while the difference between students' performance (in English-medium schools) before the introduction of the new test was not significant, at the end of the year those with an emphasis on skills improved their performance on the new exam while the students whose program emphasized SLC performed poorly. Khaniya claims that this is because of the SLC exam-oriented teaching going on in exam-emphasizing schools, due to the negative washback of the SLC test. He argues that the fact that the third group of students (in Nepalese-medium schools) also performed poorly at the end of the year further supports this claim.

Khaniya, however, does not clarify what his conception of washback is and what processes in language teaching and learning he is looking at when he makes his claims regarding the positive washback influence that the new test is exerting upon English language classes. He does not narrow down his experimental methods in such a way that the differences between the test results can be definitely attributed to the beneficial effects of the new test and not to the negative consequences of the old one. Neither does he take a clear theoretical stand with respect to the relationship between test validity and washback effect. He seems to have assumed positive washback as an indication of test validity but does not specify how in this case washback benefits from or contributes to other elements involved in the validity of a test.

***Wall & Alderson (1993)***

To investigate the impact of a new O-level exam on English language teaching, Wall & Alderson (1993 ) conducted a two-year longitudinal observational study of O-level English classes in Sri Lanka. According to the authors, the educational context in which

the study was conducted requires high school students at the end of their eleventh year of studies to take the O-level examination, whose results determine whether or not they will be allowed to enter higher education or be eligible for the available good jobs. With the old test in effect, even those few students who passed the test were not proficient enough for the situations in which they were expected to use their English language knowledge. As a result, in an attempt to increase the efficiency of the O-level classes, a new series of textbooks with an emphasis on reading, writing and oral skills were introduced. The new test was, therefore, intended to promote the skills presented by the textbook.

Unlike the previous studies which were based on questionnaires, interview results and test scores, this one adopted a qualitative approach to data gathering and, in addition to what teachers reported, based its results on direct observation of the classrooms. They conclude that the impact of the new test is less pervasive than had been expected. Their observations revealed that although the exam had affected the content of language lessons as well as the content of teaching, no changes were evident in teaching methodology. They thus suggest that tests might have an impact on what teachers teach but not on how they teach. The authors also refer to the fact that their test, which originally covered all four language skills, had to be modified into a test measuring only the written skills because of what they call practical and political reasons. They then conclude that the test is only one of the factors operative in an educational system and cannot by itself be held responsible for educational reform.

Wall & Alderson (1993) do not make any explicit reference to the learners' behaviour in their report. Classroom observations as well as questionnaires and interviews seem to concentrate on the teacher's behaviour in classes, but not on the learners' behaviour or on activities inside and outside the classroom. Also, there is no quantitative

assessment of the learners' performance on the old test versus the new one. Such information could not only give us an idea about the new test's impact on learners' achievement but it could also shed some light on how effective the teachers' methodology was. Besides, considering the fact that the test for this study was created as a reaction to and with the intention of reinforcing the new series of teaching materials that had already been in effect, the test itself was influenced by the material, thus indirectly reflecting the communicative needs of the population involved. It could be that if the test had been developed first-hand based on an assessment of the learners' needs and teachers' insights, its impact on teaching methodology would be different.

***Alderson & Hamp-Lyons (1996)***

This is another qualitative study investigating the validity of common claims that the TOEFL has an undesirable influence on language teaching. Their data consist of interviews with students and teachers as well as the observations made of four classes taught by two different teachers teaching one TOEFL preparation and one non-TOEFL preparation class each. The idea behind class observations was to separate test effects from teacher-style effects.

In their analysis of classroom data, the authors find substantial differences between TOEFL and non-TOEFL classes, regardless of who is teaching the course. For instance, in TOEFL classes, they report more occurrences of test-taking, teacher talk, the use of metalanguage, as opposed to fewer opportunities for pair work, laughter, and turn taking. Although Alderson & Hamp-Lyons (1996) are very cautious in attributing these differences to the nature of the test, they admit that:

> While the two teachers we observed were of very different personalities and teaching styles, there do seem to be some patterns of difference between TOEFL and non-TOEFL classes that are common to both teachers ... . It may be that these differences are due, or partly due, to the test....Certainly in the [TOEFL] classes we saw there was less student questioning, less time spent in student-student interaction and student-teacher interaction... (pp. 290-293).

Although the study did not aim at finding out whether TOEFL preparation courses were effective, the authors were surprised by the fact that *none* of the teachers interviewed actually claimed that TOEFL preparations classes were effective in raising students' scores in TOEFL. Nor had the institute that offered the course made any effort to gather data regarding the students' TOEFL scores before and after they took the course.

Implicit in Alderson and Hamp-Lyons' (1996) conclusions is that TOEFL washback in fact exists:

> Our study shows clearly that the TOEFL affects both *what* and *how* teachers teach, but the effect is not the same in degree or in kind from teacher to teacher ... (p. 295).

Nevertheless, they speculate that

> ... the simple difference of TOEFL versus non-TOEFL teaching does not explain *why* they teach the way they do. ... It is tempting to conclude that the TOEFL alone does not cause washback, but that it is the administration (who decree larger classes), materials writers (who provide no guidance to teachers on how to teach) and teachers themselves (who give little sign of

> thinking about how best to teach TOEFL) who cause the washback we have observed (p. 295).

While the factors that they list above are very likely to contribute in general to the washback effect exerted on the teachers and their teaching methodology, it is difficult to see how this could be a factor in this specific study. One of the teachers under observation was actually the material developer for the TOEFL preparation classes, used as an in-house book by the institute, and had been teaching TOEFL courses for over 17 years; the other teacher had in the past taught the TOEFL preparation class only once and had never taught this material before. So we can see that the teacher factor and material-writer factor referred to above are under control at least for one of the teachers participating in the study. However, given the fact that both teachers still performed homogeneously in spite of this in that they both showed similar methodological adjustments when teaching TOEFL courses, one tends to assign a more fundamental role to the test in producing washback.

***Shohamy, Donitsa-Schmidt & Ferman (1996)***

This study follows up Shohamy's attempt in 1993 to study the impact of two national tests of Arabic as a Second Language (ASL) and English as a Foreign Language (EFL) in Israel. Results then showed that the ASL test had affected teaching and learning activities by: having teachers stop teaching new material and review the old material, class textbooks being replaced with worksheets identical to old tests, class activities becoming more testlike, review sessions being added to the regular class hours, class sessions having a tense atmosphere, and teachers and students being eager to master the material. The EFL test, on the other hand, as part of the national matriculation examination, resulted in

teachers spending much of the class time on oral language skills using tasks similar to those of the test. Shohamy is, therefore, convinced that at that time both tests had some sort of influence on teaching because they brought skills that had not been taught before to the attention of the teachers.

Because such tests have become the routine practice in Israeli schools, the authors in this study seek to know if such an effect is still persistent two years later. As the main instruments of data gathering, they used questionnaires and interviews. Compared with the 1993 study, the results of this study revealed a totally different impact from the tests. They report that "the impact of the ASL test has decreased over the years to the point where it has no effect" (p. 312). Interviews with teachers, students and authorities indicated that teachers did not prepare themselves for class activities as they used to do in the past and students (and their parents) had a very low awareness of the existence and content of the test. Whereas teachers did not think of the tests as having important consequences, students considered it as eventually affecting their knowledge of Arabic and thus their future success in their studies. However, the low prestige of this subject remained unchanged. Moreover, both students and teachers rated the test as unimportant and of poor quality, one that lacks both validity and reliability. The EFL test, on the other hand, which had undergone some modifications over the years, proved a significant increase in test washback. Data revealed that the test had affected both the content and the methodology of teaching. Teachers admitted that had it not been for the test they would not have allocated so much time to teaching oral skills. Also, both students and teachers reported a high level of anxiety caused by the test. Here too, all teachers, in spite of attributing high status value to the test, reported their dissatisfaction with the test, considering it as neither valid nor reliable.

Interestingly, the authors report a discrepancy between how bureaucrats, on the one hand, and students and teachers, on the other hand, view the effects of the tests. While students and teachers considered tests as poor in quality, unreliable, and invalid, both the Arabic and the English inspectors were satisfied with the role of the tests in the educational system. The authors believe that the reason for this is that policy-makers consider the test as an "effective tool for controlling the educational system and therefore use it intensively in prescribing the behaviour of teachers and students" (p. 314).

They conclude, however, that for both tests the impact observed in this study is different from that of the 1993 study because of the presence of several factors that did not exist then. Among these factors is the lower status of Arabic versus the higher status of English in Israel. The authors also report that at the time of the study, the Arabic test was not a high-stakes test and its results were not used for any decision-making purposes, so the students did not have any reason to fear the test; on the other hand, the EFL test was a high-stakes test creating anxiety and fear in both teachers and students. They also indirectly relate the EFL test format (oral) to the test impact because it might increase anxiety. On the other hand, they consider the multi-skill format of the ASL test as a factor that might make test-takers more confident since their high proficiency in one skill can compensate for their lack of proficiency in another skill. Based on all this evidence, they suggest that the impact from a test can change over time depending on its nature, purpose, and other characteristics.

Based on an interpretation of the results, both tests should have had a negative impact on teaching and learning, each in a different way. The authors don't specify what kind of washback – positive or negative – they are referring to when they summarize their findings: "while the washback effect of the ASL test has significantly decreased over the

years the impact of the EFL test has increased" (p. 314). If, based on the data reported, (i) all EFL teachers have a negative attitude towards the EFL test and view it as neither valid nor reliable, (ii) the test creates an atmosphere of fear and anxiety, (iii) more students (46% vs. 34%) believe that the test has had little or no effect at all on their learning, (iv) lower-level teachers believe they can engage their students in more creative oral activities because they focus less on the skills tested by the exam, it is difficult to imagine how the EFL test's positive washback could have increased. Unfortunately, we don't have any other type of evidence (such as data gathered from direct observation of classrooms, or a quantitative analysis of test results) to be able to determine whether this increase in washback over the years has in fact affected the learners' command of English and the extent to which the fear and anxiety caused by changes in the test promoted or impeded learning.

***Cheng (1997)***

This study, conducted in Hong Kong, investigated how the revised Hong Kong Certificate of Education Examination (HKCEE) affected teaching English in Hong Kong secondary schools. The HKCEE is a high stake public examination which underwent major changes in 1993 in accordance with the objectives of the task-based Target Oriented Curriculum in Hong Kong. The author, however, states that despite the changes, the exam remained norm-referenced in that its scores are used primarily for the purpose of selection rather than education. In an attempt to observe how the educational system in Hong Kong would react to the new exam, Cheng (1997) used questionnaires, interviews, and classroom observations as her primary method of data collection.

The results of the study showed that the changes in the teaching materials had been more intensive than other areas. She attributed this change to the "highly commercial

nature of Hong Kong society" (p. 38). The effects on the teachers' methodology, on the other hand, were found to take place "... slowly and reluctantly and with difficulties... caused by the constraints imposed upon teaching and teachers in ....schools" (p. 38). The study does not unfortunately make a clear distinction between the effects on the material choice/development caused by the test and those brought about by non-test-related factors such as the commercial purposes of the textbook publishers. In the latter case, which based on Cheng's conclusions seems to be a very likely situation, the intensive washback cannot be attributed to the nature of the revised HKCEE or any other test for that matter. In short, it is not clear how "intensive washback" as defined by Cheng (1997) could have taken place in this area if teachers' and/or schools' choice of teaching content had been the result of their reliance on the commercial textbooks developed by publishers rather than a result of their awareness of the aims and objectives of the test.

### *3.7 The washback hypothesis proposed in this study*

The studies reviewed above adopt different research methods for examining the washback effect of high-stakes tests in mostly EFL contexts. There are, however, important theoretical considerations not taken into account by these studies. First, none of these experiments are based on an a priori theory specifying the scope and design of a washback study. Second, as extensively discussed in Chapter Two, a matter of significance in a study of washback is where the study stands with respect to the important theoretical relationship between washback and validity. Other than Wall & Alderson (1993), none of the empirical studies have so far clearly and explicitly addressed this issue. Besides, these studies have not looked into all possible areas influenced by test washback. Hughes (1989) and Khania (1990a) have studied the influence of tests on students' learning outcomes

while the rest of the studies have only focussed on teaching contents and methodology. Furthermore, with the exception of these same two studies that are based on the tests developed particularly for the purpose of the study, other experiments are mostly based on already existing tests which, according to the researchers themselves and/or teachers using them, are not valid for the purposes to which they are put. Given the importance of the testing instruments in a study of washback, this obviously affects the validity of the results and consequently the generalizations made on the basis of them. Finally, it is important to provide the reader with a clear idea of what the study's conception of washback is and how the evidence distinguishes between *washback from the test* and the impact from other non-educational factors imperative in the educational system *through the test*. This is a very important point since it is through such a distinction that we can avoid overcomplicating the notion of washback and drawing ambiguous conclusions in our studies. As Brennan (1998, p. 9) correctly argues, although "well-constructed assessments can provide remarkably good and useful snapshots of student behaviour, ... assessment procedures per se cannot reflect all that goes on in schools." But if we consider learning as the ultimate goal of teaching, and testing as a powerful means for achieving that goal, then we have to accept the reality of tests affecting teaching and learning activities. This will inevitably result in a reliance on the test scores as indicators of the extent to which learners possess the ability being tested which, in turn, underlies the decisions made about their learning (e.g., placement, achievement, admission, employment, etc.).

Given this, and to avoid further complicating an already complex phenomenon before embarking on our research into washback, we will first clarify the definition of washback based on which the scope of the study will be determined. Four generalizations form the conceptualization of washback as reflected in this study:

(i) the term *washback* is used to refer to the influence of testing on teaching and learning

(ii) this influence can be negative and/or positive

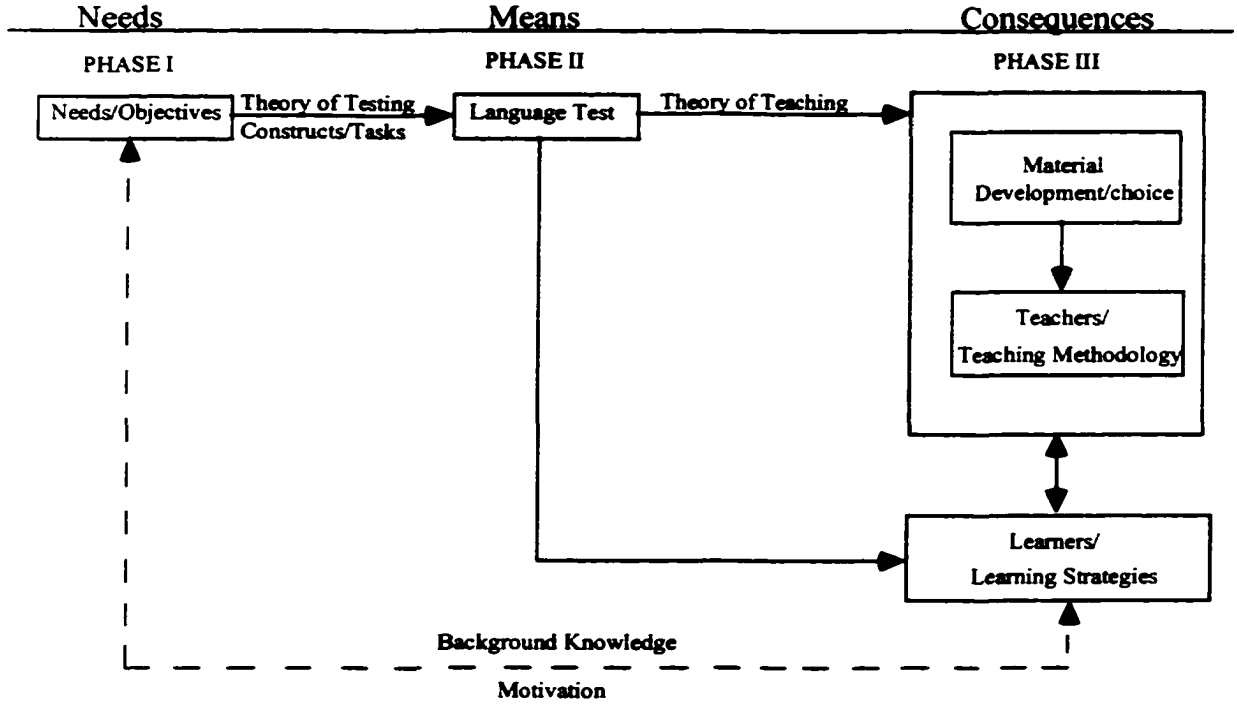
(iii) whether or not test results are valid does not necessarily depend on its positive/negative washback effect, but washback as an instance of the consequential aspect of validity is related to the validity of a test, and

(iv) the key issue in a washback hypothesis is the centrality of the test, its task(s) and objectives.

Once the phenomenon has been defined, it can be hypothesized that *a test can have a positive effect on what teachers teach, how teachers teach, what learners learn and how learners learn if its tasks and contents address the educational and functional needs of the population and the context it is intended for, and are informed by such variables as learners' background knowledge, motivations and attitudes.* This implies that in a study of washback, we have to take into consideration factors other than the test itself which contribute to or hinder the occurrence of washback. Because of this and based on our discussion of how washback works in section 3.3 above, the scheme presented in Figure 3.2 can be taken as a theoretical basis for a study of washback.

It not only clarifies the limits of the study, but also systematically presents the areas and participants that are inevitably affected in the process and are thus considered as reliable sources of evidence for any study of washback.

**Figure 3.2:**
**General Scheme for a Theory of Washback**




The model illustrates two major lines of connection with respect to the language test which should be pursued in a theory of washback: (i) the factors which directly or indirectly affect the nature of the test, its development and implementation; and (ii) the effects that the test will exert on the areas most susceptible to change. Although it is the second line which will eventually lead to the washback effect, such an outcome would be both theoretically and practically impossible without taking the first line of connection into consideration. Thus, if a test is going to have beneficial washback, its objectives should be based on a sound analysis of the "needs" of the population for whom the test is designed. It should be "learner-based." Once the objectives are clear, it is necessary to develop the test with respect to a theoretical framework – in conformity with the test objectives – so

that the same theoretical line of thought can be followed in all future decisions made with respect to material development and teaching methodology.

The second line of connection illustrated by the model is the direct effect that a language test can have upon material development, teaching methodology and learning. The first two areas are supposed to be theoretically compatible with the test[2]. As long as this happens, we can be confident that the objectives of the materials and teaching are the same as those of the test. However, the desirable learning effect might not be brought about even for learners with the same needs and in the same educational context, if the learners' experience with the target language and their topical knowledge are not taken into consideration in the development of the test. Therefore, in addition to the shared motivation – which originates from needs – the learners' background knowledge of the language to be learned and the topics to be covered by the test should be considered an integral aspect of a theory of washback.

In summary, the washback theory presented in this model, hypothesizes that a test could positively influence the material used in a classroom, the teaching activities, the methodology teachers adopt, and the learning brought about in learners, if the needs of the learners and such attributing factors as learners' motivation and experience with the target language and subject matter are taken into consideration.

With this background in mind, the next chapter will concentrate on one last theoretical consideration, performance testing. A prevailing assumption with respect to

---

[2] The arrows in the model signify the direct relationship between the components of the model in terms of the sources of input for each area. Solid arrows, however, reflect the order through which washback works.

performance-based assessments is that they encourage on the part of teachers and learners the use of techniques and strategies that are known to underlie language abilities responsible for successful communication. Such tests are potential instruments for creating positive washback.

# CHAPTER FOUR
# PERFORMANCE TESTING

## *4.1 Introduction*

The current interest in the consequences of test use resulting from developments in validity theory justifies a move toward performance assessment. This is because performance testing is generally believed to be associated with authentic, direct assessment of competence leading to positive consequences for teaching and learning. An implication of this might be that other indirect forms of assessment (e.g., paper-and-pencil multiple choice or short answer tests measuring one point at a time) are inauthentic. Such generalizations, however, do not always hold mainly for two reasons. First, considering the purpose(s) to which they are put, indirect assessments are not necessarily inauthentic. The relevant question about authenticity is therefore "authentic to what?" (Arter & Spandel, 1992, p. 38). Second, the idea of promoting performance assessments solely because they avoid isolated skills or sub-skills is equally misleading, "in that component skills and abstract problems have a legitimate place in pedagogy" (Messick, 1994, p. 13).

Then, on what reliable basis does the credibility of performance assessments lie? Messick (1994, p. 14) argues that:

> ... other forms of assessment are more constructively characterized not as inauthentic ... but as authentic to other criteria and other purposes. ... authenticity and directness map, respectively, into two familiar tenets of construct validity, that is, minimal construct underrepresentation and minimal construct-irrelevant variance.

In other words, performance assessments that are both direct and authentic differ from other types of assessment in that they aim for "complete construct representation."

In the next section of this chapter, we are going to develop this line of argument, namely, the implications of authenticity and directness for the validity of performance tests in general and the consequences of test use in particular. In the following section, characteristic features of performance tests of language ability – such as authenticity, directedness, construct-relatedness, and consequences – will be discussed in the light of a distinction made between the performance assessments evaluating the products and those assessing the constructs. We will then review the major validity criteria so far suggested for performance tests. The general concept of validity is emphasized since, as already discussed in previous chapters, test washback is only an instance of the consequential aspect of construct validity, which, together with other aspects, form the construct validity of a test. So, in order to be able to make precise inferences regarding positive or negative washback effects and to distinguish test-related consequences from the effects of other factors in the educational system regardless of test content and quality, one needs evidence from all available sources of validity.

Having characterized performance assessments of language ability, section 4.3 will focus particularly on the design of the performance tests. The framework proposed by Bachman & Palmer (1996) will be introduced as a model reflecting the main characteristics of the test-taker's ability as well as the test tasks that link test and non-test domains of language use. The framework is to be used for the development of the testing instruments in this study, since it includes potentially all conceptual notions applicable to the performance assessment of language abilities as described in this chapter.

## *4.2 Characterizing performance assessments*

### *4.2.1 Authenticity*

One reason why most educators praise the performance assessment of an individual's ability is that performance tests are more likely to produce an authentic view of what the test taker knows and can do by imposing on the testing situation the same conditions as performance in non-test situations. However, in order to avoid ambiguity and misjudgment, before making any decisions regarding the choice of test tasks, test developers have to address a fundamental question about authenticity, that is, what is indeed meant by authentic test tasks. Are they supposed to authentically reflect what has been going on in the classroom? or be an authentic representation of real life situations? (Arter & Spandel, 1992; Messick, 1994). And in the latter case, can a testing task ever be made as authentic as its counterpart in a real life situation?

Challenging the idea of boundless authenticity in language testing, Spolsky (1985) considers the issue as being unethical and pragmatically inappropriate. He cites the following discussion of illocutionary acts by Searle (1969) which clearly illustrates the problem. Searle states that a question, as a speech act, fulfills the following conditions:

> *Preparatory:* 1) S[peaker] does not know 'the answer', i.e. does not know if the proposition is true, or in the case of the propositional function, does not know the information needed to complete the proposition truly ...
> 2) It is not obvious to both S[peaker] and H[earer] that H[earer] will provide the information at that time without being asked.
> *Sincerity:* S[peaker] wants the information (p. 65, cited in Spolsky, 1985, p. 35-6).

However, the condition of the same question in the context of an examination differs from this in that real life questions are requests, whereas examination questions are in fact requests "that the hearer display knowledge":

> In real questions S[peaker] wants to know (find out) the answer; in exam questions, S[peaker] wants to know if H[earer] knows (p. 65).

Spolsky, therefore, argues that all language tests are in essence inauthentic and abnormal behaviour, unable to engage the test taker in authentic communication, however authentic their tasks might be.

Bachman (1990), on the other hand, distinguishes between two "real-life" (RL) and "interactional/ability" (IA) approaches to authenticity. Authenticity, as defined by the real-life approach, is "the extent to which test tasks replicate 'real-life' language use tasks" (p. 307). The interactional/ability approach, on the other hand, seeks authenticity in the inclusion of those features of language use that are relevant to the interpretation and uses made of test scores with an emphasis on the "interaction between the test taker, the test task, and the testing context" (p. 322). The major difference between the two approaches then is that rather than trying to replicate non-test language use in its entirety, the IA approach emphasizes the need for the recognition and employment of the test tasks that best reflect our knowledge of language abilities and uses. In the IA approach, therefore, one necessarily has to take the test-taker's ability into consideration, whereas in the RL approach, because the ability to perform non-test language tasks is the criterion for authenticity, no distinction is made between ability itself and its behavioural manifestations

in non-test situations (Bachman, 1990)[1]. This difference in the way that RL and IA approaches to authenticity view ability has very important implications for the generalizability of the results (and thus the validity of the test) in that, depending upon what approach the test is based on, it could measure performance per se or use performance as a means for measuring the ability or construct underlying it. So, considering that, generally speaking, performance assessments could be used for both purposes, before discussing the validity criteria of such tests, a distinction has to be made between "assessment of performance per se and performance assessment of competence or other constructs" (Messick, 1994, p. 13).

### *4.2.2 Performances: Means or ends?*

A preliminary issue to be considered when using performance assessments is the place of performances in the whole process of testing. Depending on what kind of test is being used, in what context and for what purpose, decisions have to be made with respect to the relationship between the performances and the meaning of scores. In other words, it has to be clear from the outset whether the scores are going to be based solely on the performance itself or on what the performance is a representation of, namely, the competence. Performances in the former case are ends in themselves, while in the latter case they are just means to other ends, namely, competences or constructs.

As for the assessment of performance as an end, Messick (1994) further

---

[1] RL and IA approaches to authenticity underlie Bachman & Palmer's (1996) definitions of 'authenticity' and 'interactiveness' as two qualities of test usefulness. They define 'authenticity' as a quality of the test that has to do with the degree to which test tasks represent the features of the tasks common to the non-test language use context. 'Interactiveness', on the other hand, accounts for the ways the characteristics of the test-takers interact with those of the test task.

distinguishes between performance and product as follows:

> In some domains such as acting and dancing, the performance and the product are essentially the same thing. ... all there is to evaluate are performances in one form or another. In other domains such as painting or creative writing, there may be so many acceptable variations in process or alternative modes of proceeding that the product is what mainly counts. In still other domains, such as ... chemical experimentation, however, both the performance and the product warrant scrutiny from the outset, because not only is the outcome at issue but so are proper procedures. (p. 14)

In such contexts then, the focus of evaluation is the performance or product, in that they are the ends, not the means of assessment. So the replicability and generalizability of the score meaning are not at issue here. Similarly, the scores obtained in this form of performance assessment are not to be used to make inferences about the competencies underlying the observed behaviour.

