T.C.
DİCLE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ
YABANCI DİLLER EĞİTİMİ ANABİLİM DALI
İNGİLİZ DİLİ EĞİTİMİ BİLİM DALI

ÜNİVERSİTELERARASI KURUL YABANCI DİL SINAVI'NIN GERİ ETKİ (WASHBACK) AÇISINDAN İNCELENMESİ (DİCLE ÜNİVERSİTESİ ÖRNEĞİ)

Emrullah DAĞTAN

Tez Danışmanı: Yrd. Doç. Dr. Nilüfer BEKLEYEN

YÜKSEK LİSANS TEZİ

DİYARBAKIR / 2012

TURKISH REPUBLIC
DICLE UNIVERSITY
GRADUATE SCHOOL OF EDUCATIONAL SCIENCES
DEPARTMENT OF FOREIGN LANGUAGE EDUCATION
ENGLISH LANGUAGE TEACHING PROGRAMME

AN EXAMINATION ON THE WASHBACK EFFECT OF THE INTERUNIVERSITY BOARD FOREIGN LANGUAGE TEST (THE EXAMPLE OF DICLE UNIVERSITY)

Emrullah DAĞTAN

Supervisor: Asst. Prof. Nilüfer BEKLEYEN

MASTER'S THESIS

DİYARBAKIR / 2012

Approval of the Graduate School of Educational Sciences

This work has been accepted as a thesis for the degree of Master of Arts in English Language Teaching Programme.

Supervisor : Asst. Prof. Nilüfer BEKLEYEN

Member of Examining Committee : Prof. Mehmet Siraç İNAN

Member of Examining Committee : Asst. Prof. Bayram AŞILIOĞLU

Approval

This is to certify that the signatures above belong to the members of examining committee whose names are written.

Director

ÖZ

ÜNİVERSİTELERARASI KURUL YABANCI DİL SINAVI'NIN GERİ ETKİ (WASHBACK) AÇISINDAN İNCELENMESİ (DİCLE ÜNİVERSİTESİ ÖRNEĞİ)

Emrullah DAĞTAN

Yüksek Lisans Tezi, İngiliz Dili Eğitimi Anabilim Dalı

Danışman: Yrd. Doç. Dr. Nilüfer BEKLEYEN

Nisan 2012, 116 sayfa

Bu çalışmanın amacı 'Üniversitelerarası Kurul Yabancı Dil Sınavı'nın (ÜDS) "washback effect" denilen "geri etki" olarak tanımlanan teknik bir boyut açısından incelenmesidir. ÜDS, temelde Yükseköğretim kurumlarında görevli olan veya bu kurumlarda çalışmak isteyenlerin yabancı dil bilgilerinin ölçülmesinde kullanılan merkezi bir sınavdır. "Doçentlik Sınav Yönetmeliği' uyarınca doçent adaylarının doçentlik başvurusu yapabilmek için bu sınavdan 65 veya üzerinde bir puan almaları gerekmektedir. Bu durumu çalışma zemini olarak kabul eden çalışmanın evreni olarak Diyarbakır İl Merkezi'nde bulunan Dicle Üniversitesi'nde görev alan öğretim üyeleri seçilmiştir. Çalışmaya toplam 575 öğretim üyesi dahil edilmiş olup bunlardan 161'i çalışmaya katılmıştır. Bunların 144'üyle birebir görüşülmüş, geriye kalan 17 kişi ise çalışmaya e-mail yoluyla dâhil edilmiştir. Merkez yerleşkede bulunan 13 ayrı Fakülte ve Yüksekokuldan çalışmaya dahil edilen katılımcılar arasında, en büyük grubu Yardımcı Doçentlerin oluşturduğu tespit edilmiştir.

Çalışmadaki nicel veriler açık uçlu, çoktan seçmeli ve Likert ölçekli sorulardan oluşan bir anket yardımıyla elde edilmiştir. Nitel veriler oluşturmak amacıyla 10 katılımcıyla bire bir görüşmeler yapılmıştır. Nicel verilerin analizinde SPSS 17.0 programı kullanılırken nitel veriler içerik analizi yöntemiyle değerlendirilmiştir.