In the performance assessment of competences, on the other hand, performances are considered as means (or "vehicles" as Messick, 1994 puts it) for the evaluation of the abilities underlying the test performances. In this case, the replicability and generalizability count since the meaning of the scores is based on the extent to which the results are representative of, consistent with, and generalizable to the tasks representing the broader construct domain. In other words, replicability and generalizability limit the meaning of the assessed construct to the situations and tasks to which it is generalizable and transferable.

Tests of language, however, fall in this latter category since in assessing a language skill, we have to go beyond the quality of the performances and evaluate the test-takers' knowledge or skill domain at issue. In doing so, rather than the simplistic view that

adopting any open-ended unstructured type of task in lieu of the more controlled multiple-choice items would serve the cause, a set of principles tying together the objectives of the test, the construct in question, and the nature of the domain area have to be taken into consideration. Consequently, in performance assessments of language skills, like many other educational tests, constructs or conceptual frameworks representing relevant skills and their component parts should be the primary drive for the design of the test. This way, test tasks will be selected solely on the basis of the nature of the construct and the type of behaviour(s) that reveal such a construct; that is, only those tasks which best elicit those behaviours will be included in the test.

Such "construct-driven" (Messick, 1994) tasks, because they replicate the performances that are typical of real-life challenges, could also enhance the authenticity of the tests even though, for the reasons mentioned in the previous section, they do not simulate all complexities of the real world. In fact, according to Messick, "what is important to simulate are the critical aspects of the criterion situation that elicit those performances from which the focal constructs of knowledge and skill are inferred" (1994, p. 17). Authenticity or the correspondence between the characteristics of the test tasks and those of the non-test language use (Bachman & Palmer, 1996) is, therefore, an important consideration in performance testing, which relates to the construct validity of the test by requiring that nothing important be left out of the definition of the focal construct; i.e., minimal construct under representation (Messick, 1994).

### *4.2.3 Directness*

Authenticity, in the above sense, while being an essential characteristic of performance assessments is not by itself sufficient for the achievement of the complete construct

representation. This is because the test performance is as much a product of the interactions between the individual test-taker and the test task as it is of the relationship between the characteristics of the test task and those of the non-test language use context. Hence, test tasks should be designed in such a way that they not only simulate as closely as possible the corresponding tasks in the real language use contexts but also involve the test takers as directly as possible in the performances whose behavioural manifestations can be used as bases for making inferences with respect to the construct at issue. This latter quality has been referred to in the literature as "directness" (Frederiksen & Collins, 1989; Messick, 1994) or "interactiveness" (Bachman & Palmer, 1996).

Directness is, therefore, a quality of the test tasks and direct tests ideally involve open-ended tasks that allow test-takers to freely express their knowledge of the ability in question. This implies that the more direct the test tasks are, the fewer constraints are imposed on test-takers' performance associated with the focal constructs; i.e., minimal construct-irrelevant variance (Messick, 1994).

However, just like authenticity, achieving absolute directness in performance tests is not attainable. Frederiksen & Collins (1989) define direct tests as tests in which "the cognitive skill ... is directly evaluated as it is expressed in the performance of some extended task" (p. 28). The definition reflects the distinction we just made between performances as targets of measurement and those as means for evaluating underlying constructs, since it implies that it is only through the direct observation of performances that cognitive skills can be evaluated. So, cognitive skills in performance tests are in fact *in*directly observed. Here again we can only come close by choosing the tasks that are direct, because generally speaking "all measurements are indirect in one sense or another" (Guilford, 1936, p. 5; cited in Messick, 1996, p. 244).

In short, from the discussion so far, we can tell that performance tests of language knowledge and skills are essentially construct-centered, so the process of designing such tests should include: (1) determining what complex of knowledge, skills, or other attributes should be assessed, (2) what behaviours/performances reveal those constructs, and (3) what tasks and situations best elicit those behaviours.

Authenticity and directness, however, as two properties of performances and situations, apply to the choice of tasks (step two above). Authentic tasks engage test-takers in realistic situations parallel to those in real-life, ensuring that all key knowledge areas in the focal construct have been covered (minimal construct under-representation), while direct open-ended tasks provide the respondents with the opportunity to freely perform their skills without being restricted by the test methods (minimal construct-irrelevant variance). In other words, once the major facets of the construct under assessment are delineated, and the behaviours representing them are determined, authenticity and directness come into play by requiring the test tasks to measure the whole construct in question, i.e., tasks should be neither too narrow, leaving out important parts of the focal construct, nor too broad, including irrelevant skills.

### *4.2.4 Test consequences*

The two sources of invalidity, construct under-representation and construct irrelevant variance, discussed above are of significance to the consequences of test use since they can distort not only the test-takers' performance on the test task(s) but also the scoring of the task and as a result, the inferences and decisions made on the basis of the test scores. If the test contains tasks requiring skills not related to the construct to be assessed (construct irrelevant variance), some students, despite their knowledge of the areas

measured by the focal construct, might perform poorly just because of the restrictions imposed on the test tasks by the irrelevantly represented skills. Some others, on the other hand, might obtain an overall high score by performing well in those irrelevant areas of knowledge even though their preparedness for the areas of knowledge tested by the focal construct might be questionable. The students' demonstration of the focal skills affected this way might lead to invalidly low/high scores upon which misleading inferences and decisions might be made.

Similarly, negative washback effects in the form of invalidly low/high grades might be caused by assessments leaving out something essential to the main construct (construct under-representation). Besides, students' and teachers' strong desire to obtain good results in the tests might drive the teachers to overemphasize and the students to overpractice the areas that are covered by the test and ignore the under-represented ones. So, in examining the consequential aspect of validity, a primary concern is that negative unintended impacts of the test use should not derive from any source of invalidity such as construct under-representation and construct-irrelevant variance (Messick, 1989).

Construct-driven performance assessments are, therefore, promising in terms of promoting positive washback because they are more likely to control the sources of invalidity, especially construct under-representation and construct-irrelevant variance, by not only basing the whole process of test development on the relevant skills and knowledge areas at issue but also employing tasks that are both authentic and direct in the sense discussed above.

### *4.2.5 Validity criteria for performance assessments: An overview*

Based on the modern views of validity proposed by Cronbach (1988) and Messick (1989), few researchers have suggested expanded validity criteria for performance testing in an attempt to balance the technical and consequential considerations. Among them are Frederiksen & Collins' (1989) principles for the development of systematically valid tests, and Linn, Baker, & Dunbar's (1991) validation criteria for complex performance-based assessments. Later on, in 1994, in an article entitled "The interplay of evidence and consequences in the validation of performance assessments," Messick briefly introduces six relevant aspects of validity. The same categories are presented in his 1996 paper in greater detail and with respect to a discussion of washback in language testing. Bachman & Palmer (1996) also suggest three principles along with six qualities of test usefulness which according to them are the key concepts in answering the question " how useful is this particular test for its intended purposes?" In what follows, the validity categories these works suggest for performance assessments will be reviewed.

As briefly referred to in the previous chapter, Frederiksen & Collins (1989) consider directness along with scope, reliability and transparency as standards for the development and evaluation of performance tests. *Direct* tests are those in which "the cognitive skill that is of interest is directly evaluated as it is expressed in the performance of some extended task" (p. 28), regardless of whether a process, a product or both is being measured. They consider directness as having important implications for learning since indirect measures could mislead the learners by leading their learning efforts towards unnecessary activities geared to the success in tests. Their second standard, *scope*, refers to the test content that should cover all the knowledge area, skills, and strategies that are necessary for the successful performance of the task, otherwise, test takers will only

partially address the task that is required of them. *Reliability* and *transparency*, on the other hand refer to the scoring of the test. According to Frederiksen & Collins, "... the most effective way to obtain reliable scoring that fosters learning is to use primary trait scoring borrowed from the evaluation of writing." (p. 30). They argue that the test must be transparent using clear terminology for judgment so that the learners can assess themselves and others with the same level of reliability as actual evaluators. Authors consider this criterion as important for systematically valid tests.

Linn et al. (1991) provide an even more detailed set of criteria for the validation of performance tests. They too consider direct assessments of performance, with open-ended problems, essays, portfolios of student work, etc., as "authentic" assessments which – in contrast to multiple-choice tests that are basically indicators of other valued performances – involve tasks that are valued in their own right. This justifies the recent move towards performance assessments since the major problem with traditional achievement tests, as the authors rightly indicate, is the common confusion of the indicators with goals, which will eventually lead to the distortion of the processes leading to the fulfillment of the goal, especially when too much emphasis is placed on the indicator. Consequently, believing that the evaluation of the adequacy of new assessment forms requires a framework that is in accordance with the broader view of validity, they suggest the following set of criteria that they believe is consistent with the nature and potential uses of performance assessments:

*Consequences*: Frederiksen & Collins's standards of directness and transparency have implications for this validity criterion according to which "high priority needs to be given to the collection of evidence about the intended and unintended effects of assessments on

the ways teachers and students spend their time and think about the goals of education" (Linn et al., 1991, p. 17).

*Fairness*: This criterion applies to the selection of performance tasks as well as to the scoring or to responses. Care should be taken that individuals or groups have equal access to resources, have similar background knowledge, exposure, and motivation, and that raters are not racially or sexually biased, etc.

*Transfer and Generalizability*: This is an expanded view of the traditional criterion of reliability. It refers to the degree to which test results can be generalized not just "from one part of a test to another or from one form to another similar (parallel) form" but "from specific assessment tasks to the broader domain of achievement" (p. 19).

*Cognitive Complexity*: Performance assessments emphasize more complex "goals" such as problem solving, comprehension, critical thinking, reasoning, and metacognitive processes. It is also important that the scoring criteria take into consideration the "processes" that students are required to exercise as well.

*Content Quality*: The quality of test content should reflect "the best current understanding of the field." This can be assured by not only relying on expert judgments about the content of the tests but also actually involving them in designing the test tasks.

*Content Coverage*: This criterion, called scope by Frederiksen & Collins (1989), ensures that, in addition to the depth of knowledge already referred to above, the test content measures the breadth of knowledge. Gaps in the content coverage could promote negative impact on teaching since teachers and learners tend to ignore the areas not included in the test.

*Meaningfulness*: More contextualized tests involve students in meaningful activities.

*Cost and Efficiency*: Practicality is a concern for large-scale performance tests. Costs need to be kept at an acceptable level by developing efficient methods of data collection and scoring.

As we can see, both sets of authors base their work on an analysis of test consequences. They also emphasize the importance of subjective judgment about the quality and comprehensiveness of test content (Moss, 1992). These categories, however, although consistent with those of the general standards for validity (APA, 1985), are more limited especially with respect to score interpretation and its value implications (Messick, 1994). So, Messick (1994, 1996) suggests that all assessments, performance or paper-and-pencil multiple-choice tests, should be uniformly evaluated by the same set of evidential and consequential criteria reflecting six aspects of construct validity, i.e., content, substantive, structural, external, generalizability, and consequential.

The *content* aspect refers to the inclusion of the tasks and activities that are comprehensive, relevant representations of the specific construct domain being tested. It is complemented by the *substantive* aspect of construct validity by appropriate sampling of the processes involved in the content domain and the empirical evidence of the actual use of such processes by test takers. The *structural* aspect, on the other hand, refers to the scoring models that are consistent with and reflect the internal structure of the behaviour domain being tested. The inferences made on the basis of the resulting test scores should be *generalizable* to broader constructs of which the assessed tasks are predictive. This aspect has implications for washback since the preparation for the assessed tasks facilitates the learning of a set of related tasks required for improved proficiency. The *external* aspect of construct validity refers to the degree to which the test scores are related to the external measures or behaviours representing the expected level of the skill being assessed.

Finally, the *consequential* aspect includes evidence related to the intended and unintended consequences of test use especially those that will lead to sources of invalidity.

Messick's six aspects of construct validity are theoretically based on his progressive matrix (1989) as introduced in Chapter Two. However, the aspects are more concrete and usable in that they can be applied to practical purposes in all educational and psychological measurement. At the same time, despite the fact that Messick introduces construct validity as including these six *distinguishable* aspects, there is enough overlap between them to allow the evidence for one aspect to be used to support some other aspects as well. For instance, evidence for generalizability and content aspects, can be especially useful for the consequential aspect and that of substantive aspect is relevant to content. So, considering that construct validity is a unifying force, evidence coming from only one of the six aspects is not sufficient to support the inferences made on the basis of the test scores. However, as long as there is enough compelling evidence to justify test interpretation and use, the lack of evidence for any specific area can be compensated for by drawing upon some other critical evidence (Messick, 1996).

These concerns have also been voiced by Bachman & Palmer (1996) in the form of the following three principles underlying their model of test usefulness.

**Principle 1** It is the overall usefulness of the test that is to be maximized, rather than the individual qualities that affect usefulness.

**Principle 2** The individual test qualities cannot be evaluated independently, but must be evaluated in terms of their combined effect on the overall usefulness of the test.

**Principle 3** Test usefulness and the appropriate balance among the

> different qualities cannot be prescribed in general, but must be determined for each specific testing situation. (p. 18)

The model itself includes six test qualities, namely, reliability, construct validity, authenticity, interactiveness, impact and practicality. The authors believe that these qualities complement each other and that the degree of the relationship between them differs with different testing situations. They consider each quality separately in its own terms, however, they agree with other researchers that "authenticity and interactiveness are related to construct validity and that impact is part of the consequential basis of test use" (p. 42, n.4). They thus emphasize that the evidence for test usefulness should come from all test qualities rather than any one individual quality.

### *4.3 Designing tests of performance assessment*

It should be clear from the discussion thus far that if the scores on performance tests are to be used to make inferences about the language ability under assessment or decisions about individuals, the development, scoring and interpretation of the test has to be construct-driven (Messick, 1994). This suggests that (as already indicated in section 4.2.3) the development of such tests should begin by clearly specifying the knowledge, skills or component skills that should be assessed, and completed by designing test tasks that best elicit the behaviours revealing those constructs.

The complexity of constructs and tasks, however, requires that test developers adopt a principled approach by following a guiding rationale that specifies the characteristics of the constructs as well as the test task(s), tying the purposes of the testing to the activities through which they are manifested and the nature of the substantive

domain. The theoretical model of language testing suggested by Bachman & Palmer (1996) is an example of such a guideline which can be applied to the design and development of the performance tests of language ability that are both direct and authentic in the sense discussed above (section 4.2). Based on the concepts presented in Bachman (1990), the model – to be discussed shortly – consists of two separate frameworks of "language ability" and "test task characteristics" that correspond to the two major stages in the design of the construct-driven performance tests introduced in 4.2.3 and include a quite comprehensive range of concepts that have to be taken into consideration in the development of tests of language ability. At the same time, as Bachman & Palmer rightly suggest, not all components of these frameworks are necessarily applicable to all testing situations. Depending on their specific purposes, test developers might, therefore, need to modify, expand or exclude some components of the models. This model can serve as a working model for the development of the testing instruments whose washback effect will be examined in this study.

In the remaining sections of this chapter, the two frameworks of "language ability" and "task characteristics" suggested by Bachman & Palmer (1996) will be presented. Chapter Six, Part III, will later illustrate in detail, how these frameworks can be applied to the practical problems in designing performance tests of language ability for the target population in this study.

### *4.3.1 Defining constructs: Test-takers' competences or abilities*

The primary consideration in designing performance tests of language ability is to determine what complex of skills and knowledge are tied to the objectives of the assessment. Because the inferences made on the basis of the test results are expected to be

generalizable across people and settings, while at the same time distinguishable from those of measures of other abilities, the specification of the skills and sub-skills to be measured by a test has to be accurate enough to serve as a basis for the inferences we can make from the test performances. On the other hand, in designing a language test, we hypothesize that the test taker's language ability will be engaged by the test tasks. This implies that a description of the abilities at issue should lead to the choice of test tasks that serve to relate abilities in actual language use to abilities involved in test performance.

Regarding this, the model of language ability that Bachman & Palmer (1996) suggest is theoretically promising in that it has two components of "language knowledge" and "metacognitive strategies"[2], a combination of which enables language users to involve in the test/non-test language use. These components, in turn, involve the integration of several other components; however, as already mentioned, it is unlikely that every single language test will be intended to measure all the sub-components in this model.

***Language knowledge***

According to Bachman (1990), language knowledge, sometimes referred to as competence in applied linguistics research, includes two broad areas of "organizational" and "pragmatic" knowledge (table 4.1), the former having to do with the formal properties of language such as grammatical and textual knowledge, the latter being the ability to create or interpret discourse such as functional or sociolinguistic knowledge.

The two areas of organizational and pragmatic knowledge are complementary, since they deal with form and meaning respectively; that is, an appraisal of language knowledge requires evidence from both areas. For both first and second language learners,

---

[2] The concepts respectively stand for the two areas of communicative competence traditionally referred to as linguistic competence and strategic competence.

the areas of language knowledge are constantly changing as new language elements are learnt or acquired, and existing elements re-structured.

**Table 4.1: Areas of Language Knowledge**
**(Bachman & Palmer, 1996, p. 68)**
Reprinted with Permission

---

**Organizational knowledge**
(how utterances or sentences and texts are organized)

**Grammatical knowledge**
(how individual utterances or sentences are organized)

Knowledge of vocabulary
Knowledge of syntax
Knowledge of phonology/graphology

**Textual knowledge**
(how utterances or sentences are organized to form texts)

Knowledge of cohesion
Knowledge of rhetorical or conversational organization

**Pragmatic knowledge**
(how utterances or sentences and texts are related to the communicative goals of the language user and to the features of the language use setting)

**Functional knowledge**
(how utterances or sentences and texts are related to the communicative goals of language users)

Knowledge of ideational functions
Knowledge of manipulative functions
Knowledge of heuristic functions
Knowledge of imaginative functions

**Sociolinguistic knowledge**
(how utterances or sentences and texts are related to the features of the language use setting)

Knowledge of dialects/varieties
Knowledge of registers
Knowledge of natural or idiomatic expressions
Knowledge of cultural references and figures of speech

***Metacognitive strategies***

In both first and second language communication, areas of strategic competence are consciously employed by speakers to better communicate their intended meaning. In second language communication, however, communication strategies are often problem-oriented; that is, the speakers resort to strategies of communication in order to avoid breakdown in communication due to a gap in their linguistic knowledge. The use of communicative strategies by non-native speakers of a language in an attempt to keep communication going is, therefore, a positive sign of their improved competence in that language. Likewise, if promoted by language tests, the use of communicative strategies by language learners could be an instance of a test's positive washback on learning.

Returning to Bachman & Palmer's model of language ability, its second macro-component is actually a model representing the processes resulting in the use of communicative strategies in language use. It is a refinement of what Bachman called "strategic competence" in his 1990 version of the model consisting of three components of strategic competence, namely "goal setting," "assessment," and "planning" (table 4.2) which correspond to the "planning" and "execution" phases of the psycholinguistic model of speech production proposed by Færch & Kasper (1983, 1984). Bachman's conception of communicative strategies, however, is more general in that it views strategic competence as part of all communicative language use, not just interlanguage communication. So, assuming that language use in all language contexts (test/non-test, first/second) requires the ability to use language both strategically and linguistically, the interactive and simultaneous use of the areas of language knowledge and metacognitive strategies in Bachman & Palmer's model should be taken into consideration in evaluating the ability of language users. On the basis of this fact, the differences in language ability

between different language users can be attributed either to their knowledge of language or their use of communicative strategies the processes of which are illustrated in the following table.

**Table 4.2: Areas of Metacognitive Strategy Use**
**(Bachman & Palmer, 1996, p. 71)**
Reprinted with Permission

---

**Goal setting**
(deciding what one is going to do)

Identifying the test tasks
Choosing one or more tasks from a set of possible tasks (sometimes by default, if only one task is understandable)
Deciding whether or not to attempt to complete the task(s) selected

**Assessment**
(taking stock of what is needed, what one has to work with, and how well one has done)

Assessing the characteristics of the test task to determine the desirability and feasibility of successfully completing it and what is needed to complete it
Assessing our own knowledge (topical, language) components to see if relevant areas of knowledge are available for successfully completing the test task
Assessing the correctness or appropriateness of the response to the test task

**Planning**
(deciding how to use what one has)

Selecting elements from the areas of topical knowledge and language knowledge for successfully completing the test task
Formulating one or more plans for implementing these elements in a response to the test task
Selecting one plan for initial implementation as a response to the test task

---

The strategies of assessment in the above model actually link the context of language use to the discourse that is used and the areas of language knowledge that the language user employs to produce or interpret utterances. As a result of the interaction

between assessment and goal-setting strategies, the goal of communication – which involves what you are going to do with the language – is determined. However, in a non-test language context, language users have more flexibility in setting goals than that in a test situation simply because language tests are usually designed to elicit a specific sample of language use. Finally, the message is linguistically formed in the planning stage by adopting those areas of language knowledge which best formulate the communicative goal.

### *4.3.2 Defining tasks: Test performances*

A second consideration in the development of a performance test is the choice of the tasks that serve to elicit the behaviours or performances that are manifestations of the abilities under assessment. Based on the discussion in 4.2, the significance of tasks for a performance test of ability comes from the fact that they reflect the competences and constructs a test is actually measuring (4.2.2), the extent to which a test relates to non-test language tasks (4.2.1), and the ways a test involves test-takers in the performance of their language abilities (4.2.3). They also affect the validity of the inferences made based on test scores (4.2.5), the contexts to which the results are generalizable (4.2.5), and the positive/negative consequences of the test use (4.2.4). A characterization of test tasks is therefore, necessary for an organized approach to test design.

Bachman & Palmer's proposed model of language testing, accounts for this by complementing the model of language ability (tables 4.1, 4.2) with a framework of task characteristics intended to relate the characteristics of the test task to the context features of target language use. It includes a set of features describing five aspects of tasks: setting, test rubric, input, expected response and relationship between input and response.

As with their model of language ability, authors suggest that test developers might need to make modifications in specific characteristics for their own purposes if necessary. The framework of task characteristics is presented in table 4.3 below.

**Table 4.3: Task Characteristics**
**(Bachman & Palmer, 1996, pp. 49-50)**
Reprinted with Permission

---

**Task characteristics**

**Characteristics of the setting**

- Physical characteristics
- Participants
- Time of task

**Characteristics of the test rubrics**

- Instructions
  - Language (native, target)
  - Channel (aural, visual)
  - Specification of procedures and tasks
- Structure
  - Number of parts/tasks
  - Salience of parts/tasks
  - Sequence of parts/tasks
  - Relative importance of parts/tasks
  - Number of tasks/items per part
- Time allotment
- Scoring method
  - Criteria for correctness
  - Procedures for scoring the response
  - Explicitness of criteria and procedures

**Characteristics of the input**

- Format
  - Channel (aural, visual)

Form (language, non-language, both)
Language (native, target, both)
Length
Type (item, prompt)
Degree of speededness
Vehicle ('live', 'reproduced', 'both')

Language of input

Language characteristic
Organizational characteristics
Grammatical (vocabulary, syntax, phonology, graphology)
Textual (cohesion, rhetorical/conversational organization)

Pragmatic characteristics

Functional (ideational, manipulative, heuristic, imaginative)
Sociolinguistic (dialect/variety, register, naturalness, cultural references and figurative language)

Topical characteristics

**Characteristics of the expected response**

Format

Channel (aural, visual)
Form (language, non-language, both)
Language (native, target, both)
Length
Type (selected, limited production, extended production)
Degree of speededness

Language of expected response

Language characteristics

Organizational characteristics

Grammatical (vocabulary, syntax, phonology, graphology)
Textual (cohesion, rhetorical/conversational organization)

Pragmatic characteristics

Functional (ideational, manipulative, heuristic, imaginative)
Sociolinguistic (dialect/variety, register, naturalness, cultural references and figurative language)

Topical characteristics

**Relationship between input and response**

Reactivity (reciprocal, non-reciprocal, adaptive)
Scope of relationship (broad, narrow)
Directness of relationship (direct, indirect)

---

The categories in the two macro-components of Bachman & Palmer's model of language testing, namely language ability and task characteristics, will be applied in Chapter Six to the actual tests developed for the purpose of the experiment in this study.

# *PART THREE*

# *EMPIRICAL CONSIDERATIONS*

## *Introduction:*

In this part of the dissertation, the conceptual framework for washback, introduced theoretically in Chapter Three, is examined empirically. From the standpoint of the present study, as extensively discussed throughout Part Two, the positive/negative washback effects of a test are related to the test and its construct validity in obvious ways even though in some cases (e.g., high-stakes tests administered at regional/national level) factors other than the test itself might be responsible for the occurrence of unintended test effects (see Wall, 1996). It is, therefore, hypothesized that the test as the central issue in a theory of washback will have positive effects on classroom teaching and learning if its contents and tasks are validated against the needs of the population for whom it has been designed. That is, tests can have positive consequences for the curriculum, the methodology and the learning outcome if they are developed on the basis of a sound assessment of the language needs of the learners and the educational system. A very important implication is that tests with beneficial washback effects can potentially be used as instruments for change in second language teaching and learning.

This hypothesis has been tested in this study through a comprehensive longitudinal multi-phase investigation conducted in different educational contexts (ESL, EFL), at different levels of proficiency (advanced, intermediate), with different tasks (oral, written) and different groups of subjects (graduate, undergraduate). The three phases of the experiment reported in this part (Chapters Five to Eight) are in fact the experimental realizations of the different levels of the theoretical framework introduced in Chapter Three. The actual process of experimentation and data collection (including needs assessment, test development, material development/selection, syllabus design, course

development, and evaluation) took place over a period of almost three years (1996-1998). The approach to data collection and analysis is both quantitative and qualitative.

# CHAPTER FIVE
# NEEDS ASSESSMENT

## *5.1 Background*

This chapter focuses on the first phase of the study, needs assessment, upon whose results the type and nature of the tests, their constructs, tasks and content will be based. Needs assessment is by far the most important determinant of language teaching and/or learning objectives. The approaches to needs assessment so far – although overlapping in many aspects – have adopted different theoretical views of language needs. Earlier trends concentrate on the language to be learned. The process of needs analysis in this sense is the process of discovering about the areas of language use already possessed as well as those needed by a group of learners. The second trend, on the other hand, interprets needs in a more humanistic way by making the individual the focus of needs assessment. According to this view, needs analysis is not just the process of finding out about a learner's language needs but that of discovering the socio-cultural factors affecting the individual's learning. These two types of analysis have been referred to in the literature as objective versus subjective (Richterich, 1980), or inductive versus deductive (Berwick, 1989) methods of needs analysis.

Exemplars of the inductive approaches to needs analysis are the studies in which language needs at an individual level are assessed through such methods as the observation of people in various settings and life situations (Freire, 1970), the analysis of communications in work situations (Chambers, 1980), case studies of individuals engaged in language-dependent tasks (Schmidt, 1981), and asking stakeholders or experts in a given field (Holmes, 1977). As for a deductive approach to needs analysis, the most

comprehensive example is the "Communicative Needs Processor" of Munby (1978) whose vast multitudes of categories aim at collecting factual information about the learners for the purpose of "finding out the communication needs that are pre-requisite to the appropriate specification of what is to be taught" (Munby, 1978, p. vi). However, the abundance of categories in Munby's model, while justified in some ways and applicable to a number of ESP situations, make it impractical for most program planners to use the model in its entirety. Besides, the model does not offer clear guidelines with respect to the specification of actual language forms needed by a group of individuals (Schutz & Derwing, 1981), nor does it examine the problem of interpreting data (Berwick, 1989). At the same time, because no language program can present the entire language, it is not sensible to ignore the objective methods of needs analysis. They help to make informed choices with respect to the content of language courses based on the current strengths and future goals of the learners. It thus appears that the two methods of needs analysis complement each other and are thus necessary for making a sound judgment regarding learners' language needs especially when there are discrepancies between what learners say they need and what teachers, authorities and experts think they really need.

In this study, however, needs assessment has been conducted primarily for the purpose of test development. Major questions at this stage are what the objectives of testing in the contexts of this study are, what kind of language abilities they are supposed to measure, and what type of tasks they should include in order to adequately measure such abilities. These questions are expected to be answered in the light of a description of the language needs in these contexts as well as a description of the individuals who need or want to function in such contexts.

Considering this, the approach adopted for identifying the needs of the learners in this study is based on not only the characteristics of the learners (as that advocated by Munby, 1978), but also the informed judgment or expert opinion. This implies that our description of test-takers' needs is based on a variety of sources and pieces of information. It is thus directed towards the common problems of larger classes of learners, rather than towards the specific needs of individual students. The choice of the instrument used for the purpose of information gathering is largely based on the nature and context of the investigation: In the EFL context, the familiarity with the educational context, test-takers, their academic and language needs enabled us to form an initial description of learners' characteristics and their language needs. This was supplemented by information obtained more systematically through further discussing the problem with the stakeholders who were familiar with the language use context. In the ESL context, on the other hand, due to widely varied nationalities, native languages, and heterogeneous fields of specialization, a combination of interview and observation proved to be more efficient. In the following sections, after a short introduction of the contexts of the study, the process of needs analysis carried out in each context of the study and the resulting description of the learners and their language needs will be presented in some detail.

## *5.2 Contexts of the study*

### *5.2.1 ESL context*

With the internationalization of the University of Victoria (UVic) as one of the goals of UVic's strategic plan, increasing numbers of international students attend the university's graduate programs in various faculties and departments each year. Many of these

candidates also function as teaching, laboratory and research assistants. The Faculty of Graduate Studies and the departments thus need to be assured that all International Teaching Assistants (ITAs) have English language proficiency skills which enable them to instruct undergraduate students effectively. However, when this study started in 1996, other than a TOEFL score of 550 required of all international graduate students (regardless of their appointment as teaching assistants) at the time of admission, the university did not have any legislation or policies regarding the assessment of potential Teaching Assistants' (TAs) oral proficiency. Neither was there any screening device or training program in effect at the English Language Centre for those new ITAs who sought out help or were sent by their graduate advisors. The only testing device available in the past had been a speaking test – developed by the English Language Centre – which was more an indirect measure of the basic components of the oral skill rather than a direct measure of the spoken abilities authenticated and validated against the needs of the ITAs in the real instructional contexts where they had to use the target language (we will return to this point later in Chapter Eight).

This situation, backed by the results of the undergraduate students' evaluations of the classes run by ITAs as well as the recommendations of some departments and graduate advisors, brought to the attention of the Faculty of Graduate Studies the need for the establishment of a testing mechanism to determine whether these students have a sufficient level of proficiency of spoken English to be able to successfully carry out their duties as TAs. Further, there was a need for a learning mechanism for those ITAs whose test scores indicated that their instructional contacts with the undergraduate population would be adversely affected without such a program. It was thus expected that such a test together

with a training program geared towards the objectives of the test would assist international graduate students in the efficient discharge of their instructional and research duties.

### *5.2.2 EFL context*

Graduate and undergraduate programs in areas such as English Language, Translation, Literature, and Applied Linguistics (TEFL), have long been taught in English in Iranian universities. Students applying to the undergraduate programs in these areas are accepted to the universities on the basis of their performance in the English language component of the University Entrance Examination. Graduate students, on the other hand, have to pass an English proficiency test as well as a specialized written test. In the past, almost all of the small number of students who entered the graduate programs in these areas already had a good command of English on arrival at the university. During recent years, however, as a result of an increase in the population of the university students, especially in graduate programs, the number of students with inadequate proficiency has increased considerably. The difficulty exhibited over the years by both undergraduate and graduate students in writing their reports, term papers and theses, has confirmed that the most problematic area for these students is the writing skill, the area tested by neither the graduate nor the undergraduate Entrance Examinations. At the same time, factors present in an EFL learning environment (such as non-English-speaking staff, sharing the same native language, and L1 dominance outside classrooms) make on-the-spot learning practically impossible.

Nevertheless, because of the Ministry of Higher Education's decision to maintain the tradition of running these programs through the medium of English and achieve a high standard of English, both graduate and undergraduate classes in the above areas continue

to be uniformly conducted in English in all universities throughout the country. Sources and course materials are in English and so are the students' reports and assignments. Graduate students are required to write and defend their theses in English as well.

As a result of this incompatible situation between the writing ability of the learners and the aims of their education, the curriculum for graduate studies in these fields has been modified by adding an obligatory "Advanced Writing" course whose purpose is to further the ability of graduate students in advanced academic writing. At the same time, the course description for undergraduate writing courses has also been improved by shifting the emphasis from teaching advanced grammar at the sentence level to the actual practice of writing long essays. Soon after these changes, the universities offering these programs realized that the only way to assure that these courses target the needs of the students was to establish a testing device that would validly measure the ability in question.

## *5.3 Describing needs*

### *5.3.1 ESL learners*

As already referred to in 5.2.1, the problem at large at the University of Victoria was the problem of ITAs having difficulty communicating with their undergraduate students during their interactions with them in instructional settings. The question to begin with was whether ITAs have a sufficient level of English language proficiency to be able to successfully carry out their duties as TAs. The objective, as put forth by the Faculty of Graduate Studies, was the development of a testing instrument that: (1) would enable test users to make inferences with respect to the language ability of the candidates, (2) could be administered to ITAs in different academic disciplines and with different language

backgrounds, (3) could serve as an entry and exit criterion for a TA preparation course, and (4) would influence the kind of teaching and learning that would bring about the language ability necessary for teaching assistantship.

The above requirements, however, gave rise to a number of open-ended questions. For the sake of needs assessment, these questions were grouped into three different categories based on whether they asked for the specification of the language use contexts, the identification of the task types, or the description of the test-takers' characteristics:

1. What exactly is the source of the communication problem? Is it merely a language problem, or lack of knowledge of the subject matter, or both?
2. What is the English language background of the ITAs in question? How long, in what capacity and in what type of environment had they been exposed to English before they came to UVic?
3. What linguistic background do they come from? Do all or most of them share the same culture and/or native language?
4. Is a language problem the only reason for undergraduate students' complaints?
5. Are cultural differences a major problem further complicating the situation?
6. What is the ITAs' motivation for accepting teaching assistantships in the departments? Is it purely academic or do they see it as one of the many ways of financing their graduate studies?
7. Are ITAs interested in improving their language proficiency level? Why?
8. Are they willing to take a course to improve their language skills?
9. In what contexts do ITAs have to display their language abilities? What are the characteristics of these contexts? Do all of them require the same level of the ability in question?

10. If language is the source of communication problems, which language ability components are most important to ITAs in their dealings with undergraduate students?
11. What are the areas of language ability in which a significant number of ITAs are seriously deficient?
12. What are the activities and tasks ITAs have to perform in such contexts?

Given the nature of the problem and the scope and objectives of the investigation, besides the investigator's observation, three major groups of participants had to be consulted. These were: (1) administrators and graduate advisors, (2) undergraduate students, and (3) ITAs themselves. The instrument used was an oral interview since it permits the interviewer to ask for clarification on certain topics and/or pursue unanticipated lines of inquiry. As for objective information regarding a participant's general background, relevant parts of Munby's 1981 framework were used.

#### *5.3.1.1 Analysis of the results*

The analysis of the general background characteristics of the non-native-speaking ITAs' at the University of Victoria – introduced to us by graduate advisors in different departments – indicated a mean age of thirty-three and some 45 percent female students. These were from different nationalities and cultures (Japan, India, Taiwan, China, Korea, Ghana, Iran, Germany, Ukraine, Poland, Bulgaria) specializing in different academic disciplines. A strong majority of them (87%), however, specialized in different areas of engineering and sciences. All of the population indicated that they had had at least five years (with an average of more than six years) of formal EFL training back home and had spent an average of more than two years in Canada.

Our survey also revealed that TA's jobs at the University of Victoria consisted mainly of assisting professors in grading exams, preparing course materials, supervising laboratory experiments, holding office hours, tutoring students, substituting for professors at times, and even teaching independent courses (especially during summer sessions). There are, therefore, three instructional contexts in which TAs have to display their academic as well as their language skills: classroom, laboratory, and office hours.

As for the source of ITAs' communication problems in the above settings, just as administrators and graduate advisors believed that teaching assistants were all academically qualified for the job to which they were assigned, almost all of the ITA population consulted described their own knowledge of the subject matter they are assigned to teach as "sufficient," "very good" or "more than necessary," no one rated him/herself as "poor." Neither was there any reference by the undergraduates in this sample to the academic incompetence of ITAs as a source of the problem.

There were, nevertheless, different views and conceptions as to the ITAs' performance as affected by their proficiency in English language. Whereas all graduate advisors supported the decision of the Faculty of Graduate Studies 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 native-speaking ones is low. Besides, the academic success and the internalization of abstract ideas in such areas as philosophy, psychology, and the general subjects of the humanities is often difficult at best without possessing the language ability to communicate, so most of the foreign students entering such programs already possess high levels of language proficiency at the time of entry. Graduate advisors in Sciences and Engineering, on the other hand, felt that the

problem was a serious one and were interested in upgrading ITAs' spoken language proficiency. This was justified by the fact that eighty-seven percent of the students referred to us by the departments were in Engineering and Science, only thirteen percent were in Humanities and Fine Arts. Similarly, undergraduate students who took part in the survey characterized ITAs' spoken language problems as the source of their communication difficulties in their dealings with undergraduates.

In ITAs' evaluations of their performance in instructional settings, however, there were mixed feelings. About one third of the respondents were confident that they had no problem communicating with native-speaking undergraduates. The rest of the sample, indicated, one way or another, that they had concerns about their spoken English, but none of them believed that cultural differences influenced their use of language in academic settings. Neither did any undergraduate student recall any cultural reasons for the communication problems in the classroom or other instructional settings. They indicated though that ITAs are more serious in the classroom than native TAs and sometimes don't understand sarcasm, but they also mentioned that they would never consider this as a source of difficulty as long as they understood what the ITAs were teaching. Some ITAs, on the other hand, believed that not all undergraduate students are so sincere in their evaluations of teaching assistants and that they sometimes wanted to blame ITAs for their own poor class attendance and grades.

When asked if they were interested in improving their efficiency as a TA through a course in spoken English, all ITAs said "yes," however, about half of the respondents believed that taking such a course should be "voluntary," not a requirement for assignment as a TA. This attitude was in part indicative of ITAs' motivation for applying for and accepting TA positions. While they all valued teaching assistantship as a great opportunity

for gaining teaching experience and an avenue for entering the teaching profession, the majority of them (85%) considered it primarily as a source of income during their graduate years.

### *5.3.1.2 Interpretation of the results*

Based on the above information, we determined that the contexts in which foreign graduate students have to communicate as TAs are those of the classroom, laboratory, and office hours. The type of discourse used in such contexts is by nature bilateral, that is, it involves the speaker not only in the production but also in the comprehension of the spoken language. It is also complex in that it draws upon areas of socio-cultural and strategic competence as well as areas of linguistic competence.

Returning to the theoretical model of language ability (Bachman & Palmer, 1996) presented in Chapter Four, the language ability required by ITAs can best be explained in terms of the following components of the model:

1. The area of language knowledge (Table 4.1) ITAs are mostly concerned with is that of Textual Knowledge. "Text" in the ITAs' case is essentially conceived as a unit of "spoken" language consisting of a number of utterances that are structured according to the rules of cohesion and rhetorical organization. Cohesion comprises ways of explicitly marking semantic relationships such as reference, ellipsis, conjunction and lexical cohesion (Halliday & Hasan, 1976), as well as conventions governing the ordering of old and new information. Rhetorical organization includes common methods of development such as narration, description, comparison, classification, analogy and process analysis.

2. Pragmatic Knowledge is another area of language knowledge evoked by the contexts in which ITAs have to perform. When they use language for performing different functions, they employ their functional knowledge of the language and when they switch registers appropriately or use cultural references, socio-linguistic knowledge is involved. Pragmatic knowledge on the part of TAs also implies the ability to use interactional language, especially in settings such as classrooms, by employing those conversational conventions Hatch (1978) refers to as attention getting, topic nomination, topic development and conversational maintenance. It is this emphasis on the interactive or reciprocal nature of the language use by TAs (in instructional settings) that necessitates the analysis of their use in their test performance of the devices that mark cohesive relationships in the oral discourse on the one hand and their organization of such discourse on the other hand.
3. Strategic Competence is by far the most important ability ITAs can draw upon to perform successfully in the highly interactive contexts described above. Examples of such strategic uses of language are rephrasing, paraphrasing, description, exemplification, generalization, circumlocution, repetition, mime ... etc. Successful communicators always exhibit the use of these compensatory strategies rather than avoidance strategies. Furthermore, since the strategic use of language also requires the simultaneous use of areas of language knowledge, the frequency and accuracy with which the speakers employ communication strategies is indicative of how advanced they are in their interlanguage use and/or how well they can manipulate the language to get around the gaps in their linguistic knowledge and thus keep the communication going.

In short, in order to function effectively in instructional contexts, teaching assistants should have the ability to produce language by processing several language tasks – such as chunking information, organizing it, providing transitional cues, and using appropriate means for best transferring it to the students – simultaneously. At the same time, it is important for them to be able to comprehend the spoken language in both formal and informal contexts. This implies that a test designed for assessing non-native TA's oral proficiency should involve them in the appropriate expression and interpretation of utterances in an interactive setting, i.e., capture their communicative language ability.

### *5.3.2 EFL learners*

The testing problem we face for this group of learners is completely different from that of ESL learners in that the language ability in question here is the writing ability and in that all learners share the same language background and academic areas of specialization. As we know from 5.2.2 above, these learners consist of two groups of advanced graduate and intermediate undergraduate subjects, the needs of whom will be dealt with separately in the following sub-sections:

#### *5.3.2.1 Advanced EFL group*

This group consists of graduate students in Applied Linguistics who had to participate in an advanced writing course in order to improve their academic writing skills. The global objective was to prepare them to write term papers, reports, proposals and eventually their theses in English. The end-of-term exam was thus supposed to involve the test-takers in

tasks representative of such activities and determine if they had properly mastered the skills necessary for such a purpose.

As in the case of ESL learners, a needs analysis was done to further specify the areas of language ability as well as language use contexts, test-takers' characteristics and the task types needed to be included in the test. The following questions were thus raised:

1. What is the population for whom the test is intended?
2. What is the general level of the test-takers' language ability?
3. Do the test-takers share the same level and type of topical knowledge?
4. In what contexts are language tasks used?
5. What is the problem area and its significance for the students' academic success?
6. What kind of writing tasks are the learners required to perform in real life: essay-type assignments, term papers, reports of experiments/observations, exam answers, note-taking, other?
7. What are the areas which EFL graduate students have problems with: specialized vocabulary, general vocabulary, writing grammatical sentences, linking sentences, organization, mechanical aspects of writing (such as punctuating, quoting, giving references and so forth)?

With respect to the personal characteristics of the learners, the population consists of male and female Masters students varying between 23 and 30 years of age. They all share the same native language, Persian, and have had a minimum of 10 years of exposure to English. They form a homogeneous group in terms of their topical knowledge since they are all pursuing a degree in the same field of study. They are also considered advanced in their general level of language ability based on the results of the English

language test – a score equivalent to TOEFL 550 – which they take at the time of admission.

The context(s) of language use for the test-takers would be those of academic contexts where they have to express their knowledge/information of the academic subjects in the form of written term papers, reports, and proposals. According to instructors and supervisors, of all in-class and out-of-class academic activities that require language ability of some sort, the activities students have most problems with include relatively long written assignments such as term papers, and proposals in which they have to produce a coherent well-developed piece of work. Faculty members also expressed their discontent with the amount of time and effort they had to spend on different language-related issues in the writings of graduate students despite the fact that these students often have a good mastery of the topical knowledge. They commented that frequent language-related problems could result in the loss of train of thought and misinterpretation of the technical information on the part of the reader.

Among the components of the language ability in question, the appropriate use of general vocabulary, organization of the written work, use of linking devices, complex grammatical structures, and punctuation are the areas mostly emphasized by the instructors.

#### *5.3.2.1.1 Interpretation of the results*

The above facts point to written language ability in formal academic contexts as the main source of learners' language difficulties. Interpreting the results in terms of Bachman & Palmer's (1996) model of language testing, the language ability to be expected from the learners in this context can be explained in the following way:

1. With respect to the areas of language knowledge, unlike our ESL context, "textual knowledge" in this context refers to the ability to produce a piece of written text with a topic developed through a number of sentences cohesively related and rhetorically organized into paragraphs. The vocabulary needed for this purpose includes both technical and general vocabulary.
2. The pragmatic knowledge of the learners involves the functional use of the language for the purpose of communicating knowledge and ideas. The relevant sociolinguistic knowledge here would be that of dialect and register appropriate to the language use context.
3. Despite its presence in all instances of language use, not all areas of strategic competence are going to be relied upon in the inferences made about the language ability of the test-takers. This is mainly because the types of tasks they are involved in here are unspeeded, non-reciprocal and less interactive compared with those of the ESL group.

In short, in order to be able to communicate their topical knowledge in the form of written texts, the population in our EFL context needs to produce a relatively long piece of work in formal standard English using both general and technical vocabulary as well as well-formed sentences which are cohesively structured. They also need to know what processes to follow to form their ideas coherently.

### *5.3.2.2 Intermediate EFL group*

This group consists of freshman undergraduate learners who have completed a course in advanced grammar and are about to take a first course in free composition. Using basically the same questions that we had used for gathering information regarding advanced EFL

learners, we found out that the situation for this group of learners resembles that of the advanced group in that the learning environment is an EFL environment, the learners share the same topical knowledge and native language, and the language ability to be tested is writing. The difference, however, lies in the fact that learners in this group are undergraduate students at an intermediate level of proficiency whose ability to write free compositions on topics of general knowledge is going to be tested.

The general level of learners' language ability is determined on the basis of the English language test that they have to pass – with a minimum score equivalent to TOEFL 500 – at the time of entry. The contexts of language use for test-takers include those non/instructional domains in which the learners have to express their thoughts, ideas, and information clearly and appropriately, though possibly with some grammatical and vocabulary mistakes that do not disrupt communication.

The types of tasks learners are required to perform in real life/instructional settings include outlining, summarizing, writing essay exams, and writing essays and/or term papers using one of the common patterns of development such as narration, comparison/contrast, description, classification, process, cause and effect, and so forth.

#### *5.3.2.2.1 Interpretation of the results*

Speaking in terms of the components of Bachman & Palmer's (1996) model of language ability, the points that have to be taken into consideration in designing a test for this group of learners are similar to those of the advanced EFL group except that here depending on what type of task the learners are engaged in (e.g., outlining vs. essay-writing), they have to draw upon their knowledge of grammar and vocabulary to different degrees. Moreover, knowledge of technical vocabulary is not needed here and textual knowledge is limited to

the use of common cohesive devices and patterns of thought development without having to resort to extensive rhetorical organization.

Also, because of the characteristics of the tasks learners have to perform (see Chapter Six for details), instances of the strategic use of the language are not used for making inferences about the written language ability of the learners.

In sum, the process of needs assessment described in this chapter provides us with information regarding both the purposes of test development and the uses to which test results will be put. Recall from Chapter Two that these are the two evidential and consequential aspects of the validity theory underlying our theory of washback. The description of the language use contexts, characteristics of the test-takers and the areas of language ability to be assessed are, therefore, of significance to the next stage of the study, test design, in that they enable us to define the constructs in question by specifying not only the characteristics of the ability or behaviour(s) to be measured by each test but also the uses to which the tests are going to be put.

# CHAPTER SIX
# TEST DEVELOPMENT

## *6.1 Introduction*

In Chapter Five, we described the characteristics of the three groups of learners for whom the tests are intended, the purposes for which the tests need to be developed, the uses to which test results will be put, and the test contexts. Based on observations and a needs analysis carried out with the stakeholders in each context, we also generally identified types of tasks and the areas of language ability that should be taken into consideration in designing test tasks.

This chapter presents systematically the procedures of test design followed for each one of the three (one ESL and two EFL) different contexts described in Chapter Five. It focuses on a detailed specification of the characteristics of the language use tasks in the corresponding contexts and, more importantly, a definition of the constructs underlying the choice of the test tasks. This will take place with direct reference to the theoretical considerations discussed in Part I, particularly the relevant components of the two models of task characteristics and language ability presented in Chapter Four, Tables 4.1, 4.2, and 4.3.

## *6.2 ESL context*

### *6.2.1 Non-test language use tasks*

Specifying and describing the tasks in the actual non-test language use context, referred to as the Target Language Use (TLU) domain by Bachman & Palmer (1996), is the first step

towards the development of test tasks that are authentic and correspond as closely as possible to the characteristics of the relevant domain.

Three types of language use tasks in the ESL context are identified by our survey as the most fundamental tasks ITAs are engaged in. In this section, the characteristics of these tasks will be formalized (Table 6.1) in terms of the theoretical model of task characteristics presented in Table 3.1. These will provide a basis for the development of the test tasks that correspond reasonably to the non-test language use tasks.

**Table 6.1:**
**Characteristics of Target Language Use Tasks in the ESL Context**

| | TLU Task 1 | TLU Task 2 | TLU Task 3 |
|---|---|---|---|
| | teaching undergraduate courses | supervising laboratory sessions | holding tutorials/office hours |
| **Characteristics of the setting** | | | |
| Physical characteristics | Location: on-campus, well-lit classroom<br>Noise level: normal<br>Temperature and humidity: comfortable<br>Materials and equipment and degree of familiarity: books, notes, handouts, blackboard, opaque projector,...etc., all familiar to the test-takers. | Location: mostly Science/engineering labs, well-lit.<br>Noise level: Varied, including quiet or relatively noisy<br>Temperature and humidity: comfortable<br>Material and Equipment: varied, including lab equipment, familiar. | Location: on-campus classroom/office, well-lit.<br>Noise level: quiet<br>Temperature and humidity: comfortable<br>Materials and equipment: same as Task 1 |
| Participants | Undergraduate students | Same as Task 1 | Same as Task 1 |
| Time of task | Monday-Friday, daytime, evenings | Same as Task 1 | Same as Task 1 |
| **Characteristics of the input** | | | |
| **Format** | | | |
| Channel | Oral/aural and visual | Same as Task 1 | Same as Task 1 |

| | | | |
|---|---|---|---|
| Form | Language/non-language (tables, pictures, equations, graphs) | Same as Task 1 | Same as Task 1 |
| Language | Target (English) | Same as Task 1 | Same as Task 1 |
| Length | Varied including short or long oral or written prompts and tasks | Same as Task 1 | Mostly short prompts (questions) |
| Type | Prompt and task | Same as Task 1 | Same as Task 1 |
| Speededness | Unspeeded | Same as Task 1 | Same as Task 1 |
| Vehicle | Live and reproduced | Same as Task 1 | Live |
| **Language of input** | | | |
| Organizational characteristics | | | |
| Grammatical | Both technical and general vocabulary, widely varied grammatical structures, Phonology: generally comprehensible | Same as Task 1 | Same as Task 1 |
| Textual | All sorts of linking devices and mostly conversational organization patterns | Same as Task 1 | Same as Task 1 |
| Pragmatic characteristics | | | |
| Functional | Ideational, manipulative (including instrumental and interpersonal) | Same as Task 1 | Same as Task 1 |
| Sociolinguistic | Variety of dialects, mostly standard Canadian<br>Register: formal and informal, natural language | Same as Task 1 | Same as Task 1 |
| Topical characteristics | Varied, mostly academic technical topics | Same as Task 1 | Same as Task 1 |
| **Characteristics of the expected response** | | | |
| **Format** | | | |
| Channel | Oral | Same as Task 1 | Same as Task 1 |
| Form | Language and non-language (tables, graphs, pictures, etc.) | Same as Task 1 | Same as Task 1 |
| Language | Target (English) | Same as Task 1 | Same as Task 1 |
| Length | Relatively long (50-100 minutes long) | Same as Task 1 | Variable (depending on the number and nature of the problem areas) |
| Type | Extended production response | Same as Task 1 | Same as Task 1 |
| Speededness | Speeded (certain amount of material has to be covered during the class time) | Same as Task 1 | Relatively speeded |

| | | | |
|---|---|---|---|
| **Language of expected response** | | | |
| Organizational characteristics | | | |
| Grammatical | Vocabulary: general and technical, Varied grammatical structures, Intelligible pronunciation | Same as Task 1 | Same as Task 1 |
| Textual | Cohesive oral text presenting well-organized pieces of information all contributing to a topic, use of common methods of development | Cohesive presentation involving a topic stated at the beginning, common rhetorical methods involve description, explanation step by step analysis, etc. | Same as Task 2 |
| Pragmatic characteristics | | | |
| Functional | Ideational, manipulative (including instrumental and interpersonal), heuristic | Same as Task 1 | Same as Task 1 |
| Sociolinguistic | Standard dialect, both formal and informal register, natural language | Same as Task 1 | Same as Task 1 |
| Topical characteristic | Academic, technical topics | Same as Task 1 | Same as Task 1 |
| **Relationship between input and response** | | | |
| Reactivity | Reciprocal | Same as Task 1 | Same as Task 1 |
| Scope of relationship | Broad | Same as Task 1 | Same as Task 1 |
| Directness of relationship | Mainly indirect | Same as Task 1 | Same as Task 1 |

The task characteristics specified in the above table serve to distinguish the set of tasks used in our ESL context from each other and other real-life tasks. They are thus referred to as "distinctive task characteristics" by Bachman & Palmer (1996). However, while these tasks are directly relevant to the purposes of the test, whether or not all or some of them are going to be considered as possible test tasks depends on the extent to which their characteristics represent the components of the construct definition underlying the test and contribute to the practicality of the test. This is a point to which we will turn

in section 6.2.3 where test task characteristics will be discussed. Meanwhile, in the upcoming section, the construct to be measured by the test is going to be defined in terms of the areas of language ability introduced in Tables 4.2 and 4.3.