Araştırmanın sonuçlarına göre, özel ders, ÜDS'ye hazırlanma yöntemleri arasında en etkili yöntem olarak belirlenmiştir. ÜDS başarısı göz önünde

bulundurulduğunda bayanların erkeklere göre, İngilizce dilini seçenlerin Almanca ve Fransızca dillerini seçenlere göre ve Sağlık Bilimleri alanında sınava girenlerin diğer alanlarda giren kişilere göre daha başarılı olduğu ortaya çıkmıştır. ÜDS'deki soru grupları düşünüldüğünde, kelime bilgisi sorularının diğerlerine göre daha zorlayıcı olduğu, ayrıca bu soruların ÜDS çalışmalarında en çok vurgulanan sorular olduğu anlaşılmıştır. ÜDS için düşünülen değişiklikler arasında ÜDS'nin dört beceriyi kapsayacak şekilde değişmesi gerektiği en çok oylanan seçenek olarak belirlenmiştir. Verilerin değerlendirilmesi sonucunda ÜDS'nin en çok okuma becerisini geliştirmeye yardımcı olduğu, dinleme, konuşma ve yazma becerilerinin ise sınava hazırlanan adaylar tarafından genellikle ihmal edildiği ortaya çıkmıştır.

Anahtar Kelimeler: Washback, Sınav Etkisi, Pozitif Washback, Negatif Washback, IBFLE

ABSTRACT

AN EXAMINATION ON THE WASHBACK EFFECT OF THE INTERUNIVERSITY BOARD FOREIGN LANGUAGE TEST (THE EXAMPLE OF DICLE UNIVERSITY)

Emrullah DAĞTAN

Master's Thesis, English Language Teaching

Supervisor: Asst. Prof. Nilüfer BEKLEYEN

April 2012, 116 pages

In this study, the effect of a problematizing foreign language test in Turkey, called IBFLE, is studied through a technical perspective called "washback effect". IBFLE is a proficiency test, which is essentially taken by university faculty members and their candidates to prove their linguistic competence. According to the "Exam Regulations for the Candidates of Associate Professor", the candidates are required to get a score of 65% or over to apply for the position of associate professor. Based on this fact, the population of the study included the full-time faculty members working at Dicle University, Diyarbakir. The total 575 faculty members were targeted, but a total of 161 participants took part. Of these, 144 were visited in person and the remaining 17 participated via e-mails. The participants were randomly chosen from 13 different faculties / schools located in the main campus of the University, and the majority of the respondents consisted of Assistant Professors.

Quantitative data was collected by means of a questionnaire, consisting of open-ended and multiple-choice sections as well as Likert-type items. For qualitative data, semi-structured face-to-face interviews were conducted with 10 participants. Statistical analysis of the quantitative data was performed using SPSS 17.0, and the analysis of the qualitative data was carried out by means of content analysis.

The results revealed that private tuition is the most common study style for IBFLE takers. In terms of IBFLE success, females seem to be better than males, the

participants taking English do better than the ones choosing German and French, and the group of Health Sciences have a higher success rate than the other two groups. Among the question types in IBFLE, the vocabulary questions are revealed as the most challenging and also as the ones that are most emphasized in IBFLE preparations. For the changes proposed for IBFLE, the respondents are most positive in the need for an all-inclusive test which would assess all four skills in the language. As regards the washback, it was revealed that the IBFLE is best useful in promoting the reading skill and that the remaining skills, *i.e.* listening, speaking, and writing are often ignored in IBFLE preparations.

Key Words: Washback, Impact, Positive Washback, Negative Washback, IBFLE

ACKNOWLEDGEMENTS

I am pleased to acknowledge the people who have granted their invaluable support throughout my thesis process.

First and foremost, I owe my deepest gratitude to my thesis advisor, Asst. Prof. Nilüfer BEKLEYEN, for her excellent guidance, caring, patience, and providing me with an excellent atmosphere. Without her, it would have been impossible to accomplish this ordeal.

Indeed, this is a great opportunity to express my respect to the committee members, Prof. M. Siraç İNAN and Asst. Prof. Bayram AŞILIOĞLU, whose passion for making contributions to my thesis has broadened my academic vision.

Finally, this thesis is dedicated to my wife and my dear daughter. I believe that their never-ending support will always be the primary source of my life energy.

TABLE OF CONTENTS

CHAPTER THREE

METHODOLOGY

CHAPTER FOUR

RESULTS

CHAPTER FIVE

DISCUSSION AND CONCLUSION

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER ONE

INTRODUCTION

1.1 Presentation

This chapter presents the background information about the content and reasoning behind the essence of the subject matter, together with its purpose, underlying motivations, research questions, and definitions of key terms and abbreviations.