### *6.2.2 Construct definition*

At this stage, based on the purposes of the test, we have to identify and define the abilities/ components of abilities that are to be measured by the test. This specification will justify the use of the test results for the intended purpose(s) of the test. Depending on the number of the components that we want to measure with respect to an ability, it will also directly affect the scoring method of the test. We should, therefore, define the ability in question precisely and clearly so that it is not only appropriate for our specific testing purpose, the population taking the test, and the testing situation, but also distinguishes one testing situation from another. This definition of the language ability is in fact the construct theoretically underlying the content and activities of the test.

Returning to our ESL context, the purpose of the test is to make inferences about test-takers' language ability to perform in a range of instructional settings in which speaking is necessary. The definition of the constructs is going to be a broad outline of the areas that should be tapped by the test since the test is not going to be used for measuring specific structures or language components the way they are measured in an achievement test. We thus use a theory-based construct definition in terms of the relevant components of the theoretical model of language ability (Bachman & Palmer, 1996) presented in Chapter Four, Tables 4.2 and 4.3. So, considering the test's purpose, the specific needs of the test-takers, and the testing context, the areas of language ability included in the

construct definition – that guides test development for this testing situation – are summarized in the following table:

**Table 6.2:**
**Constructs to be Measured in the ESL Context**

| | |
|---|---|
| **Language knowledge** | |
| Grammatical knowledge | Ability to draw upon syntactic, vocabulary and phonological knowledge simultaneously for the purpose of processing and producing 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 a 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 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 from the completion of the default task), and draw upon the areas of language knowledge for successfully implementing and completing the chosen task. |

Two points are worth mentioning at this point. First, strategic competence is included in our definition of the constructs to be measured in the ESL context since the

test-takers (i.e., ITAs) need to demonstrate their ability to adapt their language use to unpredictable situations/questions that might arise in the course of communication. Because their primary job is to transmit information and knowledge to undergraduate students, it is very important that they are able to keep communication going by making use of any verbal and non-verbal means at their disposal.

Topical knowledge, on the other hand, is not included in the construct definition primarily because ITAs come from different academic backgrounds and major in different areas. Besides, our survey showed that departments assign assistantships on the basis of the TA's academic preparedness for the job, that is, TAs are assigned to the courses for which they are academically fit. The test scores, therefore, will be used to make inferences about the language ability of the TAs not their specialized knowledge.

### *6.2.3 Test tasks*

ITAs perform a number of activities in instructional settings not all of which can be considered as possible test tasks. This is because some tasks such as grading or material preparation are not directly related to the purpose of the test of speaking ability. On the other hand, some other tasks, such as those characterized in Table 6.1, are relevant to the purpose of the test, but for obvious practicality reasons they cannot all be included in the test. So, based on the descriptions of the three representative language use tasks in Table 6.1 and the existing overlap between them, the characteristics of Task One (teaching) have been used as a basis for describing test tasks. This is because the teaching task is challenging enough to measure test-takers' ability in the areas specified by the test construct. The reverse, however, does not hold. Task Three (holding office hours), for

example, is not long enough to tap areas of language ability (such as strategic competence). Likewise, the activity in Task Two (lab supervision), because it is limited to giving guidance about an experiment or process under way, does not sufficiently cover certain areas of functional and textual knowledge.

The test (Appendix One, Part A) is designed, therefore, around a teaching task with two parts: a teaching part, and a question/answer part. In this way, we will be able to simulate more closely the natural situation of a classroom and at the same time incorporate basic properties of TLU Task Three (holding tutorials and office hours). In addition, including a question/answer part will provide a better opportunity for the test-takers to demonstrate both their language knowledge and their strategic ability.

Drawing upon the characteristics of Task One and the definition of the constructs given above, test task specifications are summarized in the following way.

**Table 6.3:**
**Characteristics of Test Task**

| | |
|---|---|
| **Characteristics of the setting** | |
| Physical characteristics | Location: on-campus classroom, well-lit, comfortable temperature<br>Noise level: normal<br>Materials and equipment: books, notes, blackboard, overhead projectors, video-camera, …etc.<br>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, 2-3 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 |
| 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 few characteristics of the test task (such as the presence of a video-camera in the classroom, the degree of familiarity with some participants, and the length of the task) are slightly different from those of the real life setting (i.e., TLU Task One in Table 6.1) due to the requirements of reliability and practicality. However, there are some measures that can be taken in order to assure that the test-taker's performance is not adversely affected by these factors, for example, the video-camera can be removed for certain test-takers who express concern over its operation while they are performing and the test can be preceded by a short warm-up for the purpose of familiarizing test-takers with ESL raters.

The time allotted to the whole task is 15 minutes, during which the test-taker presents a 10-minute lesson and answers questions for 5 minutes. Because topical knowledge is not part of the construct definition, the topic of the presentation is chosen by the test-taker. This enhances the authenticity of the task since in real-life situations instructors determine the content of the syllabus and prepare for the class in advance.

Like the test task, the scoring of the test is affected by the definition of the construct or language ability. This means that the rating instrument used for the test includes as many components as that of construct definition (see Appendix One, Part B) and the performance of the students on each component will be analyzed in terms of the levels of ability exhibited in fulfilling the test task. A 5-point ability-based scale ranging from lowest ability level "no production at all" to highest level "excellent performance" is thus used for this purpose (Appendix One, Part C).

Appendix One reflects the finalized version of the test itself, its rating instrument, the rating scale, as well as a description of the ability components measured by the test and listed in the rating instrument.

## *6.3 EFL context: Advanced group*

### *6.3.1 Non-test language use tasks*

The second test developed for the purpose of this study, as discussed in Chapter Five, targets the writing ability of advanced EFL learners sharing the same native language and area of specialization. The purpose of the test is to serve as an advanced writing course exit criterion measuring test-takers' ability in academic writing.

Here also we are going to initiate the process of test design by specifying the distinctive characteristics of the tasks that test-takers are likely to perform in real language use contexts. As in the case of the ESL test above, these characteristics, together with a definition of the focal constructs to be measured by the test, will constitute a basis for the characterization and development of the test tasks. Thus, following our theoretical model of task characteristics (Table 4.1), Table 6.4 summarizes the characteristics of the three TLU tasks in our EFL advanced language use context.

**Table 6.4:**
**Characteristics of the TLU Tasks in the Advanced EFL Context**

| | TLU Task 1 | TLU Task 2 | TLU Task 3 |
|---|---|---|---|
| | writing term papers on academic topics | writing take-home/essay exams | writing proposals |
| **Characteristics of the setting** | | | |
| Physical characteristics | Location: home, library<br>Noise level: quiet<br>Temperature and humidity: comfortable<br>Materials and equipment and degree of familiarity: books, dictionaries, notes, | Location: home, class-room.<br>Noise level: quiet<br>Temperature and humidity: comfortable<br>Material and Equipment: same as Task 1. | Location: home, library<br>Noise level: quiet<br>Temperature and humidity: comfortable<br>Materials and equipment: same as Task 1 |

| | | | |
|---|---|---|---|
| | journals, pen/pencil, paper, word processors, all familiar to the test-takers. | | |
| Participants | Supervisor, course instructor, librarians, classmates | Course instructor | Same as Task 1 |
| Time of task | Any time | Tests: weekdays, daytime, evenings, Take-home exams: anytime | Same as Task 1 |
| **Characteristics of the input** | | | |
| **Format** | | | |
| Channel | Both visual and oral | Same as Task 1 | Same as Task 1 |
| Form | Language | Same as Task 1 | Same as Task 1 |
| Language | Target (English) | Same as Task 1 | Same as Task 1 |
| Length | Varied including short topics or long oral or written prompts | Short prompts such as questions | Mostly short prompts such as research questions |
| Type | Prompt and task | Same as Task 1 | Same as Task 1 |
| Speededness | Unspeeded | Same as Task 1 | Same as Task 1 |
| Vehicle | Live and reproduced | Reproduced | Same as Task 1 |
| **Language of input** | | | |
| Organizational characteristics | | | |
| Grammatical | Both technical and general vocabulary, widely varied grammatical structures, typewritten | Same as Task 1 | Same as Task 1 |
| Textual | All sorts of linking devices and organization patterns | Same as Task 1 | Same as Task 1 |
| Pragmatic characteristics | | | |
| Functional | Ideational, manipulative (mostly instrumental) | Same as Task 1 | Same as Task 1 |
| Sociolinguistic | Standard dialect, Register: formal, natural language | Same as Task 1 | Same as Task 1 |
| Topical characteristics | Various topics in applied linguistics | Same as Task 1 | Same as Task 1 |
| **Characteristics of the expected response** | | | |
| **Format** | | | |
| Channel | Visual (written) | Same as Task 1 | Same as Task 1 |
| Form | Language | Same as Task 1 | Same as Task 1 |
| Language | Target (English) | Same as Task 1 | Same as Task 1 |
| Length | Long | Short (1-3 paragraph | Relatively long but |

| | | | |
|---|---|---|---|
| | | long for essay exams) and relatively short (maximum 5 pages for take-home exams) | shorter than Task 1 |
| Type | Extended production response | Same as Task 1 | Same as Task 1 |
| Speededness | Unspeeded | Relatively speeded | Same as Task 1 |
| **Language of expected response** | | | |
| Organizational characteristics | | | |
| Grammatical | Vocabulary: general and technical, Varied grammatical structures, typewritten | Same as Task 1, both hand and typewritten | Same as Task 1 |
| Textual | Cohesive written text presenting well-organized pieces of information all contributing to a topic, use of common methods of development, subtitles often used, research problem or hypothesis usually stated at the beginning | Cohesive written text usually arguing for or against a theory, explaining/defining a concept, or reviewing some academic work | Same as Task 1 |
| Pragmatic characteristics | | | |
| Functional | Ideational, manipulative (mostly instrumental), heuristic | Same as Task 1 | Same as Task 1 |
| Sociolinguistic | Standard dialect, formal register, natural language | Same as Task 1 | Same as Task 1 |
| Topical characteristic | applied linguistics related topics | Same as Task 1 | Same as Task 1 |
| **Relationship between input and response** | | | |
| Reactivity | Non-reciprocal | Same as Task 1 | Same as Task 1 |
| Scope of relationship | Broad | Same as Task 1 | Same as Task 1 |
| Directness of relationship | Indirect | Same as Task 1 | Same as Task 1 |

### *6.3.2 Construct definition*

Just like our ESL context, our definition of the construct to be measured here is theory based. Considering the test purpose, i.e., making inferences about a test-taker's ability to

write academic papers and proposals, Table 6.5 summarizes the components of the language abilities to be measured by the test.

**Table 6.5:**
**Constructs to be Measured in the Advanced EFL Context**

| | |
|---|---|
| **Language knowledge** | |
| Grammatical knowledge | Ability to draw upon both syntactic and vocabulary knowledge simultaneously for the purpose of processing and producing well-formed comprehensible sentences: knowledge of grammatical structures, accurate use of them for the purpose of communication; knowledge of general and specialized vocabulary; knowledge of the proper use of punctuation |
| Textual knowledge | Ability to organize sentences to form a text: knowledge of cohesive devices used to mark the relationships; knowledge of common methods for organizing thoughts |
| Functional knowledge | Ability to create a piece of writing using different functions common to academic texts: 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 standard dialect, formal register |
| **Strategic competence** | Ability to set communicative goals and make plans |
| **Topical knowledge** | Ability to express the knowledge of applied linguistic issues through language |

Unlike the ESL context, topical knowledge has been included as part of the construct definition here. This is mainly because the purpose of the test is to make inferences about test-takers' ability to use language in writing with reference to applied linguistic issues. Besides, all the members of the population to be tested in this context specialize in the same area and thus have homogeneous topical knowledge.

As for strategic competence, it is included in the above construct definition, but at a level of specification different from that of the ESL context. Because of the unspeeded, less interactive nature of the test of writing (compared with that of speaking) and the test-takers' opportunity to use a dictionary during the test, the students' ability to use communication strategies at the lexical level is not of much interest in this context. On the other hand, respecting the fact that their ability in academic writing is being tested, it is important that they can set communicative goals and make plans for the better expression of their academic knowledge. This is why strategic competence as a component of the above construct is defined more specifically than that of the speaking test.

### *6.3.3 Test tasks*

Just like our ESL language use context, here too the students may use a variety of writing tasks – like note-taking or letter writing – that are not representative of academic writing. Because of this, we have limited our specification of the non-test language use tasks to the three basic writing tasks described in Table 6.4. Yet again, for practical considerations, we cannot include all three tasks as test tasks. So, considering the similarity between the characteristics of the three tasks, Task Three (writing proposals) has been chosen as a basis for describing the test task whose specifications are summarized in Table 6.6 below. It is very similar in nature to Task One (writing term papers) except that it is shorter and more suitable for a testing context.

**Table 6.6:**
**Characteristics of the Advanced EFL Test Task**

| | |
|---|---|
| **Characteristics of the setting** | |
| Physical characteristics | Location: on-campus classroom, well-lit, comfortable temperature<br>Noise level: quiet<br>Materials and equipment: English-English dictionaries, booklets, pens, pencils.<br>Degree of familiarity: everything is familiar to test-takers |
| Participants | Course instructor |
| Time of task | Exam period at the end of fall/spring semester; weekdays, working hours |
| **Characteristics of the input** | |
| Format | |
| Channel | Visual (written) |
| Form | Language |
| Language | English |
| Length | Short |
| Type | Complex prompt on a technical topic providing a context for the writing task, simple instructions for writing the essay |
| Speededness | Unspeeded |
| Vehicle | Mainly reproduced |
| Language of input | |
| Grammatical | General and technical vocabulary, varied grammatical structures |
| Textual | Sentences within the prompt properly linked and organized |
| Functional | Ideational, manipulative |
| Sociolinguistic | Standard dialect, formal register |
| Topical characteristics | Prompt related to topics in applied linguistics |
| **Characteristics of the expected response** | |
| Format | |
| Channel | Visual (written) |
| Form | Language |
| Language | English |
| Length | Moderate |
| Type | Extended production response |
| Speededness | Moderately speeded |
| Language of expected response | |

| | |
|---|---|
| Grammatical | General and technical (Applied Linguistic) vocabulary, varied grammatical structures |
| Textual | Cohesive, well-organized piece of written text |
| Functional | Ideational, manipulative (instrumental), heuristic |
| Sociolinguistic characteristics | Standard dialect, formal register, natural language |
| Topical characteristics | Topic selected by the examiner, has to be related to applied linguistics |
| **Relationship between input and response** | |
| Reactivity | Non-reciprocal |
| Scope of relationship | Broad |
| Directness of relationship | Indirect |

A comparison of the test task characteristics in the above table with those of real-life Task Three in Table 6.4 reveals the extreme similarity between them, which in turn implies the authenticity of the test task. The test is going to consist of a single written task (Appendix Two, Part A) to be completed in 120 minutes. Since (for the reasons discussed in section 6.3.2) topical knowledge is part of the construct definition of this test, all the test-takers are supposed to perform in response to the same prompt assigned by the examiner and representing the type of topics normally discussed in applied linguistic texts.

Just like the test in the ESL context, the scoring of this test is also affected by and based on the components of the construct definition, that is, the rating is done componentially. The rating instrument thus includes as many components as the construct definition itself, and the students' performance on each component will be rated on the basis of a 5-point ability scale ranging from 0 (none) to 5 (mastery).

For a detailed account of the test, its rating instrument, the rating scale, and a description of the ability components included in the rating instrument, see Appendix Two.

## *6.4 EFL context: Intermediate group*

### *6.4.1 Non-test language use tasks*

In this section we turn to the third and last language learning context for which we are designing a test. As discussed in Chapter Five, the test is intended for undergraduate students majoring in English Language & Literature, and Translation. The purpose of the test – to be administered at the end of an essay-writing program – is to decide whether or not the students are ready to undertake the written activities they are expected to perform during their undergraduate years. The procedures of test development are identical to those followed for the above two contexts. In this case, however, we are dealing with a different set of test specifications due to a change in the context and purpose of the test. Table 6.7 summarizes the characteristics of the language use tasks in the intermediate EFL context upon which the characteristics of the test tasks are based.

**Table 6.7:**
**Characteristics of the TLU Tasks in the Intermediate EFL Context**

| | TLU Task 1 | TLU Task 2 | TLU Task 3 |
|---|---|---|---|
| **Characteristics of the setting** | writing term papers on literary topics | writing essay exams | writing reports and essays on various topics |
| Physical characteristics | Location: home, library<br>Noise level:mostly quiet<br>Temperature and humidity: comfortable<br>Materials and equipment and degree of familiarity: books, dictionaries, notes, journals, pen/pencil, paper, word processors, | Location: class-room.<br>Noise level: quiet<br>Temperature and humidity: comfortable<br>Material and Equipment: same as Task 1. | Same as Task 1 |

| | | | |
|---|---|---|---|
| | all familiar to the test-takers. | | |
| Participants | Course instructor, librarians, classmates | Course instructor | Same as Task 1 |
| Time of task | Any time | Weekdays, daytime, evenings | Same as Task 1 |
| **Characteristics of the input** | | | |
| **Format** | | | |
| Channel | Both visual and oral | Same as Task 1 | Same as Task 1 |
| Form | Language | Same as Task 1 | Same as Task 1 |
| Language | Target (English) | Same as Task 1 | Same as Task 1 |
| Length | Varied including very short topics or longer oral or written prompts | Short prompts such as questions | Short topics for essays, but longer written or oral prompts for reports |
| Type | Prompt and task | Same as Task 1 | Same as Task 1 |
| Speededness | Unspeeded | Moderately speeded | Same as Task 1 |
| Vehicle | Live and reproduced | Reproduced | Same as Task 1 |
| **Language of input** | | | |
| Organizational characteristics | | | |
| Grammatical | Both literary and general vocabulary, widely varied grammatical structures, typewritten | Varied vocabulary including technical (literary, and applied linguistic) and general, widely varied grammatical structures, type-written | Same as Task 2 |
| Textual | All sorts of linking devices and organization patterns | Same as Task 1 | Same as Task 1 |
| Pragmatic characteristics | | | |
| Functional | Ideational,manipulative, imaginative | Same as Task 1 | Same as Task 1 |
| Sociolinguistic | Standard dialect, Register: moderately formal, natural language | Same as Task 1 | Same as Task 1 |
| Topical characteristics | Literary topics | Varied | Same as Task 2 |
| **Characteristics of the expected response** | | | |
| **Format** | | | |
| Channel | Visual (written) | Same as Task 1 | Same as Task 1 |
| Form | Language | Same as Task 1 | Same as Task 1 |
| Language | Target (English) | Same as Task 1 | Same as Task 1 |
| Length | Maximum 10 pages | Short (1-3 paragraph) | Same as Task 1 for reports, medium length for essays |
| Type | Extended production | Same as Task 1 | Same as Task 1 |

| | | | |
|---|---|---|---|
| | response | | |
| Speededness | Unspeeded | Relatively speeded | Same as Task 1 |
| **Language of expected response** | | | |
| Organizational characteristics | | | |
| Grammatical | Vocabulary: general and literary<br>Varied grammatical structures, type-written | Same as Task 1, hand-written | Same as Task 1 |
| Textual | Cohesive written text presenting well-organized pieces of information all contributing to a topic, use of common methods of development, subtitles often used, with thesis usually stated at the beginning | Cohesive written text usually arguing for or against a theory, explaining/defining a concept, or reviewing some academic work | Same as Task 1 |
| Pragmatic characteristics | | | |
| Functional | Ideational, manipulative, and imaginative | Same as Task 1 | Same as Task 1 |
| Sociolinguistic | Standard dialect, relatively formal register, natural language | Same as Task 1 | Same as Task 1 |
| Topical characteristic | Literary topics | Varied | Varied |
| **Relationship between input and response** | | | |
| Reactivity | Non-reciprocal | Same as Task 1 | Same as Task 1 |
| Scope of relationship | Broad | Same as Task 1 | Same as Task 1 |
| Directness of relationship | Indirect | Same as Task 1 | Same as Task 1 |

### *6.4.2 Construct definition*

Based on the test purposes, the components of the language ability to be measured by this third test are summarized in the following table. Unlike Test 2 above, in this context, topical knowledge and strategic competence are not included as part of the construct definition. This is because test-takers are not exactly homogeneous in terms of their

topical knowledge, and test scores are supposed to be used for making inferences about test-taker's language ability regardless of the topic they are writing on. Similarly, considering the fact that the students are allowed to consult dictionaries during the test, no specific inferences are intended to be made with respect to their ability to use communication strategies or other aspects of strategic competence.

**Table 6.8:**
**Constructs to be Measured in the Intermediate EFL Context**

| | |
|---|---|
| **Language knowledge** | |
| Grammatical knowledge | Ability to draw upon both syntactic and vocabulary knowledge simultaneously for the purpose of processing and producing well-formed comprehensible sentences: knowledge of grammatical structures, accurate use of them for the purpose of communication; knowledge of general and specialized vocabulary; knowledge of the proper use of punctuation |
| Textual knowledge | Ability to organize sentences to form a text: knowledge of cohesive devices used to mark the relationships; knowledge of common methods for organizing thoughts |
| Functional knowledge | Ability to create a piece of writing using different functions: 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 standard dialect, moderately formal register |
| **Strategic competence** | Not included |
| **Topical knowledge** | Not included |

### *6.4.3 Test tasks*

Following the same procedure that we adopted for determining the test tasks for ESL and advanced EFL tests, we decided that Task 3 (writing essays) in Table 6.7 be included as the test task since it shares almost all of the characteristics of the other two tasks and is of a suitable length for a test task.

**Table 6.9:**
**Characteristics of the Intermediate EFL Test Task**

| | |
|---|---|
| **Characteristics of the setting** | |
| Physical characteristics | Location: on-campus classroom, well-lit, comfortable temperature<br>Noise level: quiet<br>Materials and equipment: English-English dictionaries, booklets, pens, pencils.<br>Degree of familiarity: everything is familiar to test-takers |
| Participants | Course instructor |
| Time of task | Exam period at the end of fall/spring semester; weekdays, working hours |
| **Characteristics of the input** | |
| Format | |
| Channel | Visual (written) |
| Form | Language |
| Language | English |
| Length | Short |
| Type | Short prompt providing one or more topics for the writing task and simple instructions for writing the essay including the method to be followed |
| Speededness | Unspeeded |
| Vehicle | Mainly reproduced |
| Language of input | |
| Grammatical | General vocabulary, varied grammatical structures |
| Textual | Sentences within the prompt properly linked and organized |
| Functional | Ideational, manipulative |
| Sociolinguistic | Standard dialect, relatively formal register |
| Topical characteristics | One/or more prompts for written essay |

| | |
|---|---|
| **Characteristics of the expected response** | |
| Format | |
| Channel | Visual (written) |
| Form | Language |
| Language | English |
| Length | Medium |
| Type | Extended production response |
| Speededness | Moderately speeded |
| Language of expected response | |
| Grammatical | General vocabulary, varied grammatical structures |
| Textual | Cohesive, well-organized piece of written text |
| Functional | Ideational, manipulative, heuristic |
| Sociolinguistic characteristics | Standard dialect, relatively formal register, natural language |
| Topical characteristics | If given a choice, test-takers choose from among several prompts given |
| **Relationship between input and response** | |
| Reactivity | Non-reciprocal |
| Scope of relationship | Broad |
| Directness of relationship | Indirect |

The test is going to consist of a single written task (Appendix Three, Part A) to be completed in 120 minutes. Since topical knowledge is not part of the construct definition of this test, test-takers will be asked to perform on a topic which does not require any specialized knowledge. They might also be given a chance to choose from among a few topics assigned by the examiner.

Just like the other two tests, the scoring of this test is also affected by and based on the components of the construct definition; that is, the rating is done componentially. The rating instrument thus includes as many components as the construct definition itself and the students' performance on each component will be rated on the basis of a 5-point ability scale ranging from 0 (none) to 5 (mastery).

For a detailed account of the test, its rating instrument, the rating scale, and a description of the ability components included in the rating instrument, see Appendix Three.

# CHAPTER SEVEN
# THE EXPERIMENT

## *7.1 Objectives*

As part of the research for examining our proposed theoretical framework, in the third phase of the study, an experiment was conducted in both EFL and ESL contexts (see 5.2 for a description of the contexts of the study) to examine empirically the relationship between the tests developed in Chapter Six on the one hand, and the teaching and learning activities taking place in the corresponding English language classrooms on the other hand. It was primarily intended to find out whether the tests developed on the basis of the learners' and educational systems' needs were having the washback effects hypothesized by the theoretical framework presented in this study or not. And if so, to what extent the tests could be held responsible for this outcome.

The specific questions addressed were, therefore, those regarding the presence of washback, i.e., whether or not the tests had any effects whatsoever on the language program and if so, what form(s) washback would take. Would it appear in the form of the material reflecting major/minor aspects of the course objectives, in/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 educational system?

## *7.2 Subjects*

Three groups of non-native speaking populations with different needs and proficiency levels, and in different learning environments participated in the study.

Group one (G1) originally consisted of 47 male and female foreign graduate students functioning as teaching assistants (known as International Teaching Assistants or ITAs) in an ESL environment at the University of Victoria, Canada. They constitute a group of language learners who need to communicate ideas orally at an advanced level with native speakers of English, but whose TOEFL scores (that they are required to provide at the time of admission) do not necessarily reflect their oral communicative abilities. These students, selected from among those referred to us by the graduate advisors in the corresponding departments, came from different language backgrounds. They were nationals of China, Hong Kong, Japan, Taiwan, India, Germany, Iran, Russia, Ghana, and Pakistan, and their areas of specialization included visual arts, biochemistry and microbiology, physics, chemistry, biology, computer science, health information science, mechanical engineering, computer and electrical engineering, geography, sociology, and linguistics. They were at an advanced level of proficiency with a TOEFL score of minimum 550[1] and an average age of 32.

Group Two (G2) consisted of 42 male and female graduate students in an EFL environment at Free University, Iran. Unlike the subjects in group one, these students were all majoring in Applied Linguistics and shared the same native language. They were at an advanced level of English language proficiency – with a score equivalent to TOEFL 550 – determined by the results of the English language test they had taken at the time of admission.

---

[1] This is the minimum TOEFL score required by the University of Victoria for admission to the Faculty of Graduate Studies.

The third group of subjects (G3), were 58 male and female undergraduate EFL students majoring in English and Translation at Allameh Tabatabai University, Iran. Based on the results of the English language component of the University Entrance Examination, they were at an intermediate level of proficiency – with a score equivalent to TOEFL 500. They all shared the same native language and were of an average age of 22.

## *7.3 Methods and procedures*

The data for the experiment were gathered in several stages. Quantitative research methods were used to find out whether or not by the end of the course learning had taken place, and what ability areas had been most affected, positively or negatively. Qualitative research methods, on the other hand, were used to determine the teaching and learning strategies adopted by the teachers and the learners and, more importantly, if they were in any way related to or affected by the tests. The procedures and instruments used in this experiment are as follows:

Because the University of Victoria admits ITAs on the basis of their TOEFL scores and does not require them to present proof of their spoken language abilities, in order to have a homogeneous sample, the first step was to determine that the subjects had the general oral English language proficiency required for the TA program. So, the SPEAK (Spoken Proficiency English Assessment Kit) test[2], a context-free standardized

[2] SPEAK is the institutional version of the TSE (Test of Spoken English) and is usually rated by trained raters at the institution administering the test. TSE is the most commonly used measure of the spoken ability by the universities that have TA programs (Bauer and Tanner, 1994), however, both TSE and SPEAK are considered as indirect measures of communicative ability since they are tape-recorded tests in which the examinee's responses are also tape-recorded. Educational Testing Service (ETS) recommends that TSE scores should not be considered as the only measure for evaluating ITAs and that other relevant information should also be taken into consideration (1990).

test, was administered to all 47 of them about two weeks before the start of the TA program. It was used as a screening device with a passing score of 220 out of 300[3].

In the next stage, about a week after the administration of the SPEAK test, 28 ITAs who had passed the test were required to take the performance test of spoken language ability developed for the purposes of this study. There were two main reasons for this: (i) to exclude from the program those candidates who already possessed the abilities measured by the test (5 were excluded), and (ii) to have a set of scores for the candidates who were going to take the course for the purpose of comparison with their end-of-term scores on the same test. A similar procedure was also conducted in the EFL context. Both intermediate and advanced tests of writing ability – discussed and developed in Chapter Six – were administered to the subjects before and after the corresponding writing courses to see if any learning had taken place as a result of the program.

In order to be able to attribute unambiguously any improvement in the subjects' performance at the end of the term to the type of training they had been subjected to, the subjects in each group were randomly divided into experimental and control groups. The members of the control groups had thus the same level of English language proficiency and readiness as those in the experimental groups before the start of the courses. The subjects in the ESL control group (n=9) were chosen from among the ITAs who had been screened by the SPEAK test and had taken the pre-test. However, partly because of the limitations in the class size and partly because of the conflict of the class time with their schedules, they were asked to come back in four months when they would take the post-

---

[3] There are no passing/failing scores on TSE. Institutions using TSE set their own standards depending on their purposes. In this case we set the cut-off score at 60% acceptance level which according to the Manual for Score Users (1982, 1992) is equivalent to 220 on TSE.

test. In the meantime, they engaged only in their routine graduate programs and did not take any formal English language training. It was thus intended to see if simply by living and working in an ESL environment they would show any improvement in the spoken language ability areas required for functioning as a TA and measured by the oral performance test or not. EFL control groups, on the other hand, were functioning in a context where everybody else spoke in the native language outside the program. The EFL advanced control group (n=20) was engaged in the graduate program in applied linguistics while the EFL intermediate control group (n=28) was taking courses that focused on language skills other than writing ability. In short, other than the fact that the experimental groups underwent the one-semester-long courses in English, they were otherwise treated the same as their corresponding control groups.

With respect to the administration and scoring of the tests, different procedures were followed for different tests. The SPEAK tests were administered to all examinees simultaneously in the language listening laboratories. Their performance 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 ESL oral performance tests of the same subjects, on the other hand, were administered over a period of one week and were rated by a panel of raters comprised of two ESL instructors and three native-speaking undergraduate students introduced to us by the graduate advisors in the departments to which ITAs belonged. The undergraduate students' participation in the testing session was an important consideration in the design of the oral test since it would add to the authenticity of the test task and the testing context by providing an atmosphere similar to that in the real life situation. Their involvement during the presentation and

question-answer phases of the test would generate a lot of spontaneous speech on the part of the examinee from which his/her level of comprehensibility and success in communication could be assessed. To make sure that the raters understood the rating procedure and the areas of ability they should look for in the performance of the test-takers, the researcher met with the ESL instructors and potential undergraduate participants from each department, explained the rating procedures, and provided them with copies of the rating instrument, a description of the ability components in the rating instrument, and the rating scale (see Appendix One for details) days before the administration of the test. The performance of each TA was rated either during his/her performance or shortly after it was over. Nevertheless, due to the transient nature of the oral production, the entire testing session was videotaped in case the raters missed some parts of the production or major disagreements were later found in their ratings of the same individual. As for the EFL writing performance tests for both advanced and intermediate groups, the tests for each group were administered during the same session before and after the corresponding writing courses. They were rated later on by two ESL/EFL instructors using the rating instruments, descriptions of the ability components in the rating instrument, and the rating scales (see Chapter Six, Appendixes Two and Three respectively).

The results of the quantitative data thus gathered would shed light on the degree of learning taking place in the areas of knowledge specified by the test. However, to determine whether or not the tests had any effect(s) on the teachers' syllabi, their approach to the language course, their teaching methodology, their choice of the materials, and the type of in-class and out-of-class activities, qualitative research

instruments such as interviews and questionnaires with open-ended questions were adopted. Before and during the course of the experiment, teachers were interviewed for their opinions regarding the tests, their component abilities, their rating instruments, the time allotment to the administration and rating of the tests, the preparation of the students for the test, the activities that they did because of the test and the things that they would or would not do if they did not have to adhere to the test. Besides teachers, both groups of ESL/EFL instructors as well as undergraduate raters were interviewed for their reactions towards tests and their rating instruments. Moreover, all ESL subjects as well as randomly selected EFL subjects[4] in both the experimental and the control groups were interviewed for their reaction towards the performance tests and the program they had participated in, the content of the tests, its relevance to what they were doing in class, the objectives of the course, and the material being used. ITAs were also asked about their attitudes towards the test and how and why they liked it in comparison with the SPEAK test. Class observations were also conducted directly by the researcher or indirectly through independent observers. The information gathered from the observations was intended to supplement the opinions of teachers and students regarding the teaching and learning activities, especially those with respect to and revealing the direct and/or indirect effects of the tests. Finally, it should be emphasized that none of the participants knew what the hypothesis of the experiment was and everyone was blind to the expected outcome.

---

[4] Due to the accessibility limitations, only those EFL subjects for whom an email address was available and those I visited during my short stay in Tehran during the summer of 1997 were consulted.

In short, in this investigation of washback phenomenon, the qualitative and quantitative data obtained through all these methods were expected to complement each other in an attempt to provide a comprehensive view of how direct authentic tests operate in the context of foreign/second language teaching as well as the breadth and depth of the effect(s) that such tests can have on different aspects of the teaching and learning activities taking place in a language classroom.

## *7.4 Reliability of the tests*

Reliability in general refers to the degree to which scores are reliable indicators of the ability being tested. Based on the performance of the students on the pre-test administration of the tests, reliability analyses were conducted to determine the sources of inconsistencies, if any, within and among the scores. We were concerned especially with (i) the consistency of items within each category, (ii) the consistency of categories within each test, and (iii) the consistency among raters. We, therefore, estimated the internal consistency of the items and categories and the interrater reliability for each test using the "coefficient alpha" formula. The results are reported below for each test.

**Table 7.1:**
**Reliability Coefficients for the ESL Test***

| | |
|---|---|
| **Within-Category Reliabilities**** | |
| Grammatical Knowledge | $\alpha$= .8811 (No. of items=3) |
| Textual Knowledge | $\alpha$= .8945 (No. of items=2) |
| Sociolinguistic Knowledge | $\alpha$= .5780 (No. of items=2) |
| Strategic Competence | $\alpha$= .8924 (No. of items=5) |
| **Overall Test Reliability** | $\alpha$= .9083 (No. of categories=6) |
| **Correlation Between Raters** | r = .9131 (No. of raters=5) |

* No. of cases = 22

** The two categories of Functional Knowledge and Overall Performance are not included since they consist of only one item.

**Table 7.2:**
**Reliability Coefficients for the Advanced EFL Test***

| | |
|---|---|
| **Within-Category Reliabilities**** | |
| Grammatical Knowledge | $\alpha$= .8979 (No. of items=2) |
| Textual Knowledge | $\alpha$= .8203 (No. of items=2) |
| Sociolinguistic Knowledge | $\alpha$= .9500 (No. of items=2) |
| Strategic Competence | $\alpha$= .7792 (No. of items=2) |
| **Overall Test Reliability** | $\alpha$= .9029 (No. of categories=7) |
| **Correlation Between Raters** | r = .8042 (No. of raters=2) |

* No. of cases = 42

** The three categories of Functional Knowledge and Topical Knowledge and Overall Performance are not included since they consist of only one item.

**Table 7.3:**
**Reliability Coefficients for the Intermediate EFL Test***

| | |
|---|---|
| **Within-Category Reliabilities**** | |
| Grammatical Knowledge | $\alpha$= .9000 (No. of items=2) |
| Textual Knowledge | $\alpha$= .9241 (No. of items=2) |
| Sociolinguistic Knowledge | $\alpha$= .8014 (No. of items=2) |
| **Overall Test Reliability** | $\alpha$= .9484 (No. of categories=5) |
| **Correlation Between Raters** | r = .8095 (No. of raters=2) |

* No. of cases = 58

** The two categories of Functional Knowledge and Overall Performance are not included since they consist of only one item.

As can be seen from the tables above, the analysis of the students' scores on the test items indicated a high level of reliability both for the items within each category and for the categories within each test, indicating that the items and categories in each test measure the same ability. This internal consistency further suggests that the constructs underlying the components of each category as well as those underlying each test's choice of categories are well-defined.

Besides, the high inter-rater reliabilities indicate that the tests' detailed rating scales and the descriptions of the components included in the rating instrument were simple,

clear, and specific enough to prevent raters from subjectively scoring the students' performances.

# CHAPTER EIGHT
# WASHBACK: THE EFFECTS OF TESTS ON TEACHING AND LEARNING

## *8.1 Introduction*

This chapter presents the study's findings on the washback effects of the tests developed based on the academic needs of the learners for whom they were intended and the objectives of the educational systems in which they were used. The impact is described here in terms of the choice of materials, the teaching methodology, and the learning strategies and outcomes – the three major areas of the general model of washback introduced in Chapter Three.

It is worth repeating here that the tests, designed for one ESL context and two different EFL contexts, differed in major ways: they measured different language abilities at different levels of proficiency. Besides, while the ESL test was a high-stakes placement test scored by a panel of raters, the EFL tests were both achievement tests scored by the teachers themselves[1]. These variables as well as such interfering factors as the learners' topical knowledge and experience with the target language were properly addressed in the design of each test. Still, the variation in the choice of the subjects and the testing contexts helps us to better understand how washback works and to what extent a test might be responsible for teaching and learning consequences. Here, the success or failure of the tests in bringing about desirable consequences was sought in the choice of materials for use inside (and outside) the classroom, in the teachers' attitudes towards the test, in the

[1] However, to ensure reliability both EFL exams were also scored by two raters.

methodology adopted by the teachers in the classroom, in the learners' behaviour, their achievement of the course content, and the uses to which they could put their language outside the classroom. The results are reported separately for each context.

## *8.2 ESL context*

### *8.2.1 Materials*

The main text chosen for the ITA training course was Smith, Meyers & Burkhalter (1992). The book consists of ten units, each of which centres around one of the most common teaching tasks in university classrooms. Language skills and grammar sections in each unit discuss the problems common among ITAs and are related to the rhetorical teaching task of the unit. The section on cultural awareness, on the other hand, does not present any information but refers to the immediate subculture in which ITAs find themselves and helps them to examine their own values and beliefs about teaching against those of their fellow native speaking assistants, professors, and undergraduate students, and thus assess for themselves what they need to learn about the culture in which they will be working. The book, therefore, assumes that in addition to language proficiency, ITAs need to develop teaching skills as well as an awareness of cultural differences. Another piece of material chosen to supplement the main text for this course was a videotape (Douglas & Myers, 1989) intended to develop communication strategies for coping with the teaching needs of the teaching assistants.

To examine the extent to which the materials chosen for the course could potentially promote the abilities measured by the test, we will evaluate the textbook in terms of the focal constructs underlying the choice of test tasks. We will analyze the

structure of the units in the textbook to see how the constructs represented as performance categories in the rating instrument are promoted and developed by the book.

The ten units in the book are centred around common rhetorical teaching tasks in university classrooms: introducing a syllabus, explaining a visual, defining a term, teaching a process, fielding questions, explaining something at a basic level, presenting information over several class periods, and leading a discussion. The main body of each unit consists of related sub-sections addressing the language used for teaching as well as the cultural issues.

The *focus* sections introduce and provide practice with specific teaching skills. The main emphasis in this section is on the development and practice of the language abilities related to teaching, cognitive strategies, as well as the areas of textual knowledge corresponding to the performance categories categorized under textual, functional and strategic knowledge in our rating instrument (see Appendix One for a description of these categories). For example, Focus 2 in Unit Three basically encourages the abilities listed under textual knowledge in the rating instrument. Here, the students are taught to present their intended meaning in an organized manner, the common methods of organization employed in both written and oral forms are introduced, and exercises are presented to help ITAs practice the use of each method.

Similarly, the *functional language* sections present and provide practice with general vocabulary and expressions necessary to perform the teaching task of each unit. For instance, this section in Unit 4 presents organizational cues that ITAs need to produce texts that attain the level of cohesion and coherence necessary for easy interpretation of

the teaching task already discussed in the previous unit. Therefore, common words and phrases used to signal ideas in presentations are presented and practiced in this section.

The *assignment preparation* and *assignment presentation* sections guide ITAs in preparing short presentations related to the teaching task around which each unit is centred. ITAs are supposed to employ the functional language they have already learned, and learn from and support each other as they develop their classroom communication skills.

The section on *language skills* includes work on the development of grammar, vocabulary and pronunciation. It deals with the grammatical and pronunciation problems common to ITAs in general and those related to the rhetorical teaching task of the unit in particular. The grammar section in Unit 5, for example, teaches the verb form and the sentence structure usually used in classroom or lab sessions for giving instructions. In Unit 4, for example, relative clauses are introduced in relation to a defining task.

For pronunciation, too, the emphasis is mostly on the problems of stress, rhythm, and intonation that are common to ITAs of most language backgrounds rather than changing the pronunciation which requires long-term training. Consequently, areas of utmost importance for the teaching task such as useful communication strategies – the employment of which can make ITAs' speech more comprehensible even if they still have pronunciation problems – and regular patterns of stress are introduced and practiced. In short, the sections on language skills in each unit serve to develop not only the categories listed under grammatical knowledge in our rating instrument, but also those of strategic knowledge.

The final section of each unit refers to the immediate sub-culture in which ITAs have to function, i.e., the university and their specific department at the university. This section helps ITAs to collect and analyze cultural information about the environment in which they are working and to examine it against their own cultural values. The candidates are supposed to work on these sections themselves, while at the same time issues of general interest can be discussed in class sessions with the course instructor participating as a consultant.

Besides, since almost all ITAs show gaps in the areas of linguistic competence, this component is a crucial one in ITA training programs. Through teaching them how to use both verbal and nonverbal strategies in order to make up for their knowledge or abilities that are weak in other areas, we can help them learn to use compensatory techniques to increase their communicative effectiveness. To this end, the textbook was supplemented with a videotape (Douglas & Myers, 1990) exclusively designed "for training ITAs in definable skills in their various technical areas, while at the same time developing communication strategies for coping with a range of situations that cannot be predicted in detail" (p. 169).

On the whole, the design of the textbook and of the supplementary videotape potentially allow the user to develop proficiency in almost all areas of the focal construct of our performance test. Just like the performance test, the textbook does not assume the knowledge of a particular field of specialization. Communication strategies which are of particular significance to ITAs' communicative success are encouraged and practiced by the text. There are, however, areas such as cultural sensitivity and teaching skills that are equally emphasized by the text, although the test is primarily concerned with language

abilities. This could potentially lead to an emphasis on topics irrelevant to the test constructs and thus under-representation of the focal constructs if the teacher felt constrained to cover everything in the book. Given this situation, an examination of the materials alone so far clarifies that the materials chosen for the course could potentially reflect, promote and practice, among other things, the areas of language ability measured by the test. However, in order to find out the extent to which the test and its purposes have affected the content of teaching and learning activities inside and outside the classroom, we need further evidence based on teachers' and learners' comments and classroom observations.

### *8.2.2 Teaching and methodology*

The information necessary to investigate the effect(s) of the exam on the content of the lessons, the methodology the teacher used, and the attitudes of the teachers and raters was collected over time – before, during and after the training program – using different qualitative approaches the nature and results of which are summarized below.

***The background study***

In the spring and summer of 1996, almost a year before the beginning of the course, a preliminary survey was carried out in order to find out what the ITA assessment and teaching program was like before the introduction of the new exam. This included holding several group and individual meetings with the stakeholders at the Faculty of Graduate Studies and the staff at the English Language Centre (ELC) responsible for the direction, coordination and teaching of ESL courses as well as for the analysis of the test, the syllabus and the materials (if any) used for ITA training in the past. The information and

the material provided during these interviews were all recorded and transcribed for further analysis by the researcher.

The results of this aspect of the data revealed that at the time the Faculty of Graduate Studies, in consultation with departmental graduate advisors, assessed the English language proficiency skills of international graduate students based on their TOEFL scores and educational background when they applied for admission and positions as TAs. However, all parties in the Faculty of Graduate Studies and the ELC believed that the various assessment procedures that were used to identify graduate candidates with acceptable/unacceptable English language proficiency skills were not effective in determining whether graduate students with low or marginal proficiency in spoken English should be appointed as TAs. Neither did the speaking test used by the ELC adequately measure the spoken language ability of those ITAs who were sent to the ELC at the recommendation of their departments/supervisors or of their own free will.

The test was an in-house test intended as an exit criterion for ITA programs and was made up of three sections (interview, narration, and discussion) administered to students in pairs. In the interview section, the students are 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 are required to tell a story based on a series of connected pictures. The discussion part engages pairs of students in a three-minute discussion of a topic such as "Do you think there is too much violence on T.V.?" or "Do you support capital punishment?" The students are then rated on a 0-6 scale for pronunciation, grammar, vocabulary, and fluency. While grammar, fluency, and vocabulary are each accorded 30% towards the final total, pronunciation is

given 10%. The test kit does not provide any information regarding the rationale underlying the choice of the test tasks, the time allocated to each task, or the theoretical basis of the rating scale. Furthermore, the test tasks and topics are in no way related to those of the real-life contexts in which the students are expected to function as teaching assistants. The test thus seems inadequate for measuring the oral language abilities of this group of learners.

Consequently, ITA courses offered in the past 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 necessarily cover the needs of the ITAs. One of the ESL teachers said that she would rather base her teaching on "improving the teaching abilities of the TAs" than the test. 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 and participation in class activities. Besides, no specific material had been chosen or generated to promote the objectives of either the test or the ITAs. In general, teachers and program directors believed that the test was neither authentic nor valid for use in ITA programs. It was then quite clear that the old test had adversely affected teaching activities, in that it had neither promoted teaching nor material development. Instead, in the absence of a meaningful assessment device, the teachers had used whatever methodology they were convinced would best prepare their students for the instructional settings they were supposed to function in. The Faculty of Graduate Studies and the ELC were thus convinced that what was needed was an assessment device to serve as an entry and exit criterion for ITA training programs.

***Reaction to the new test***

Altogether, two teachers[2] and 15 raters (from different departments) were involved in the teaching, administration, and scoring of the test. The teacher whose class was being studied was also involved in the administration and scoring of the test. This enhanced familiarity with the test could affect not only what was taught, but also the teaching strategies used in the classroom. The teachers' and raters' reactions to the test were examined through interviews, observation of class activities, observation of the rating process, analysis of actual teaching content, and the students' input. The results are summarized in the rest of this section in the order they were received.

1) A preliminary survey was carried out among 17 raters about a month before the training program began. The two ESL teachers were interviewed and observed for their reaction towards the administration and scoring of both the SPEAK and the performance test, as were the undergraduate raters. The following table shows their original reaction to the different aspects of the performance test and the testing process:

**Table 8.1:**
**Raters' Reaction to the Performance Test**

| | yes | no |
|---|---|---|
| Raters understood all the ability components of the rating instrument | 76% | 24% |
| Raters regarded the performance categories as adequate for measuring ITAs' spoken language ability | 59% | 41% |
| Raters believed that the test was a practical one | 88% | 12% |
| Raters believed that the 0-4 rating scale was reasonable and clear | 71% | 29% |
| Raters regarded the test task as closely related to the real-life tasks | 88% | 12% |
| Raters believed that the test content would motivate ITAs to improve their spoken English | 94% | 6% |
| Raters thought that on-the-spot scoring was practical | 76% | 24% |
| Raters needed to go over the video-tape again | 0 | 100% |

[2] These same ESL teachers had been trained for and scored the SPEAK test used as the initial screening device for the ITA program.

When asked about the adequacy of the performance categories, a number of raters (41%) including the teacher of the course indicated that they believed that the ability components measuring teaching skills should have been included in the rating instrument[3, 4]. It can also be seen from the table that the majority of raters believed that the test and its scoring system were practical. This result was backed by the researcher's observation of the whole rating process during which it was noticed that all raters managed to score the test confidently during or immediately after the examinee's presentation. Despite the fact that the testing sessions were videotaped, none of the raters felt the need to review the students' performance on video[5]. It was noticed, however, that after the first few cases, once the raters became more adept at the process, they were able to go back and forth between different ability components and thus complete the rating instrument more quickly and simultaneously with the students' performance.

As for the authenticity of the tasks, 88% of the raters considered the test tasks and the testing environment as closely related to the actual settings in which ITAs have to function. Observations further revealed that the undergraduate raters became involved in genuine discussions with the ITAs in relation to the topics already presented, particularly

---

[3] McNamara (1996) distinguishes an individual's ability to use the target language from his ability to perform future job tasks successfully in that language by referring to the former as the *Weak Performance Hypothesis* and the latter as the *Strong Performance Hypothesis*.

[4] It should be noted at this point that just like the topical knowledge, teaching abilities are not part of the test's construct. Our test solely measures the spoken language abilities that the learners should possess to perform TA tasks, not those abilities needed to perform job-related tasks that are basically not related to language abilities. In other words, using this test, we are not evaluating ITAs for abilities that native-speaking TAs are not expected to possess. This was an important matter to be taken into consideration at the time of test development since it would have affected the test's validity, acceptability and fairness.

[5] All raters were paid on an hourly basis for the administration, scoring and, if necessary, reviewing the videotapes, so time allotment was not an issue here.

in the third part of the test (question and answer). This often provided excellent opportunities for other raters, especially ESL teachers, to better evaluate the various ability areas in the spontaneous speech of the examinees. The raters had often implied this in their written comments on individual examinees' performances:

> ... [the student] 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.
>
> ... got carried away with the subject while answering the question.
>
> ... wrote too much during the presentation, very little actual speaking until he had to answer questions.

At this stage, the ESL teachers who had also rated the performance of the same population on the SPEAK test, were also questioned for their reaction to the SPEAK test as compared with the performance test. In their verbatim answers to the questionnaire, they commented on different qualities of the test:

> This job has been extremely tedious and time-consuming. ... rating takes forever because of the numerous pauses in the tape and long introductions to each section. ... and you have to listen to each answer three times before you can decide it is incomprehensible.

> How can one score short correct answers relative to more complex ones?
>
> Sometimes the students avoided production because of their unfamiliarity with the topic being asked about, not their inability to speak in English ... this is not fair.

Given the time they had invested on the scoring of the SPEAK test, the raters decided that despite the ease of administration, rating the SPEAK test was much more time-consuming than the performance test.

2) In the next stage of studying the reaction of the different parties to the test, we examined the contents of teaching, and the teaching methodology. As stated in 8.2.1, the textbook chosen for the course had the potential to reinforce not only the constructs measured by the test but also other related areas such as cultural sensitivity and teaching techniques. Given this fact, the aim of this part of the study was to see whether the test was having the influence it was intended to have. Was the teacher using the textbook, the supplementary material and the *Instructor's Manual* to promote the language skills encouraged by the test, or simply going over the material from cover to cover? Either way, was the test in any way responsible for this?

The analysis of the field notes obtained from class observations and the after-class interviews with the students carried out during the first, fifth and eighth week of the program identified the following variables as the most important aspects of the classroom activities that could be traced to the test in one way or another:

- awareness of course objectives
- the material covered in the classroom

- the use of the textbook
- the use of the supplementary video
- the use of the extra material
- abilities mostly emphasized and practiced in the classroom
- time allotment to different activities
- students' participation in class activities
- class evaluations

We observed the first session of the course to find out how the objectives of the course were introduced and the context for subsequent classes was set up. There was a direct reference to the test at this stage when the instructor reminded the students that the final evaluation would be the repeat of the initial performance test. The students were given a copy of the test's rating instrument and the ability components were all defined, and, in some cases, exemplified for the students. The instructor then introduced the main text for the course along with a brief introduction of the contents of the text. At this point, however, the teacher made another direct reference to the test by emphasizing that only those parts corresponding to the abilities measured by the test would be covered and practiced in class, and that omissions, additions, and modifications would be introduced wherever appropriate in due time.