1.2 Statement of the Problem

Foreign language proficiency is a criterion for various processes, such as advancing in an academic career, personal development, professional promotion, prestige and so on. Indeed, proficiency in English, which is regarded as the foremost second language in the world, certainly has more to do in widening a learner's horizon in every phase of the educated world (Crystal, 1989: 358; 1995: 106; Cook, 2003: 25, 26). More and more sophisticated materials are being produced to promote the quality of the process of learning English with the aim of luring as many people as possible. As a process, it has gained worldwide popularity, creating a sense of competition, in which there are not only competitors, *i.e.* learners, but also some mechanisms which obligatorily come up depending on the necessities of beneficiaries. Of these, *assessment* plays the most important part (Cheng, 1998: 254; Madsen, 1983: 4). Just like other processes, language learning process, too, entails a formal and standardised testing system which could be employed both to test the real outcome of this process and to promote higher scores through the sense of competition (Cheng, 2000: 3). These testing systems have gradually been transformed to cram their test takers into a challenging atmosphere of *win-or-lose*, rather than pledging them a hope of good opportunities. Coined as 'high stakes' tests, these assessment systems "influence the way that students and teachers behave as well as their perceptions of their own abilities and worth" (Hayes, 2003: 1). As the name suggests, these are the "tests from which results are used to make significant educational decisions about schools, teachers, administrators, and students" (Amrein & Berliner,

2002a: 1). They affect processes of teaching and learning as well. Used as the indicators of winning or losing, they impose a minimum score, accepting the higher scores as 'passed' and casting the lower ones as 'failed'. Hence they inevitably have to be quite challenging to be reasonably eliminative.

Among the high-stakes tests in Turkey, *Üniversitelerarası Kurul Yabancı Dil Sınavı* (ÜDS; hereafter referred to as *Interuniversity Board Foreign Language Examination* – IBFLE) is a well-known test which is a prerequisite for gaining an academic title at universities or applying for a position in state institutions and private sectors. IBFLE is not only *well-known* but also *notorious* for test takers, which pertains to the dichotomy of *necessity* vs. *challenge*: its widespread implementation due to the greatness of the number of applicants denotes its well-known necessity while its challenging style which causes many test takers to fail successively refers to its problematizing side. IBFLE takes much of its criticism upon its *allegedly* mechanical and recitative style, which is overwhelmed with vocabulary and grammar, both by the ones who have passed and the ones who have not. Test takers go beyond criticism, maintaining that IBFLE is aimless since it does not fully meet the linguistic standards of academicians, who exclusively need productive skills in performing their academic activities. Depending on these needs, the test is said to fall behind in representing the essential qualities of its target group. Suen and Yu (2006: 51) devise this matter as "construct underrepresentation", explaining that a test is devoted to assess a "construct" in order to represent the *must-have* qualifications of its test takers. When it does not, the test is considered to "underrepresent" the targeted qualifications. Centring upon only one of the four skills, *i.e.* reading, IBFLE assumes this description unless it is complicated with other skills.

Related to the targeted outcome of a test, IBFLE inherits another technical term by Suen and Yu (2006): a test is intrinsically expected to assess some required knowledge or abilities; however, it may reveal some "irrelevant" details about its test takers, such as their socioeconomic status, gender, or marital status. Labelled as "construct irrelevance", this procedural mishap inadvertently causes drawbacks in the reliability and validity of the test (pp. 50, 51). As for the irrelevance in IBFLE, it is something other than personal information. It principally discloses to what extent test takers are familiar with its testing

style and intricacies rather than proving the knowledge or skills assessed in the test. Therefore, IBFLE is better known for its intriguing style rather than its educational power of assessment.