Later class observations revealed that the teacher actually did this by using the book as a resource for addressing the students' problems. Selected exercises and activities from each unit were routinely practiced in the class while some other parts had been either omitted or left to students to work on by themselves. Sections practicing common pronunciation problems, complex sentence structures, organizational methods, cohesive

devices and achievement strategies were the parts mostly discussed and practiced in class. Furthermore, supplementary material had also been introduced for both in-class and out-of-class practice in these areas. For example, while teaching the common methods of organization, the teacher provided the students with a list of common cohesive devices used most for each method. For further out-of-class practice in pronunciation and comprehension, she referred them to resources in the CALL Facility.

As for class activities, an average of four to five hours were spent per unit with almost two hours of group instruction (including work on the selected sections of each unit) and two to three hours of in-class teaching presentations performed by the students and immediately followed by teacher and peer feedback. The teacher specifically asked the students to try to apply what they had learned and practiced during group instruction (the functional language, commonly used grammatical patterns, organizational patterns and pronunciation strategies) to their presentations. The components of the rating instrument, not the *instructor feedback form* suggested by the textbook, were used as a basis for teacher feedback.

In the teacher's account of what was going on in the classroom, she emphasized that the class activities were mostly learner-centred and that the students differed in their teaching and presentation skills simply because some of them had had years of teaching experience in their home countries while some others were new to the practice. However, she agreed that they all needed to work on their language skills, especially the abilities related to pronunciation, organization, coherence and communicative strategies. 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 resources that addressed such problems more thoroughly. On the other hand, she had decided not to use the supplementary video for in-class instruction and practice because she thought that the video's examples of TAs' teaching skills were not didactic, and would not introduce anything more than the textbook did. The teacher was also very satisfied with the students' participation in the class discussions and reaction to the class activities. She said that towards the middle of the term the students started to engage seriously in class discussions, especially after the presentation sessions and during the instructor and peer feedback time. She described her role in those discussions as the provider of feedback and instruction with respect to those language areas that most interfered with the students' role as efficient TAs. To this end, and believing that TAs usually share similar language problems, she preferred to address the students' problems in the form of group instruction in the classroom rather than conducting tutorials for individual students. She especially emphasized that having the students practice their spoken language abilities by giving mini-lectures and by simulating teaching sessions had a significant effect on their progress and boosted their confidence as TAs. Nevertheless, she expressed her preference for a more relaxed seminar-type classroom focusing on teaching skills – rather than language abilities – had she not been expected to prepare the students for the exam.

### *8.2.3 Learners and learning outcomes*

Our study also looked at any effects the test might have had on the learners and the learning outcome. We did this through a quantitative study of the test results administered before and after the training took place, as well as the qualitative study of the students'

reactions to the test and class activities. The data for this part of the study was thus collected at four different stages before and after the training program:

1. Preliminary administration of SPEAK and the Performance test;
2. Students' interviews conducted before the course;
3. Final administration of SPEAK and the Performance test;
4. Students' interviews conducted after the course.

Table 8.2 summarizes the results of a survey conducted among the 35 students who took part in the preliminary administration of both the SPEAK and the Performance tests before the start of the program.

**Table 8.2:**
**Learners' Reaction to the Performance Test**

| | yes | no |
|---|---|---|
| Students regarded the test as more challenging than SPEAK | 86% | 14% |
| Students regarded the test as directly related to their real-life TA tasks | 63% | 37% |
| Students regarded the performance categories as adequate for measuring their spoken language abilities | 71% | 29% |
| Students thought that the test was fair and acceptable | 69% | 31% |
| Students felt uncomfortable being videotaped and speaking in front of the panel | 14% | 86% |
| Students felt uncomfortable with the tape-recorded format of SPEAK | 46% | 54% |
| Students believed that preparation for the test would require them to improve their spoken language abilities | 91% | 9% |

When asked about their reaction to the performance test as opposed to SPEAK, 86% of the students responded that they found it more challenging in terms of the spoken language abilities than SPEAK. In their comments, they also added that because of the interactive nature of the performance test, it provided them with a better chance to demonstrate their knowledge of language. They also added that as opposed to the artificial

situation created by SPEAK, the tasks and topics in the performance test were all meaningfully connected creating a sense of purpose during the test. 37% of the students, however, thought that the tasks in the test were not directly related to their real-life TA responsibilities since, according to them, newly appointed TAs in their departments only did marking. Still, this group of learners were motivated to participate in the course to improve their spoken language abilities. A majority of students (69%) also believed that the format of the test and what it measured was acceptable and fair. Interestingly, most of the learners (86%), including those with a lower proficiency level, did not express any concern about their performance being videotaped while, on the other hand, 46% of them expressed their dislike for the tape-recorded format of the SPEAK. In their comments, they described it as a "dry," "controlled," "confusing" and "unrealistic" form of measuring speaking.

During the second survey conducted few months later, the learners who had participated in the course were questioned with respect to their reactions and contributions to classroom teaching, materials, and activities. During classroom observations, we noticed that, except for group instruction time during which the teacher was the one who talked most, the students contributed most to class activities. This contribution took the form of either individual presentations, or participation in group discussions and feedback. When asked to rank the abilities promoted by the test which they considered as the most effective in their everyday interactions with undergraduates, 90% of the students referred to strategic knowledge. They respectively ranked "functional knowledge" and "textual knowledge" as the second and third most important language abilities. They described their class activities as practice in these areas and claimed that

they tried to apply the compensatory strategies they had learned in class not only to their class presentations but also to their everyday communications in academic contexts. A majority of the students (75%) had also found the textbook exercises (especially in the areas of pronunciation and functional language) very useful. However, all of the students commented that they had consulted the supplementary pronunciation and listening comprehension materials introduced by the teacher. On the other hand, nobody had seen the supplementary video originally suggested along with the textbook. Finally, when asked whether or not they would use the material and do the activities the same way if they were not required to take the performance test as the exit test, 65% said "no," and the rest of the respondents were not sure about it.

To further determine whether any learning had taken place, that is, whether the students who had to take the test did succeed in raising their scores on the test, a Paired Samples T-Test procedure was performed. Paired Samples T-Test computes the differences between the values of the two related variables and tests whether the average differs from 0. Here, it was based on the scores the students in both the experimental and control groups had obtained during the two administrations of the test before (Time 1) and after (Time 2) the training program on the five major 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 8.3 displays the mean values and the standard deviations for the pairs of variables compared for the purpose of the Paired Samples T-Test procedure:

**Table 8.3:**
**Paired Samples Statistics for Experimental and Control Groups**

| | | Experimental (n=13) | | Control (n=9) | |
|---|---|---|---|---|---|
| | | Mean | Std. Deviation | Mean | Std. Deviation |
| 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 |

As expected from the table above, the results of the Paired Samples Test showed that for the experimental group there was a significant difference between the mean values of the Time 1 and Time 2 performances for all ability areas (Table 8.4). 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 8.5).

**Table 8.4:**
**Paired Samples T-Test for the Experimental Group**

| | Paired Differences | | | | | t | df | Sig. 2tail |
|---|---|---|---|---|---|---|---|---|
| | Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference | | | | |
| | | | | Lower | Upper | | | |
| GK1-GK2 | -.5154 | .3349 | 9.288E-02 | -.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 8.5:**
**Paired Samples T-Test for the Control Group**

| | Paired Differences | | | | | t | df | Sig. 2tail |
|---|---|---|---|---|---|---|---|---|
| | Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference | | | | |
| | | | | Lower | Upper | | | |
| GK1-GK2 | -4.44E-02 | 6.009E-02 | 2.003E-02 | -9.06E-02 | 1.747E-03 | -2.219 | 8 | .057 |
| TK1-TK2 | 1.111E-02 | .1537 | 5.122E-02 | -.1070 | .1292 | .217 | 8 | .834 |
| FK1-FK2 | 2.222E-02 | .1922 | 6.407E-02 | -.1255 | .1700 | .347 | 8 | .738 |
| SK1-SK2 | -.1111 | .3333 | .1111 | -.3673 | .1451 | -1.000 | 8 | .347 |
| SC1-SC2 | 4.444E-03 | 9.475E-02 | 3.158E-02 | -6.84E-02 | 7.728E-02 | .141 | 8 | .892 |
| OP1-OP2 | -6.67E-02 | .1414 | 4.714E-02 | -.1754 | 4.204E-02 | -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 Analysis of Variance, ANOVA, and the GLM Multivariate Tests[6] of Between-Subjects Effects both revealing a group effect on the performance of the learners in the second administration of the test (Table 8.6).

[6] GLM Multivariate procedure provides regression and analysis of variance for multiple dependent variables by one or more factor variables. Using this model, one can investigate the effects of individual factors as well as the interactions between factors on the means of various groupings of a joint distribution of dependent variables.

**Table 8.6:**
**Differences Between Groups in Time 2 Administration**

| Variables | t-test for Equality of Means* | | | | 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 |

* No. of cases = 22

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 8.7). The results were consistent with those of the ANOVA Repeated Measures procedure.

**Table 8.7:**
**Tests of Within-Subjects Effects***

| 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 | 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 |

* No. of cases = 22

The significant time effect shows that there has been a difference in learning taking place as a result of time. However, the results in Table 8.4 confirm that this is 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.

### *8.2.4 Discussion of the results*

The results obtained from both the qualitative and the quantitative data suggest that the test has had a washback effect on all three major areas (materials, teaching, and learning) that we have been concerned with. However, we also noticed that the depth, extent and directness of the impact differs with the area being affected.

***Washback on the materials***

The content of teaching seems to be the area directly affected by the test. As reported in 8.2.2, classroom observations revealed that the teacher did not go through the prescribed textbook chapter by chapter, paid less attention to the sections that did not practice the oral skills evaluated by the test, and encouraged those activities which required more practice. They also showed that additional tasks and exercises, especially those practicing strategic competence, were added to some chapters to enhance students' skills in that area. Also introduced by the teacher were supplementary materials for listening comprehension practiced in class and materials for pronunciation widely used by the students for out-of-class practice. The prescribed supplementary video, however, was never used for in-class activities.

These changes to the contents of teaching, if not directly a result of the test, definitely addressed the test in many directions. However, to answer why these changes to materials took place, we turned to our discussions with the teacher regarding this matter. She believed that, given the objectives of the course (which, as stated in 8.2.2, were

introduced with reference to the test objectives in the very first session of the course), some parts of the textbook required more emphasis than other parts. She explained as well that based on the students' performance in the preliminary administration of the test, she believed that extra material focussing on listening comprehension and pronunciation skills needed to be covered inside and outside the classroom. Here again there was a direct reference to the test as the source of input. As for the supplementary video, the teacher emphasized that she wouldn't allocate class time to it since it was not didactic and just contained examples of the teaching strategies used by some experienced TAs. Although there was not a direct reference to the test at this point, considering that the test primarily measured language abilities rather than teaching skills, this choice again implies the test's effect on how the teacher chose the course content. The test's washback on the choice of materials was, therefore, *direct* in that all the changes were in the direction of the test, and *extensive* in that it affected both the main and the supplementary material used inside and outside the classroom.

***Washback on teaching***

As can be seen from the survey results in 8.2.2, teachers, raters and administrators showed a positive attitude towards the test. The teachers and raters especially praised the test's tasks, rating instrument, directness, interactive nature, and ease of administration and scoring relative to the standardized alternatives, TSE and SPEAK. As for the class activities, both observations and student interviews point to the fact that they were, to a large part, adapted to the goals of the test: the textbook was customized according to the aims of the course, supplementary materials were introduced, a substantial amount of time was allocated to the students' presentations and feedback sessions, and the test's rating

instrument was used for class evaluations. This implies that the way of organizing class activities was affected by the test. In addition, given that the teacher was personally in favour of a seminar-type course primarily concerned with the teaching skills, these adaptations imply the test's *direct* and *deep* impact on the teacher's choice of class activities. However, this very reason – i.e., the teacher's preference for a teaching-based syllabus – could also be responsible for the teachers' communicative approach to those class activities which happened to be in line with the objectives of the test. In other words, there is no evidence based on class visits or teacher interviews that the test was necessarily responsible for the teacher's methodology.

***Washback on learners***

As can be seen from the results of the needs assessment (Chapter Five), ITAs were eager to eliminate the language barrier affecting their functioning as a TA for different academic, professional and financial reasons. They actively participated in different phases of the experiment and were exposed to both the standardized SPEAK and the Performance Test. In general, the students' reaction to the test was positive. They described the test as a meaningful device for measuring their language abilities, were comfortable with the undergraduate students and teachers acting as raters and considered the whole testing context as natural. Their awareness of the test, as reflected in both their in-class and out-of-class activities, was high. This was partly due to the teaching effect which itself was in the direction of the test and partly because of the importance the students attached to the test. For instance, student interviews revealed that from among the material the teacher had excluded from the textbook but had assigned for out-of-class practice, only those directly related to the abilities measured by the test had been adopted

by the students. These, for example, included the materials and tapes on listening comprehension and pronunciation strategies but not the section on cultural notes appearing at the end of each chapter in the textbook. Besides, both the teacher interviews and class observations indicated that the students' participation in class activities and exercises that practiced abilities like textual knowledge, the use of functional language, and strategic competence was noticeable. This was in line with the students' rating of these abilities as the most important abilities evaluated by the test. Raters' comments following the Time 2 administration of the test also referred to such factors as better organization, improved fluency, and comprehensibility in the speech of the learners.

Moreover, the influence of the test on the choice of the teaching content, teaching activities, and learning strategies is further confirmed by the results of the statistical analysis of the scores the students obtained before and after taking the course (section 8.2.3). 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 superior performance of the experimental group against that of the control group in the Time 2 administration of the test suggests that this change was the result of the training that took place in the meantime. Mere living and studying in an ESL environment would not necessarily bring about the language abilities needed for efficiently functioning as a TA.

In general, considering the significant change in learners' ability levels measured by the test, the impact of the test on learning was *deep*. However, the impact on the learners was both *direct* through their adoption of learning strategies, and *indirect*, through teaching activities and teaching contents.

## *8.3 EFL context*

### *8.3.1 Advanced group*

The selection of course content and particular materials for the graduate advanced writing course is normally done by the teachers themselves. However, teachers believed that the course description issued by the Ministry of Higher Education which is supposed to guide teachers in determining the textbook and the course outline is very general and uninformative. It summarizes the aims of the course as "writing and analyzing different essays on advanced issues in different styles (theoretical, descriptive, explanatory, and argumentative)," as well as "gaining mastery over different writing types, with an emphasis on theses writing." Given this situation and because we wanted to see how the teacher whose classes were being pilot-studied would react to the performance writing test as opposed to the ministry's course description, no specific material was prescribed as courseware.

The material the teacher had chosen as the main text was Bailey & Powell (1989). The book is a step-by-step approach to writing beginning with essay fundamentals and personal experiences and ideas, and moving towards 1000-to-2000-word term papers. Such topics as organization, unity, and coherence are covered as part of essay fundamentals, while punctuation and advanced grammatical structures most commonly used in writing are dealt with in separate chapters, not as parts of the step-by-step approach. More formal language and strategies for organizing and presenting topical knowledge are specifically addressed in the chapters dealing with research papers. Returning to the focal constructs underlying the choice of the test tasks for this group of learners, we can say that the book does promote the abilities measured by the test.

Nevertheless, it starts with a more fundamental tightly structured one-paragraph essay approach to writing, the ability that the students in the advanced writing program are already expected to possess.

The analysis of the course syllabus and class observation notes, however, revealed that the parts of the book actually covered in class were those dealing with the development of the 5-paragraph essay and the multi-paragraph research essay. According to the teacher, skills such as how to achieve organization, unity, coherence, emphasis, and variety in essays as well as efficient introduction and presentation of support based on the learners' topical knowledge were promoted in class. In addition to the main text introduced above, supplementary materials such as the APA research manual, some sample writings, and articles from different sources were also used.

Class observations further indicated that the teaching methodology and teaching activities in class were in fact in the direction of the test. Class activities included the discussion of essay development techniques, actual writing performances based on previously read materials, and providing feedback to the students by having samples of their writings duplicated and discussed in class. Also noted by the teacher and acknowledged by the observations was the regular evaluation of the students' writings by the teacher with respect to such factors as organization, content, coherence, language and style. The final evaluation of the students only partially counted towards their total grade, a grade which was based on several essay assignments, mid-term and final exams, as well as a term project whose format was based on the APA style sheet. The final exam, just like the papers practiced during the term, addressed the students' topical knowledge by asking

them to elaborate on one of the topics in language teaching. These observations imply the direct effect of the test and its format on instruction and class activities.

Like the ESL group, we studied the learners' reactions to the test by looking at their performance before and after taking the course. We also asked for their opinions about the test and class activities, and compared their performance on the test with that of a control group randomly selected from among the TEFL students who were enrolled in the first-year Master's courses other than advanced writing. The results of a survey conducted among 50 – control and experimental – students originally taking part in the preliminary administration of the test showed that 82% of the students considered the test task as exactly reflecting the type of the written activity they routinely engage in at the Master's level. A total of 74% believed that a course aimed at promoting the abilities measured by the test was necessary, and 60% of them said that they tried to take the course as early as possible in their graduate studies.

The students who eventually took the course, were aware of the aims of the course and the content of the test. They especially appreciated the fact that the topics for class activities were all related to their area of expertise, thus making both in-class practice and out-of-class assignments more interesting to them. In addition, both the observers and the students noted the feedback the students got on their performance during the course. They had found the use of the students' writing samples as class material was a very effective method. It not only brought the problems common to students to their attention, but also helped the individual writers to get teacher and peer feedback about their writing. Although observations indicated that the highlighted areas were in line with the ability areas measured by the end of the term performance test, the students mentioned that the

most emphasis was on such areas as rhetorical organization and overall unity and coherence of the writing, and areas such as vocabulary and grammar were not directly addressed or discussed in class. They, nevertheless, did not rank grammar and vocabulary as areas where they thought they had a major problem.

When asked if they thought their writing skill had improved, 82% of the students answered positively. They considered factors such as the syllabus, class practice and frequent writing assignments responsible for their improvement. In answer to the question whether they thought the test task had affected their learning efforts, some of them commented:

> what we did in the exam was nothing different from what we were doing all along
> sure, I think it affected the students' and teacher's performance during the course
> writing is difficult, honestly I don't think I would go ahead with all that work if I were tested otherwise
> you bet, who wants to fail?

Furthermore, the statistical analysis of the scores the students had obtained during the two administrations of the test before and after the course, revealed that the learners had in fact improved their scores on the six major ability areas (Grammatical Knowledge (GK), Textual Knowledge (TK), Functional Knowledge (FK), Sociolinguistic Knowledge (SK), Strategic Competence (SC), and the use of Topical Knowledge (Tp.K))as a result of the course (Table 8.8). The results of the Paired Samples Test showed that, for the experimental group, there was a significant difference between the mean values of the Time 1 and Time 2 performances for all ability areas. The control group, however, did not

show any significant progress in any one of the ability areas during that time frame (Table 8.9). They even performed significantly poorer with respect to the two areas of SK and SC.

**Table 8.8:**
**Paired Samples Statistics for Experimental and Control Groups**

| | Experimental (n=22) | | Control (n=20) | |
|---|---|---|---|---|
| | Mean | Std. Deviation | Mean | Std. Deviation |
| Pair 1 GK1 | 2.1432 | .3630 | 1.9175 | .5209 |
| GK2 | 2.7591 | .3386 | 1.9175 | .5154 |
| Pair 2 TK1 | 1.2068 | .4599 | 1.0700 | .3526 |
| TK2 | 2.7591 | .4455 | 1.0178 | .3931 |
| Pair 3 FK1 | 1.2273 | .3239 | 1.1100 | .3837 |
| FK2 | 2.4818 | .4305 | 1.0000 | .3539 |
| Pair 4 SK1 | 2.1364 | .3155 | 2.1875 | .4507 |
| SK2 | 2.9091 | .1974 | 2.0500 | .3940 |
| Pair 5 SC1 | 1.0682 | .4363 | 1.0650 | .3884 |
| SC2 | 2.6659 | .4584 | .8975 | .5051 |
| Pair 6 Tp.K1 | .5545 | .5722 | .5850 | .4945 |
| Tp.K2 | 2.9955 | .4134 | .5250 | .4993 |
| Pair. 7...OP1 | 1.7773 | .3380 | 1.7700 | .4846 |
| OP2 | 2.8659 | .3698 | 1.6550 | .4751 |

**Table 8.9:**
**Paired Samples Test for the Experimental and the Control Group**

| | Experimental | | | Control | | |
|---|---|---|---|---|---|---|
| | t | df | sig. | T | df | sig. |
| GK1-GK2 | -8.247 | 21 | .000 | .000 | 19 | 1.000 |
| TK1-TK2 | -13.512 | 21 | .000 | .903 | 19 | .378 |
| FK1-FK2 | -12.925 | 21 | .000 | 1.655 | 19 | .114 |
| SK1-SK2 | -10.803 | 21 | .000 | 2.463 | 19 | .024 |
| SC1-SC2 | -14.614 | 21 | .000 | 4.030 | 19 | .001 |
| Tp.K1-Tp.K2 | -18.770 | 21 | .000 | .536 | 19 | .598 |
| OP1-OP2 | -14.361 | 21 | .000 | 2.562 | 19 | .019 |

Table 8.10 displays the results of the One-Way ANOVA test, Independent Samples Test, and GLM Tests of Between-Subjects Effects implying a significant group

effect on the performance of the learners in the second administration of the test. In other words, 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.

**Table 8.10:**
**Differences Between Groups in Time 2 Administration**

| Variables | t-test for Equality of Means* | | | | Test of Between-Subjects Effects* Source: Group | |
|---|---|---|---|---|---|---|
| | t | df | Sig. (2-tailed) | Mean Difference | F | Sig. |
| GK2 | 6.310 | 40 | .000 | .8416 | 17.710 | .000 |
| TK2 | 13.374 | 40 | .000 | 1.7413 | 72.299 | .000 |
| FK2 | 12.226 | 40 | .000 | 1.4818 | 64.800 | .000 |
| SK2 | 9.060 | 40 | .000 | .8591 | 17.216 | .000 |
| SC2 | 11.897 | 40 | .000 | 1.7684 | 50.456 | .000 |
| Tp.K2 | 17.525 | 40 | .000 | 2.4705 | 92.016 | .000 |
| OP2 | 9.263 | 40 | .000 | 1.2109 | 25.254 | .000 |

* No. of cases = 42

Multivariate Tests of Within-Subjects Effects and the ANOVA Repeated Measures procedure also showed a significant ($p<.01$) time effect, as well as a combined time and group effect on the performance of the learners in all category areas.

### *8.3.2 Intermediate group*

As in the EFL graduate level writing course discussed above, to observe the extent to which the end-of-term test affected the content of teaching, the choice of material for the undergraduate writing course was left to the teacher once she accepted to give the performance test to her classes before and after instruction. Interestingly, we found out that the main textbook chosen for this course was the same as that of the graduate

advanced writing course, i.e., Bailey & Powell (1989). This was despite the fact that the teachers were teaching writing courses with different objectives, at different levels and at different institutions. So, at this point, it was important to find out why the teachers had chosen the material they were using, what areas they had chosen to emphasize, and what teaching approach they had adopted towards the content of the course.

The complementary data obtained from observations and teacher interviews revealed that in the undergraduate class, the parts of the text mostly focused on were those dealing with the simple one-paragraph essay and the organization of slightly more sophisticated longer essays. According to the teacher, the second section of the book dealing with the development of a more scholarly type of writing had not been included in the course content. The selected parts of the book resembled the exam in that they promoted the fundamentals of essay writing while the support for the main idea came from the writer's own experiences, imagination, or general knowledge. When asked why she was using Bailey & Powell (1989) as the main text for her class, she replied that the book provided plenty of useful patterns and writing models not to mention that it was also commercially available with a reasonable price. The teacher had also introduced to the class supplementary material (Langen, 1988) parts of which had been used for both in-class and out-of-class activities.

To investigate the relationship between the test and the teacher's methodology, if any, we looked at the course description prescribed by the ministry, the teacher's syllabus and actual class activities. The course description issued by the Ministry of Higher Education, requires teachers to provide "a comprehensive definition of essay types" and "present samples of great English writers." The teacher's own syllabus, however, includes

"reading, examining and evaluating of model essays" as well as "extensive writing, and teacher and peer feedback" as the main method used in class. It lists such areas as content, organization, development, transition, vocabulary, language structure and mechanics, as the main topics to be dealt with in the course. This discrepancy between the ministry's course description and the teacher's syllabus was later justified by the teacher's comment that "... the [ministry's] course objectives are designed with no reference to the students' objectives."

According to the class notes, class activities at the beginning of the term mostly consisted of brainstorming and free writing of ideas. Later course activities consisted of both reading and writing activities. Previously read passages were discussed with reference to the writer's method of development, main and supporting ideas, linking words, transitional sentences, idiomatic expressions, figurative language and choice of words. Writing activities included providing support for a given topic sentence or developing essays on a given topic. Both the teacher and the learners considered class assessments as an effective part of the methodology. When asked what the relationship between her methodology and the final test was, the teacher commented: "...in the test, they write on a given topic, the students expect such a topic since in class they learn about parts of the [test] task ,... the quizzes and the mid-term do the same [thing that the test does] in more controlled ways... So, the relationship between the teaching method and the test type and content is obvious."

The learners in general considered the exam as an effective means of learning simply because it made them do more and more writing and made their teacher correct their work and point out their problems. They all referred to their everyday academic need

to do some sort of writing as their main motivation for wanting to improve their writing skill. Most of them mentioned that their main problem was that they didn't know how to express their intended meaning in writing in a clear way and thought that class activities helped them with this. As for the learners' attitude towards the test, the teacher's personal belief was that students rarely show a positive attitude towards tests but in this case they had found the test congruent with their expectations. The students themselves commented that they had found the exam fair and realistic and thought that they had improved.

The scores they had obtained during the two administrations of the test before and after the course, revealed that the learners had in fact improved their performance in the four major ability areas – Grammatical Knowledge (GK), Textual Knowledge (TK), Functional Knowledge (FK), Sociolinguistic Knowledge (SK) – as well as their overall performance (OP) as a result of the training they had received.

**Table 8.11:**
**Paired Samples Statistics for Experimental and Control Groups**

| | Experimental (n=30) | | Control (n=28) | |
|---|---|---|---|---|
| | **Mean** | **Std. Deviation** | **Mean** | **Std. Deviation** |
| Pair 1 GK1 | 1.8167 | .4997 | 1.5714 | .4707 |
| GK2 | 2.6417 | .4289 | 1.4429 | .4180 |
| Pair 2 TK1 | 1.4333 | .5371 | .7411 | .4929 |
| TK2 | 2.7833 | .5323 | .6125 | .5215 |
| Pair 3 FK1 | 1.4833 | .6497 | .8571 | .5066 |
| FK2 | 2.6833 | .4450 | .8036 | .5153 |
| Pair 4 SK1 | 1.7500 | .4453 | 1.1036 | .3960 |
| SK2 | 2.6167 | .4583 | 1.0179 | .3531 |
| Pair. 5...OP1 | 1.7000 | .5099 | 1.4036 | .4282 |
| OP2 | 2.9400 | .4223 | 1.3429 | .3360 |

The results of the Paired Samples Test showed that for the experimental group there was a significant difference between the mean values of the Time 1 and Time 2

performances for all ability areas. The control group, however, did not show any significant progress in any one of the ability areas during that time frame (Table 8.12).

**Table 8.12:**
**Paired Samples Test for the Experimental and the Control Group**

| | Experimental | | | Control | | |
|---|---|---|---|---|---|---|
| | t | df | sig. | T | df | sig. |
| GK1-GK2 | -10.131 | 29 | .000 | 1.932 | 27 | .064 |
| TK1-TK2 | -14.840 | 29 | .000 | 2.260 | 27 | .032 |
| FK1-FK2 | -11.029 | 29 | .000 | .902 | 27 | .134 |
| SK1-SK2 | -9.957 | 29 | .000 | 1.544 | 27 | .131 |
| OP1-OP2 | -14.967 | 29 | .000 | 1.559 | 27 | .375 |

Besides, as it is clear from Table 8.13, there was a significant group effect on the performance of the learners in the second administration of the test, i.e., 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.

**Table 8.13:**
**Differences Between Groups in Time 2 Administration**

| Variables | t-test for Equality of Means* | | | | Test of Between-Subjects Effects* Source: Group | |
|---|---|---|---|---|---|---|
| | t | df | Sig. (2-tailed) | Mean Difference | F | Sig. |
| GK2 | 10.767 | 56 | .000 | 1.1988 | 45.172 | .000 |
| TK2 | 15.673 | 56 | .000 | 2.1708 | 129.530 | .000 |
| FK2 | 14.898 | 56 | .000 | 1.8798 | 99.488 | .000 |
| SK2 | 14.805 | 56 | .000 | 1.5988 | 136.448 | .000 |
| OP2 | 15.864 | 56 | .000 | 1.5971 | 84.779 | .000 |

* No. of cases = 58

This difference between the two groups was also referred to by the raters who noticed a considerable difference between the length of the essays written by the two groups in the

Time 2 administration of the test. The experimental group's essays had been judged as longer and more fully developed compared with those of the control group.

Just like the advanced group, here too, Multivariate Tests of Within-Subjects Effects and the ANOVA Repeated Measures procedure showed a significant ($p<.01$) time effect, as well as a combined time and group effect on the performance of the learners in all category areas.

### *8.3.3 Discussion of the results*

The above results from the study conducted in the EFL context suggest that both intermediate and advanced tests of writing ability had affected the following areas of teaching and learning to some degree:

***Washback on the materials***

A very interesting and at the same time revealing observation in the EFL context was that both of the teachers whose classes had been pilot-studied chose the same book as the main text for their classes despite the fact that they did not work together and were preparing different groups of learners at different levels for different types of tests. So, given the fact that the choice of the material was completely up to the teachers, what could have possibly affected their decision in this regard? The results in the previous sections point at three factors that could potentially have influenced the teachers' choice of materials: non-academic factors outside the classroom, the course description prescribed by the ministry, and the test format and content.

As one of the teachers had indicated, her choice of material had partly been influenced by such non-academic factors as the commercial availability of the book and its

price. Yet we cannot ascribe the teachers' choice of course content to the limitations in the educational context in which they were functioning because, as discussed in 8.3.1 and 8.3.2, the content of the textbook in both teachers' classes had been narrowed in ways indicative of the level and purpose of the course. There were also supplementary materials used for each course that served to further practice and complement the selected parts of the textbook and promote the objectives of the course. These objectives could have been determined by the course description defined by the ministry or the educational needs of the learners. Based on the contents of the teachers' syllabi, their direct comments, and the type of materials they chose, the former case is clearly not an option. On the other hand, as we know from the discussion in the previous sections, there were no conflicts in the aims and activities of the course and those of the tests, i.e., the abilities the courses were aimed at were also those measured by the tests. This implies that the materials used by the teachers – whether influenced by the test or directly by the needs of the learners and/or the educational system – reflected the contents of the tests.

***Washback on teaching***

It was very clear from the examination of the course syllabi, and the teachers' and raters' survey results that the teachers in both graduate and undergraduate levels had a very positive attitude towards the tests' tasks and formats. However, here it is difficult to tell from the data whether the teachers' methodology was affected by the test or whether they simply did what they thought they were supposed to do considering the objectives of the program and their students.

At both levels, the teachers assessed the students regularly throughout the course, and as expected the final exam constituted only a portion of the students' final grade. But

these other means of assessment all reflected the tasks and contents of the exam. For example, we noticed that all writing assignments and class discussions for the advanced graduate group had been done with reference to the students' topical knowledge, while on the other hand, the undergraduate group's teacher said that she tried to assign topics of general interest that encouraged a response. Class tests had also been marked using the same criteria used for the exam. When asked how they felt about the rating scale, the raters replied that it was realistic and sensible. One of the teachers also added that after a while she could work with the scale with ease so she had adopted the same scale for marking the students' work throughout the term. This implies the exam's direct impact on how teachers evaluated students and marked their tests.

As for the teachers' methodology, their choice of materials, class activities, assignments, and topics for class work and exam performance reflected their awareness of the students' levels and needs as well as their understanding of the test requirements. The data in this respect point both to the exam and the academic needs of the students as powerful determinants of the teachers' methodology. Course syllabi, class teaching and testing activities, as well as teachers' own comments reflect an obvious relationship between the exam and the teaching methods. Nevertheless, when asked if they would change the teaching contents and their methodology had the students been required to write a different type of test (such as short-answer), the answer was that they couldn't think of a multiple-choice or short-answer type of test for a writing course. One of the teachers further elaborated on her reasons for adopting a methodology in the direction of the exam by stating:

> the very existence of writing courses in the curriculum is suggestive of the fact that the *need* is there. One should see the real more tangible *needs* as required by the students, ... that the students *need* to express themselves in writing later on in more advanced courses... (emphasis added)

So, as it should be clear by now, the issue we face in this context is disentangling the influence of the test per se from that of the students' and educational context's needs. In other words, it is difficult to say whether the teachers' adoption of the general approach and the methodology promoted by the exams was because they thought it was the best way to prepare their students for the abilities measured by the exams or because they were merely doing what they would ordinarily have done to meet the academic needs of the students. This, however, is not a problem from the standpoint of this study given the fact that the new tests were theoretically meant to reinforce the needs of the educational contexts for which they had been designed. So, either way the results were positive, since the needs of the learners which for the purpose of the study and the test development were taken to be fundamental had been appropriately addressed by the general approach and the methodology adopted by the teachers.

***Washback on learners***

In the process of our data collection, we heard claims that the writing courses had helped the students to improve their writing ability. Also, according to the teacher for the undergraduate group who appeared to regularly monitor students' progress through classroom tests, the students' progress had been helped by the exam since, according to her, they had found it "congruent with their expectations." Class observations, too, reported on students' active participation in such class activities as peer correction and

content revision resulting in the improvement of the ability areas directly measured by the test. For instance, as evidenced by the statistical results as well as the raters' observations, there was a significant difference between the Time 1 and Time 2 performances of the graduate group on the use of topical knowledge, while for the undergraduate group the length of the essays had changed considerably.

Students' evaluations of the exams and their significance also reflected a similar picture. While they all expressed their strong desire to pass the exam, they never expressed any anxiety or fear resulting from the exams. This could have been the result of the teachers' approach to the course evaluation which based the students' final grades on multiple assessments given during the course rather than a single administration of the test. These frequent class evaluations – which as we know from our discussion above had all been affected in form, content and scoring procedure by the exam – indirectly helped the students become aware of the abilities assessed by the exam and thus try to master them through class activities and practice.

The students' improvement in the areas measured by the test, as evidenced by their improved scores in the Time 2 administrations of the exams, further supports the effectiveness of both the teaching methodology and the learning strategies which – as we know from our discussion above – had been influenced, in one way or another, by the exams.

# *PART FOUR*

# *CONCLUDING REMARKS*

# CHAPTER NINE
# CONCLUSIONS AND IMPLICATIONS

## *9.1 Assumptions and expectations*

In this section, we return to the theoretical framework proposed in Chapter Three for examining the washback effect. Our findings based on observations and other forms of data collection will then be listed in terms of this theoretical framework.

The model presupposes that washback, as a form of the consequential aspect of validity, has to be addressed in the design of a test, and that tests with minimal construct under-representation and construct irrelevant variance are more likely to produce beneficial washback (Messick, 1996). The idea is incorporated in the model by basing the development and scoring of a test (Phase II) on the needs and objectives of the language learners and the educational context in which they are expected to function (Phase I).

Addressing the learners' needs is of significance to a study of test consequences in that it can shed light on not only the fundamental test-related factors likely to create positive washback, but also those non-test variables that can promote its occurrence. An awareness of the educational needs of the learners also helps test makers to delineate the tasks, contexts, and constructs to be included in the test. This will in effect result in the choice of the test tasks and contexts which are closely related to the real world tasks, processes and contexts (i.e., authenticity), and constructs which represent all of the important abilities and component abilities to be tested (i.e., minimal construct under-representation). A consideration of the learners' needs further helps test developers to adopt tasks with response formats that most directly elicit the behaviour through which

inferences can be made about the abilities being assessed, thus avoiding assessments which are too broad and increase the effect of construct-irrelevant variance on the examinee's behaviour (i.e., directness).

The two validity standards discussed above (namely, minimal construct under-representation and construct-irrelevant difficulty) may not, however, readily lead to positive washback if some variables present in the educational system are not strengthened and exploited in the process of test design. Relying on information related to the needs of the learners and to the educational system also helps us to gain insight into factors other than those related to the test validity that can enhance the occurrence of positive washback. We can then find ways to incorporate such factors into the process of test design. Although the number and type of these factors might differ with different contexts and different learners, some ways to accomplish this are by using controlling variables such as learners' background knowledge, by determining learners' affective factors, and by involving teachers in the choice of the test tasks and topics. The model thus assumes that while the test is the central component in a theory of washback, the occurrence of washback depends on factors including but not limited to the test's validity. It also suggests that a test designed on the basis of the needs of the target population and context is likely to affect the areas of material development, teaching, and learning and that these are the areas where evidence for washback should be sought (Phase III).

## *9.2 Conclusions*

This study tested the proposed theoretical framework for washback over time, across groups and settings, and in response to experimental interventions such as contextual factors, instructional treatment, proficiency levels, and motivational conditions[1]. The findings are thus categorized here in terms of the components of the model, as well as those factors found by the study to be instrumental in promoting or hindering positive washback.

*1. Materials choice/development*

This is the area that seems to be most affected by the test. The results of the study show a direct and intense relationship between the contents of the tests and those of teaching. Given that the tests were theoretically in no conflict with the aims and objectives of the program, teachers' efforts to modify the contents of teaching with respect to the contents of the tests and to complement the textbooks with additional materials and exercises can be considered as positive steps towards achieving the academic objectives of the learners. The information gathered in this regard generally pointed to the tests as being responsible for this effect:

- The teacher in the ESL context did not cover everything in the prescribed textbook but modified it with respect to the contents of the test.

---

[1] As extensively discussed in Chapters Five through Eight, this research is based on both qualitative and quantitative data collected at the institutional level from all participants at different times. The final statistical data is, however, based on three single student populations within each program. We are, therefore, not generalizing the results to contexts beyond the three universities where we conducted the research, although we have no reason to believe that, for example, the needs, objectives, and working environment of ITAs in other Canadian universities are different from those of ITAs in the University of Victoria. Besides, the apparent homogeneity of the student population subjected to the training course, and the comprehensive nature of the qualitative data, support the representativeness of the subject population.

- The ESL teacher skipped the material practicing teaching strategies (this could be a negative test consequence if students' inability to teach interfered with the expression of their language abilities).
- Teachers in both contexts paid less attention to areas of grammatical knowledge than to those of strategic and textual knowledge.
- EFL teachers adopted different selection approaches to the same material.

*2. Teaching activities and methodology*

Observations revealed that teaching tasks and activities in class had to a large part been positively affected by the tasks and activities required by the tests. This was reflected in the following classroom activities:

- Class evaluations reflected the form and content of the test.
- Teachers used the same rating scale used by the test.
- Class activities and practice were in general concentrated on the areas to which more weight had been given by the exam (negative when certain ability areas were over-practiced).
- Students' work was used as class materials.

Our results further showed that compared with the contents and activities of teaching, teaching methodology is an area where a direct effect of the test was less evident. Although this does not necessarily imply the absence of test impact on this area, it shows that the washback effect on methodology is not as easily traceable to the test. For example, in the ESL context, class observations backed the idea of the test's impact on the teacher's methodology while, on the other hand, the teacher interview revealed that had she not been supposed to prepare the students for the exam, she would prefer to present a

seminar-type course primarily addressing ITAs' teaching strategies than focussing on the language problems per se. Given this, despite the fact that her approach to *what* should have been taught in the course was influenced by the exam, we don't know if *how* she taught the course, i.e., her adoption of a communicative approach, was influenced by the exam or whether she would have taught any ITA course the same way because it was her preferred method of teaching that had just happened to be in line with the objectives of the test.

In the EFL context, too, we were faced with an uncertainty of a different nature with regard to the influence of the exam on the teachers' methodology. Here the teachers referred to the *needs* of the learners as the main reason for why they were teaching the way they were teaching. Since there was no conflict between the goals of the tests and the needs of the learners – from which they had theoretically been derived – the results were positive and supportive of our proposed model for a theory of washback.

*3.Learning strategies and outcome*

There was evidence that the effect of the exams over learning activities was both direct and indirect (through teaching practice and materials). Class observations, teacher interviews, and student questionnaires showed that this effect was mostly positive:

- The learners made extensive use, for extra learning activities, of the materials and exercises that the teachers had introduced based on the objective of the course which itself had been derived from the exam.
- The learners provided feedback in class using the components of the rating instrument used for the exam.

- The learners were motivated to pass the test.
- EFL learners didn't express any sense of anxiety towards the test (negative if originating from an ignorance of the exam and its contents).
- All learners performed significantly better at the end of the term especially with respect to those areas emphasized by the test.

Our study also shed light on factors that potentially can facilitate or hinder the occurrence of the washback effects of a test. Examples of such factors are the status of the test, its format, and the educational system in which the test is administered.

*Low/high stakes*

We noticed similar patterns in the washback effect of both ESL and EFL tests on the immediate educational environment – classroom teaching and learning activities – despite the fact that the ESL test was a high-stakes placement test and the EFL tests were classroom achievement tests. This can be ascribed to the design of the tests being informed by the needs of the educational systems and specific variables operating in those contexts. However, we also noticed that with the high-stakes test, so many factors originating from sources other than those in the classroom environment were critical not only for delineating needs before the design of the test but also for helping positive washback continue to happen once the test was in effect. While all stake-holders in the educational system acknowledged the significance of the ITAs' language training for the improvement of undergraduate programs, the performance test has never been used since then on a campus-wide basis to screen international graduate students for the TA job. There are financial and ethical issues that have to be resolved before that happens. For example, there are questions as to whether or not passing the test should be mandatory for

assignment to TA duties, or who should financially sponsor ITA language programs. In the meantime, the teachers in the EFL context continue to administer the performance tests to their writing classes. We conclude, therefore, that while language tests (especially those with high stakes) designed within the theoretical framework introduced here are very likely to bring about positive washback on teaching and learning activities, a change in the educational system definitely requires input from various resources, including but not limited to the test. In other words, various non-test related factors in the educational system might facilitate positive washback or prevent it from happening.

*Testing theory*

The tests used for the purpose of this study were communicative tests measuring both complex skills and their component skills delineated by the learners' needs. A direct test based on a comprehensive theory of testing which targets learners' communicative competence can better incorporate different aspects of learners' needs and thus promote positive washback.

*Resources in the educational system*

Our study also pointed at the resources in the educational context such as the availability or absence of materials, teachers' resourcefulness, teachers' beliefs, and teachers' and learners' awareness and understanding of the test as important factors affecting and interfering with the use of a new test and its washback effect.

## *9.3 Pedagogical implications*

An awareness of the positive consequences of testing, how they work, and how we can promote them has serious implications for educational testing in general and language

testing in particular. Traditionally speaking, testing, especially achievement testing, is considered as part of the curriculum which occurs at a later stage and follows teaching. In that case, it is possible for teaching to take place for some time without teacher(s) having the slightest idea about the contents, tasks and format of the test to be administered at the end of the term. It is also likely that teachers do not cover certain parts of the syllabus in class (for reasons such as inadequate planning, time limits, difficulty level or learners' disinterest in the subject) and consequently do not include those parts in the test.

The model of washback examined here, while serving as a workable model for promoting positive washback, implies that the instructional process does not necessarily have to be unidirectional, going from syllabus design to testing. Tests, be they high-stakes external examinations or classroom achievement tests, can serve the needs of learners and the educational system very well as a source of input for material development, motivation for learners, and guidance for teachers. In other words, in most educational systems, tests can be used as instruments that keep teaching and learning activities focussed.

Furthermore, the study points to the fact that we need more authentic performance-based types of language tests that measure both complex abilities and their component abilities. Such tests provide the score users – teachers, learners, and institutions – with useful information about what the score means in terms of the ability areas (or constructs) that underlie test performance. This means that for the tests to be positively involved in the curriculum and help teaching and remediation, they have to be designed with reference to the cognitively-based theories of language test performance that picture the language as a complex, multi-dimensional and dynamic interaction between a variety of knowledge sources.

## *9.4 Suggestions for research*

Clearly, further research is needed into "a phenomenon on whose importance all seem to be agreed, but whose nature and presence have been little studied" (Alderson & Wall, 1993, p. 115). Given the complexity of the phenomenon, long-term research programs are needed to study and further examine the nature, scope and limits of washback.

The published literature on washback has been, to a large extent, limited to the study of the effects of language tests administered in one foreign language environment. It is desirable to conduct extensive longitudinal studies on the washback effect of high-stakes language tests, such as TOEFL, MELAB, IELTS, TSE, etc., across settings (EFL/ESL), cultures and language backgrounds and to evaluate the positive/negative consequences that such tests and the importance given to the scores obtained on those tests have had on the practice of language teaching and learning in those environments. Do the preparation courses and extensively published commercial materials that claim to increase the applicants' scores on such tests during a short period of time really teach English? Do the learners have to increase their proficiency in language in order to be able to increase their scores on these tests? Is that really what preparation courses do? If the answers to these questions are "no," then the question is what the consequences are. Do undesirable consequences originate from the invalidity of such test scores for the purposes they are used or from other educational factors operative in the system?

Finally, and in addition to teachers' and learners' behaviour, teaching methodology, and the contents of teaching, it is important that the research on washback looks into un/desirable learning outcomes in terms of their relationship with the test. This requires an examination of what Frederiksen & Collins (1989) call "systemic validity" of

the test, the evidence for which comes from the learners' progress in the ability areas measured by the test "after the test has been in place within the educational system for a period of time" (Frederiksen & Collins, 1989, p. 27).

# BIBLIOGRAPHY

Alderson, J. C. (1986). Innovations in language testing. In M. Portal (Ed.), *Innovations in language testing* (pp. 93-105). London: NFER/Nelson.

Alderson, J. C. (1991). Language testing in 1990s: How far have we come? How much further have we to go? In S. Anivan (Ed.), *Current developments in language testing* (pp. 1-26). Singapore: SEAMEO Regional Language Centre.

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. (1990). *The Sri Lankan O-level evaluation project: Second interim report.* Lancaster University.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 41-69.

American Educational Research Association, & National Council on Measurements Used in Education. (1955). *Technical recommendations for achievement tests.* Washington, DC: National Education Association.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*(2), Part 2.

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals.* Washington, DC: APA.

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests.* Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.

Anastasi, A. (1961). *Psychological testing* (2nd ed.). New York: Mac-Millan.

Anastasi, A. (1968). *Psychological testing* (3rd ed.). New York: Mac-Millan.

Anastasi, A. (1976). *Psychological testing* (4th ed.). New York: Mac-Millan.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1-15.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 19-32). Hillside, NJ: Erlbaum.

Arter, J. A., & Spandel, V. (1992). Using portfolios of student work in instruction and assessment. *Educational Measurement: Issues and Practice, 11*(1), 36-44.

Bachman, L. (1988). Problems in examining the validity of the ACTFL oral proficiency interview). *Studies in Second Language Acquisition, 10*(2): 149-164.

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

Bachman, L. (1991). What does language testing have to offer? *TESOL Quarterly, 25*(4), 671-704.

Bachman, L., & Palmer. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bailey, K. (1984). A typology of teaching assistants. In K. M. Bailey, F. Pialorsi, & J. Zukowski/Faust (Eds.), *Foreign teaching assistants in U.S. universities* (pp. 110-125). Washington, DC: NAFSA.

Bailey, K. (1985). If I had known then what I know now: Performance testing of foreign teaching assistants. In P. Hauptman, R. Leblanc, & M. Wesche (Eds.), *Second*

*language performance testing* (pp. 153-180). Ottawa: University of Ottawa Press.

Bailey, K. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*(3), 257-79.

Bauer, G., & Tanner, M. (Eds.). (1994). *Current approaches to instructional TA preparation in higher education: A collection of program descriptions*. Seattle: Center for Instructional Development and Research.

Berwick, R. (1989). Needs assessment in language programming: From theory to practice. In R. K. Johnson (Ed.), *The Second language curriculum* (pp. 48-62). Cambridge: Cambridge University Press.

Bornstein, R. F. (1996). Face validity in psychological assessment: Implications for a unified model of validity. *American Psychologist, 51*(9), 983-84.

Brennan, Robert L. (1998). Misconceptions at the intersection of theory and practice. *Educational Measurement: Issues and Practice, 17*(1), 5-9, 30.

Brindley, G. (1989). The role of needs analysis in adult ESL program design. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 63-78). Cambridge: Cambridge University Press.

Brown, G., & Yule, G. (1983). *Teaching the spoken language*. Cambridge: Cambridge University Press.

Brown, H. D. (1993). *Principles of language learning and teaching*. Englewood Cliffs, NJ: Prentice-Hall.

Buck, G. (1988). Testing listening comprehension in Japanese university entrance examinations. *JALT Journal, 10*, 15-42.

Byrd, P., & Constantinides, J. (1992). The language of teaching mathematics: Implications for training ITAs. *TESOL Quarterly, 26*(1), 163-167.

Canale, M. (1983). On some dimension of language proficiency. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 333-342). Rowley, MA: Newbury House.

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

Cattell, R. B. (1964). Validity and reliability: A proposed more basic set of concepts. *Journal of Educational Psychology*, *55*, 1-22.

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

Clark, J. L. D., & Swinton, S.S. (1980). *The test of spoken English as a measure of communicative ability in English-medium instructional settings* (TOEFL Research Report 7). Princeton, NJ: Educational Testing Service.

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). New York: Harper.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

Cronbach, L. J. (1971). Validity. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 443-597). Washington, DC: American Council on Education.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? New directions for testing and measurement: Measuring achievement over a decade. In *Proceedings of the 1979 ETS Invitational Conference* (pp. 99-108). San Francisco: Jossey-Bass.

Cronbach, L. J. (1984). *Essentials of psychological testing*. New York: Harper & Row.

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillside, NJ: Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement theory and public policy: Proceedings of a symposium in honor of Lloyd G. Humphreys* (pp. 147-171). Urbana: University of Illinois Press.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621-694). Washington, DC: American Council on Education.

Davies, A. (1968). *Language testing symposium: A psycholinguistic approach.* Oxford: Oxford University Press.

Davies, A. (1985). Follow my leader: Is that what language tests do? In Y. P. Lee, A. C. C. Y. Fok, R. Lord, & G. Low (Eds.), *New directions in language testing.* Oxford: Pergamon.

Douglas, D., & Myers, C. (1990). *Teaching assistant communication strategies* (Videotape and instructors manual). Ames: Iowa State University Media Production Unit.

Educational Testing Service. (1990). *Test of Spoken English: Manual for score users.* Princeton, NJ: ETS.

Eisemon, T. O. (1990). Examinations policies to strengthen primary schooling in African countries. *International Journal of Educational Development, 10,* 69-82.

Færch, C., & Kasper, G. (1983). Plans and strategies in foreign language communication. In C. Færch & G. Kasper (Eds.), *Strategies in interlanguage communication* (pp. 20-60). London: Longman.

Færch, C. & Kasper, G. (1984). Two way of defining communication strategies. *Language Learning, 34*(1), 45-63.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Fullan, M. G., & Stiegelbauer, S. (1991). *The new meaning of educational change.* London: Cassell.

Gokcora, D. (1992). *The SPEAK test: International teaching assistants' and instructors' affective reactions.* Paper presented at the 26th Annual TESOL Convention, San Francisco.

Goodlad, J. I., Klein, M. & Associates. (1970). *Behind the classroom door.* Worthington, OH: Charles A. Jones.

Green, Donald R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement: Issues and Practice*, *17*(2), 16-19, 34.

Guilford, J. P. (1936). *Psychometric methods*. New York: McGraw-Hill.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, *6*, 427-439.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman

Hambleton, R. K. (1984). Validating the test scores. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 199-230). Baltimore, MD: The John Hopkins University Press.

Hatch, E. (1978). Discourse analysis and second language acquisition. In E. Hatch (Ed.), *Second language acquisition: A book of readings* (pp. 401-35). Rowley, MA: Newbury House.

Heaton, G. B. (1990). *Writing English language tests*. London: Longman.

Heyneman, S. P., & Ransom, A. W. (1990). Using examinations and testing to improve educational quality. *Educational Policy*, 177-92.

Hoekje, B., & Linnell, K. (1994). Authenticity in language testing: Evaluating spoken language tests for international teaching assistants. *TESOL Quarterly*, *28*(1), 103-126.

Hoekje, B., & Williams, J. (1992). Communicative competence and dilemma of international teaching assistant education. *TESOL Quarterly*, *26*(2), 243-269.

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* (ELT Document 127, pp. 134-153). Modern English Publications.

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

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

Johnson, K. (1991). Modifying the SPEAK test for international teaching assistants. *TESOL Matters*, 8.

Kane, M. T. (1992). An argument based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kellaghan, T., & Greaney, V. (1992). *Using examinations to improve education: A study of fourteen African countries*. Washington, DC: The World Bank.

Kennedy, C. (1988). Evaluation of the management of change in ELT projects. *Applied Linguistics, 9*(4), 329-42.

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

Khaniya, T. R. (1990b). *Examinations as instruments for educational change: Investigating the washback effect of the Nepalese English exams*. Ph.D. Dissertation. University of Edinburgh.

Lane, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17*(2), 24-28.

Langan, John. (1988). *College writing skills*. USA: McGraw-Hill, Inc.

Lees-Haley, P. R. (1996). Alice in validityland, or the dangerous consequences of consequential validity. *American Psychologist, 51*(9), 981-983.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*(2), 14-16.

Linn, R. L. (1979). Issues of validity in measurement for competency-based programs. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based measurement* (pp.108-23). Washington, DC: National Council on Measurement in Education.

Linn, R. L. (1983). Curricular validity: Convincing the court that it was taught without precluding the possibility of measuring it. In G. F. Madaus (Ed.), *The courts, validity, and minimum competency testing* (pp. 115-32). MA: Kluwer-Nijhoff.

Linn, R. L. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17*(2), 28-30.

Linn, R. L., Baker, E .L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Lynch, B. K., & Davidson, F. (1994). Criterion-referenced language test development: linking curricula, teachers, and tests. *TESOL Quarterly, 28*(4), 727-743.

Madaus, G. F. (1983). Minimum competency testing for certification: the evolution and evaluation of test validity. In G. F. Madaus (Ed.), *The courts, validity, and minimum competency testing* (pp. 21-61). MA: Kluwer-Nijhoff.

Markee, N. (1993). The diffusion of innovation in language teaching. *Annual Review of Applied Linguistics, 13*, 229-43.

McChesney, B. (1990). University office hours: What professors and teaching assistants say to students. In G. Barnes, M. Berns, & C. Madden (Eds.), *Training of International Teaching Assistants*. Preconference symposium presented at the 24th Annual TESOL Convention, San Francisco.

McNamara, T. (1996). *Second language performance measuring*. London and New York:Longman.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*(2), 16-18.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-27.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessment. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241-56.

Morris, B. (1972). *Objectives and perspectives in education: Studies in educational theories*. London: Routledge and Kegan Paul.

Morrow, K. (1986). The evaluation of tests of communicative performance. In M. Portal (Ed.), *Innovations in language testing*. London: NFER/Nelson.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*, 229-258.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5-12.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17*(2), 6-12.

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

Myers, C., & Douglas, D. (1991). *The ITA lab assistant: Strategies for success*. Paper presented at the Annual NAFSA Convention, Boston.

Pearson, I. (1988). Tests as levers for change. In D. Chamberlain, & R. Baumgardner (Eds.), *ESP in the classroom: Practice and evaluation* (ELT Document 128, pp. 98-107). London: Modern English Publications.

Ponder, R. (1991). *The TSE: A language teacher's critique*. Paper presented at the 25th Annual TESOL Convention, New York.

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13.

Popham, W. J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice, 17*(2), 13-16.

Rounds, P. (1987). Characterizing successful classroom discourse for NNS teaching assistants. *TESOL Quarterly, 21*(4), 643-671.

Schutz, N. W., and Derwing, B. L. (1981). The problem of needs assessment in English for specific purposes: Some theoretical and practical considerations. In R. Mackay, & J. P. Palmer (Eds.), *Languages for specific purposes: Program design and evaluation.* Rowley, MA: Newbury House Publishers, Inc.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language.* Cambridge: Cambridge University Press.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19,* 405-450.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-8, 13, 24.

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

Smith, H. J. (1989). *ELT project success and the management of innovation.* Unpublished Manuscript. University of Reading: Centre for Applied Language Studies.

Smith, J., Meyers, C., & Burkhalter, A. (1992). *Communicate: Strategies for International Teaching Assistants.* Englewood Cliffs, NJ: Regents/Prentice Hall.

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing, 2* (1), 31-40.

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

Swain, M. (1984). Large-scale communicative language testing: A case study. In S. J. Savignon, & M. S. Berns (Eds.), *Initiatives in communicative language testing* (pp. 158-201). Reading, MA: Addison-Wesley.

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

Taleporos, E. (1998). Consequential validity: A practitioner's perspective. *Educational Measurement: Issues and Practice, 17*(2), 20-23, 34.

Tanner, M. (1991). *NNSTA-student interaction: An analysis of TA's questions and students' responses in a laboratory setting*. Unpublished Doctoral Dissertation. University of Pennsylvania, Philadelphia.

Underhill, N. (1989). *Testing spoken language*. Cambridge: Cambridge University Press.

Upshur, J. A. (1979). Functional proficiency theory and a research role for language tests. In E. Briere and F. B. Hinofotis (Eds.), *Concepts in language testing: Some recent studies* (pp.75-100). Washington, DC: TESOL.

Valdman, A. (Ed.). (1988). The assessment of foreign language oral proficiency. *Studies in Second Language Acquisition, 10*(2).

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

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

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

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

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

Woods, A., Fletcher, P., & Hughes, A. (1991). *Statistics in language studies*. Cambridge: Cambridge University Press.

Zimiles, H. (1996). Rethinking the validity of psychological assessment. *American Psychologist, 51*(9), 980-81.

# APPENDIX ONE
## Part A

### Test of Spoken Language Ability for International Teaching Assistants (ITAs)

**General Directions:** In this test, the test-takers will be required to demonstrate how well they can talk about themes and topics in their own field of specialization using English language as the medium. The approximate time for the entire test is between 15-20 minutes. The whole process in sections two and three will be videotaped for the purpose of revision and precision. The test will be rated by a panel of observers including three undergraduate students from the test-taker's department and two ESL instructors.

**I. Introduction phase**

In this section of the test, the test takers will be required to answer some questions about themselves. The purpose of this phase, which should not last more than five minutes, is to allow the candidates to establish themselves in readiness for the main part of the test. The questions are asked by ESL instructors and depending on the time allocated to this part, test takers can give shorter answers to two or more questions or a longer answer to only one question.

Questions in this phase can be related to the students themselves, their educational background, their home country, their interests, their hopes and future plans, the relevance of what they are doing here to their life in their country, their reasons for choosing Canada in general and UVic in particular for studying ... etc. Test takers are not scored for what they say in this section since it is a quick warm-up before the main phase. This phase might

be waived for those candidates who have taken the test at least once in the past or are familiar enough with the panel members and test format.

**II. Presentation**

In this section the test taker will be required to present a maximum 10-minute talk related to his/her major field of specialization as if s/he were talking in front of his/her undergraduate students in one of the course sessions to which s/he is or expects to be assigned as a TA.

The subject is a topic of his/her own choice for which s/he has prepared in advance. The setting is a classroom setting including necessary accessories such as blackboard, over-head, ... etc. The test taker should be informed about all these at least 24 hours before the test. S/he should also be instructed that s/he will be graded both on the accuracy and the appropriateness of his/her English as well as how well s/he plans and presents the idea. They should also expect questions or requests for clarification in the middle of their talk.

**III. Question/Answers**

In this phase, the panelists ask questions based on what has been presented in section two. The questions might require the test taker to elaborate the original topic or be involved in a new unprepared but related topic. The time allocated to this phase is at most 5 minutes.

# APPENDIX ONE
## Part B

### Rating Instrument

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 viewing the videotapes after the test is over. Raters are supposed to make sure that they understand both the ability components listed here and the rating scale used for rating (Appendix One, Parts C and D) before the test.

Name: Date: Rater:

Directions: Please circle 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 | | | | | |

# APPENDIX ONE
## Part C

**Rating Scale:**

| Ability Levels | None | Limited |
|---|---|---|
| | 0 | 1 |
| **Ability areas** | | |
| A. Grammatical knowledge | | |
| 1. Vocabulary | Not demonstrated | Limited grasp of vocabulary, many instances of inaccurate use of words interfering with the intelligibility |
| 2. Grammar | No control of grammatical rules | Few basic syntactic structures accurately used, limited knowledge of syntax interferes with intelligibility |
| 3. Pronunciation | Unintelligible speech | Limited knowledge of phonological rules and foreign stress and intonation resulting in occasional unintelligibility |
| B. Textual knowledge | | |
| 4. Cohesion | Not demonstrated | Few cohesive devices used, mostly detached unrelated sentences |
| 5. Conversational organization | Not demonstrated | Topic is roughly introduced, no clear way of presenting details |
| C. Functional knowledge | | |
| 6. Use of ideational, manipulative, and heuristic functions | Not demonstrated | Limited use of functions, often unclear or irrelevant to the topic |
| D. Sociolinguistic knowledge | | |
| 7. Dialect | No evidence of knowledge of standard language | Little evidence of the use of the standard language |

| Moderate<br>2 | Extensive<br>3 | Complete<br>4 |
|---|---|---|
| Moderate knowledge of vocabulary, often used incorrectly | Vast knowledge of vocabulary, seldom used incorrectly | Complete knowledge of vocabulary used with accuracy |
| Knowledge of a medium range of syntactic structures used with good accuracy | Vast knowledge of syntactic structures, few errors | Complete knowledge of syntax, evidence of accurate use of all structures with no limitation |
| Medium range of phonological rules correctly applied, some consistent phonemic errors, foreign stress and intonation | Vast knowledge of phonological rules, few non-native pronunciation errors but always intelligible | Complete control of phonological rules, pronunciation as clear and smooth as that of a native speaker |
| Moderate range of cohesive devices used, more common ones often used correctly | Vast knowledge and accurate use of the textual cohesion, few errors | Complete knowledge of cohesion, control and accurate use of all cohesive devices |
| Evidence of moderate knowledge of conversational organization methods, procedures are set up but their relationship to each other and to the topic are not clear enough | Extensive knowledge of organizational methods, step by step presentation, divisions relevant to the topic, high level of accuracy | Complete knowledge of organizational methods, clear relevant procedures, smooth native-like transitions between divisions |
| Moderate knowledge of functional language, formulaic patterns used accurately, little evidence of the accurate use of multiple functions outside the range | Extensive use of different functional expressions in connected speech, few errors | Complete knowledge of functional language, accurate and appropriate use of individual and multiple functions |
| Evidence of both standard and non-standard language | Extensive use of standard dialect, occasional uses of non-standard expressions | Complete knowledge of standard dialect, no instances of non-standard language use |

| | | |
|---|---|---|
| 8. Register | No awareness of register variations | Limited knowledge of relatively formal register, inaccurate use of register in either or both formulaic and substantive language |
| E. Strategic competence | | |
| 9. Goal-setting | Not demonstrated | Little evidence of deliberate goal setting, generally irrelevant to the topic, no reference to the old information |
| 10. Use of verbal strategies | Not demonstrated | Limited knowledge of effective strategies resulting in frequent communication breakdowns |
| 11. Use of non-verbal strategies | Not demonstrated | Gestures/teaching aids rarely used to supplement verbal communication |
| 12. Achievement of communicative goal through production | Total avoidance/ communication breakdown | Frequent interruptions, long pauses and unsuccessful completion of the communication |
| 13. Achievement of communicative goal through comprehension | No evidence of understanding the language of input | Limited ability to relate to the audience resulting in insufficient and/or irrelevant response |

| | | |
|---|---|---|
| Moderate knowledge of formal and informal register, occasional inappropriate use of either register | Extensive knowledge of moderately formal register in both formulaic and real discourse, few errors | Accurate and appropriate use of a moderately formal register in all forms of discourse |
| Moderate use of goal-setting strategies, communicative goals occasionally unclear or irrelevant to the topic | Communicative goals set with high level of accuracy, few errors | Completely successful goal-setting, clear relationship provided between the old and new information |
| Moderate knowledge of basic verbal strategies used with accuracy, occasional inappropriate communication strategies | Extensive knowledge of types and effectiveness of communication strategies, few errors | Complete knowledge of verbal strategies, accurate and appropriate use of a variety of effective strategies |
| Effective use of a limited number of non-verbal strategies, occasionally ineffective | Extensive use of a variety of non-verbal strategies, few unsuccessful attempts | Complete knowledge of types and efficiency of non-verbal strategies, all used effectively |
| Occasional incomplete tasks due to inadequacy of language knowledge and prompt communication strategy use | Communicative goal achieved through efficient strategic use of areas of language knowledge, few errors | Complete ability to use the knowledge of language accurately and appropriately to achieve the communicative goal |
| Moderate comprehension of the language of input, occasional requests for clarification or repetition | Extensive comprehension and interpretation of the language of input, few errors | Complete ability to understand the language of input, no repetition or elaboration required |

# APPENDIX ONE
## Part D

### Description of the Ability Components in the Rating Instrument

A. Grammatical Knowledge

1. Vocabulary: control of general and field specific vocabulary, choice of semantically appropriate words

2. Grammar: control of syntactic structures and morphological rules

3. Pronunciation: including vowel and consonant sounds, and syllable stress to the extent that they interfere with the communication of meaning

B. Textual Knowledge

4. Cohesion: the use of overt linking devices and appropriate transitions which add to the clarity of expression and thus help the communication run more smoothly

5. Conversational organization: including the techniques the examinees use to open, develop, and terminate the discussion; use of common methods of organization

C. Functional knowledge

6. Use of ideational, manipulative, and heuristic functions: whether or not the utterances are appropriate for performing specific functions such as the expression and exchange of ideas and knowledge, making suggestions and comments, establishing relationships and so forth

D. Sociolinguistic knowledge: extent to which utterances are appropriately related to the Characteristics of the setting

7. Dialect: standard/non-standard English; standard English is the kind of English that educated people use in public and accept as appropriate for almost any situation. It includes formal and informal levels of language but not slang

8. Register: appropriate use of formal/informal register depending on the context of language use

D. Strategic competence

9. Goal-setting: ability to relate to the audience by using appropriate communicative goals

10. Use of verbal strategies: the extent to which the examinee makes use of verbal communication strategies either to make his/her point more forcefully or to overcome possible linguistic gaps (e.g., paraphrase, circumlocution, exemplification,... etc.)

11. Use of non-verbal strategies: the extent to which the examinee supplements his verbal language by non-verbal communicative strategies (e.g., gestures, pauses)

12. Achievement of communicative goal through production: examinee's ability in matching his/her communicative goals and the linguistic devices at his/her disposal to the purpose of production

13. Achievement of communicative goal through comprehension: examinee's success in understanding the verbal/non-verbal language of input (questions, comments, requests for clarification, gestures, ...etc.)

# APPENDIX TWO
# Part A

## Test of Written Language Ability for Advanced EFL Learners

**Directions to test users:** In this test, the test-takers will be required to demonstrate how well they can write a proposal about themes and topics in their field of specialization (Applied Linguistics) using English language as the medium. The topic/prompt will be selected (by the instructor) from among the technical topics known to the test-takers. The test has only one phase to be completed in no more than 120 minutes. It will be rated by three ESL instructors with background in applied linguistics using the rating instrument in Appendix Two, Part B. As one of the components in the rating instrument deals with the "topical knowledge" of the test-takers, their performance will be rated not only for their ability to use language in writing but also for how they use it in relation to topics in Applied Linguistics.

# APPENDIX TWO
## Part B

### Rating Instrument

Name: Date: Rater:

Directions: Please circle 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 | | | | | |
| B. Textual knowledge | | | | | |
| 3. Cohesion | | | | | |
| 4. Rhetorical organization | | | | | |
| C. Functional knowledge | | | | | |
| 5. Use of ideational and manipulative functions | | | | | |
| D. Sociolinguistic knowledge | | | | | |
| 6. Dialect | | | | | |
| 7. Register | | | | | |
| E. Strategic competence | | | | | |
| 8. Goal-setting | | | | | |
| 9. Achievement of communicative goal through written production | | | | | |
| F. Topical knowledge | | | | | |
| 10. Use of topical knowledge as the information base | | | | | |

# APPENDIX TWO
## Part C

**Rating Scale:**

| Ability Levels | None | Limited |
|---|---|---|
| | 0 | 1 |
| **Ability areas** | | |
| A. Grammatical knowledge | | |
| 1. Vocabulary | Not demonstrated | Small range of vocabulary, a noticeably inappropriate choice of words |
| 2. Grammar | No control of grammatical rules | Few basic syntactic structures accurately used, limited knowledge of punctuation rules |
| B. Textual knowledge | | |
| 3. Cohesion | No evidence of knowledge of cohesion | Few instances of the use of cohesive devices, sentences mostly detached and unrelated |
| 4. Rhetorical organization | Not demonstrated | Limited use of organizational patterns, no clear way of presenting details |
| C. Functional knowledge | | |
| 5. Use of ideational, manipulative, and heuristic functions | No evidence of functional knowledge | Limited use of very few formulaic functions, often unclear or irrelevant to the topic |
| D. Sociolinguistic knowledge | | |
| 6. Dialect | No evidence of knowledge of standard language | Little evidence of the use of the standard language |
| 7. Register | No awareness of register variations | Limited knowledge of formal register, inaccurate use of register in either or both formulaic and substantive language |

| | | |
|---|---|---|
| E. Strategic competence | | |
| 8. Goal-setting | Not demonstrated | Little evidence of deliberate goal setting, generally irrelevant to the topic, no reference to the old information |
| 9. Achievement of communicative goal through written production | Total avoidance/ communication breakdown | Frequent interruptions and unsuccessful completion of the communicative task resulting in incomprehensibility |
| F. Topical knowledge | | |
| 10. Use of topical knowledge as the information base | Not demonstrated | Very limited knowledge of the topic or part of the topic |

| Moderate | Extensive | Complete |
|---|---|---|
| 2 | 3 | 4 |
| Demonstrates some range of either or both general and technical vocabulary, inappropriate word choice outside the range | Vast knowledge of vocabulary, seldom used incorrectly | Complete knowledge of general and technical vocabulary used with accuracy |
| Demonstrates some variety of syntactic structures and rules of punctuation with good accuracy | Vast knowledge of syntactic structures and rules of punctuation, few errors | Complete knowledge of syntax, evidence of accurate use of all structures and punctuation rules with no limitation |
| Moderate range of cohesive devices used, more common ones often used correctly | Vast knowledge of accurate use of textual cohesion with few errors | Complete knowledge of cohesion, control and accurate use of all cohesive devices |
| Evidence of moderate knowledge of textual organization methods, procedures are set up but their relationship to each other and to the topic are not clear enough | Extensive knowledge of organizational methods, clear step by step presentation, divisions relevant to the topic, high level of accuracy | Complete knowledge of organizational methods, clear relevant procedures, smooth transitions between divisions |
| Moderate knowledge of functional language, formulaic patterns used accurately, little evidence of the accurate use of multiple functions outside the range | Extensive use of different functional expressions in connected discourse, few errors | Complete knowledge of functional language, accurate and appropriate use of individual and multiple functions |
| Evidence of both standard and non-standard language | Extensive use of standard dialect, occasional uses of non-standard expressions | Complete knowledge of standard dialect, no instances of non-standard language use |
| Moderate knowledge of formal register, occasional inappropriate use of informal register | Extensive knowledge of formal register in both formulaic and real discourse, few errors | Accurate and appropriate use of a formal register in all forms of discourse |

| | | |
|---|---|---|
| Moderate use of goal-setting strategies, communicative goals occasionally unclear or irrelevant to the topic | Communicative goals set with high level of accuracy, few errors | Completely successful goal-setting, all tasks appropriately chosen |
| Occasional incomplete tasks and conceptual confusion due to inappropriate planning | Communicative goal achieved through efficient planning and appropriate use of areas of language knowledge, few errors | Complete ability to plan and use the knowledge of language accurately and appropriately to achieve the communicative goal |
| Relevant topical knowledge moderately used with accuracy | Vast knowledge of the relevant topic, few limitations | Complete mastery of the relevant topical information |

# APPENDIX TWO
## Part D

### Description of the Ability Components in the Rating Instrument

A. Grammatical Knowledge

1. Vocabulary: control of general and field specific vocabulary, choice of semantically appropriate words

2. Grammar: control of syntactic structures and morphological rules

B. Textual Knowledge

3. Cohesion: the use of overt linking devices and appropriate transitions which add to the clarity of expression and thus help the communication run more smoothly

4. Rhetorical organization: the overall conceptual structure of the text including the techniques the examinees use to open, develop, and terminate the discussion; use of common methods of organization

C. Functional knowledge

5. Use of ideational, manipulative, and heuristic functions: whether or not the sentences are appropriate for performing specific functions such as the expression and exchange of ideas and knowledge, making suggestions and comments, establishing relationships and so forth

D. Sociolinguistic knowledge: extent to which utterances are appropriately related to the characteristics of the setting

6. Dialect: use of standard English, the kind of English that educated people use in public and accept as appropriate for almost any situation. It includes formal and informal levels of language but not slang

7. Register: use of formal register

E. Strategic competence

8. Goal-setting: ability to relate to the readers by using appropriate communicative goals

9. Achievement of communicative goal through written production: examinee's ability in carefully planning and matching his/her communicative goals and the linguistic devices at his/her disposal to the purpose of the production

F. Topical knowledge

10. Use of topical knowledge as the information base: knowledge and use of the relevant topical information (applied linguistic issues in this case)

# APPENDIX THREE
## Part A

### Test of Written Language Ability for Intermediate EFL Learners

**Instructions:**

In this test, you are required to demonstrate how well you can write an essay about one of the following topics. You will have 120 minutes to complete the task. Your essay will be graded for clarity of expression, logical development of ideas, their relationship to each other, the general organizational method, and control of grammar and vocabulary; not for the specific ideas expressed or implied.

**Prompt:**

Addressing the general public in a moderately formal language, write an essay of classification or comparison/contrast on one of the following topics:

1. Types of friends / two friends
2. Types of books / two books
3. Characteristics of a good marriage / two couples
4. Characteristics of a successful student / two students

# APPENDIX THREE
## Part B

### Rating Instrument

Name: Date: Rater:

Directions: Please circle 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 | | | | | |
| B. Textual knowledge | | | | | |
| 3. Cohesion | | | | | |
| 4. Rhetorical organization | | | | | |
| C. Functional knowledge | | | | | |
| 5. Use of ideational and manipulative functions | | | | | |
| D. Sociolinguistic knowledge | | | | | |
| 6. Use of figurative language and cultural references | | | | | |
| 7. Register | | | | | |

# APPENDIX THREE
## Part C

**Rating Scale:**

| Ability Levels | None | Limited |
|---|---|---|
| | 0 | 1 |
| **Ability areas** | | |
| A. Grammatical knowledge | | |
| 1. Vocabulary | Not demonstrated | Small range of vocabulary, many instances of inaccurate use of words |
| 2. Grammar | Not demonstrated | Only a few syntactic structures accurately used, limited knowledge of syntax and punctuation |
| B. Textual knowledge | | |
| 3. Cohesion | No evidence of knowledge of cohesion | Few instances of the use of cohesive devices, sentences mostly detached and unrelated |
| 4. Rhetorical organization | Not demonstrated | Limited use of organizational patterns, no clear way of presenting details |
| C. Functional knowledge | | |
| 5. Use of ideational and manipulative functions | No evidence of functional knowledge | Limited use of very few formulaic functions, often unclear or irrelevant to the topic |
| D. Sociolinguistic knowledge | | |
| 6. Use of figurative language and cultural references | Not demonstrated | Little evidence of the use of figurative language and cultural references |
| 7. Register | No awareness of register variations | Limited knowledge of relatively formal register, inaccurate use of register in either or both formulaic and substantive language |

| Moderate | Extensive | Complete |
|---|---|---|
| 2 | 3 | 4 |
| Moderate use of either or both general and technical vocabulary, often used incorrectly | Vast knowledge of vocabulary, seldom used incorrectly | Complete knowledge of general and technical vocabulary used with accuracy |
| Use of a medium range of syntactic structures and punctuation with good accuracy | Vast knowledge of syntactic structures and punctuation, few errors | Complete knowledge of syntax, evidence of accurate use of all structures and punctuation with no limitation |
| Moderate range of cohesive devices used, more common ones often used correctly | Vast knowledge of accurate use of textual cohesion with few errors | Complete knowledge of cohesion, control and accurate use of all cohesive devices |
| Evidence of moderate knowledge of textual organization methods, procedures are set up but their relationship to each other and to the topic are not clear enough | Extensive knowledge of organizational methods, clear step by step presentation, divisions relevant to the topic, high level of accuracy | Complete knowledge of organizational methods, clear relevant procedures, smooth transitions between divisions |
| Moderate knowledge of functional language, accurate use of formulaic patterns, multiple functions outside the range often used inaccurately | Extensive use of different functional expressions in connected discourse, few errors | Complete knowledge of functional language, accurate and appropriate use of individual and multiple functions |
| Moderate knowledge of figurative language and cultural references, often used appropriately | Extensive knowledge of figurative language and cultural references, occasionally used inappropriately | Complete knowledge and accurate use of figurative language and cultural references |
| Moderate knowledge of relatively formal register, occasional use of inappropriate register | Extensive knowledge of relatively formal register in both formulaic and real discourse, few errors | Accurate and appropriate use of a relatively formal register in all forms of discourse |

# APPENDIX THREE
## Part D

### Description of the Ability Components in the Rating Instrument

A. Grammatical Knowledge

1. Vocabulary: control of general and specific vocabulary, choice of semantically appropriate words

2. Grammar: control of syntactic structures, morphological rules, and rules of punctuation

B. Textual Knowledge

3. Cohesion: the use of overt linking devices and appropriate transitions which add to the clarity of expression and thus help the text flow more smoothly

4. Textual organization: including the techniques the examinees use to open, develop, and terminate the discussion; use of common methods of organization

C. Functional knowledge

5. Use of ideational and manipulative functions: whether or not the sentences are appropriate for performing specific functions such as the expression and exchange of ideas and knowledge, making suggestions and comments, establishing relationships and so forth

D. Sociolinguistic knowledge: extent to which utterances are appropriately related to the characteristics of the setting

6. Use of figurative language and cultural references: the ability to show – rather than tell – the ideas by going beyond the literal meaning of words through the appropriate use of figurative language and cultural references

7. Register: relatively formal register