The remaining question to be asked refers to *why* and *how* the IBFLE takers – no matter whether they be the ones that passed or failed – find the test challenging and even aimless despite being educated to it. Obviously the answer has something to do with the influence of the test on the processes of teaching and learning. The notion to clarify this interwoven relationship is called *washback*, or *backwash*, which signifies "the influence of tests on teaching" (Alderson & Wall, 1992: 2), "the impact that tests have on teaching and learning" (Alderson & Banerjee, 2001: 214), or "the effect of testing on teaching and learning" (Hughes, 1992: 1). The concept is concerned both with curriculum and testing. The curriculum aspect is related to the materials and methodologies that are specifically arranged up to the requirements of a test, whereas the latter is about the materials and precautions which are taken into consideration during the pre- and post-test periods. As Cheng (1997: 38) puts it; washback is "an active direction and function of intended curriculum change by means of the change of public examinations", appraising washback as a pivotal element that governs curriculum changes. The concept of washback exists in two fold (Alderson & Wall, 1992: 3). The first yet the indirect one relates to its educational changes, which would co-affect its by-products such as syllabus design, textbook, course design, teaching methodologies and approaches. Most of these elements, if not all, go through a partial change upon the reverberations of a test. As for the second, it is rather individual and direct. Each test taker feels the influence of a test throughout his/her entire academic life even after he/she has achieved the test. Suitably, the significance of this dichotomy necessitates a two-way inspection through the complexity of the notion of washback, which has penetrated into the test in discussion. Washback could be either in a positive or negative way (Cheng, 1997: 40). Just because the term *washback* mistakably associates with the word *influence*, it is mostly regarded as something bad or harmful; *i.e.* negative. However, even a poor exam could prove positive if it causes the learners and teachers to do good things (Alderson & Wall, 1992: 6). This being the case, one should dismiss the same myth for IBFLE, which refers to the unofficial allegations regarding its *so-called* negativity. The virtue of the present study is

powered by these allegations, which stand against the well-known indispensability of IBFLE.

1.3 Purpose and Significance of the Study

A literature review spanning over the last decade has revealed a scarcity in the number of studies specifically focussing on washback effect itself and more importantly on the washback effect of standardized tests in Turkey. The present study takes its virtue from this scarcity and specifically from the absence of such studies on IBFLE. With this regard, the study aims an exhaustive scrutiny into the test takers' preparation styles as well as their educational backgrounds. Targeting a group of IBFLE-wise university faculty members and structured on an institution-wide needs-analysis research, the present study is dedicated to find out the underlying influence of the test on the test takers' educational and academic lives, in an attempt to suggest practical solutions for future test takers and also for policymakers. Further, it is aimed to delve into the deep-rooted drawbacks of the test with a view to creating a public awareness towards their problematizing sides. In light of these targets, the study will focus on both the pre- and post-test periods of the respondents, together with their academic lives and experiences. Moreover, the study will elicit essential tactics and strategies for further test takers, compiled from the respondents' IBFLE experiences, and will also provide practical information and suggestions for policymakers.

1.4 Research Questions

The present study not only examines the washback effect in essence but also looks to find practical solutions for future takers and policymakers. These two aims are formulated in five questions, which pertain to both pre- and post-test periods:

Research Question 1: Is there a relationship between IBFLE success and factors such as gender, IBFLE languages, IBFLE modules, private tuition, private courses / classes, internet usage, overseas education, and undergraduate foreign language education?

Research Question 2: What language (sub)skills are emphasized in IBFLE?

Research Question 3: What is the effect of IBFLE test organization on the preparation of the students?

Research Question 4: According to test takers, what changes does IBFLE need?

Research Question 5: What competences are gained through IBFLE?

1.5 Definition of key terms and abbreviations

High-stakes test: the standardized test whose results are used as the sole determining factor for making a major decision (MSN Encarta Dictionary).

KPDS (SEFLE): the high-stakes foreign language test administered for state employees in Turkey. (KPDS: *Kamu Personeli Yabancı Dil Sınavı*; SEFLE: *State Employees Foreign Language Examination*)

Faculty Member: an instructor at Turkish universities who hold the academic titles of professor, associate professor, or assistant professor.

Standardized test: a test, administered according to standardized procedures, which assesses a student's aptitude by comparison with a standard (MSN Encarta Dictionary).

ÜDS (IBFLE): the high-stakes foreign language test in Turkey which is mainly administered for academicians and candidates of associate professor. (ÜDS: *Üniversitelerarası Kurul Yabancı Dil Sınavı*; IBFLE: *Interuniversity Board Foreign Language Examination*)

Washback (effect): the impact that tests have on teaching and learning (Shohamy, 1993: 4, cited in Bailey, 1999).

Foreign Language Test for the Candidates of associate professor (FLTAP): the high-stakes test once administered for the candidates of associate professorship in Turkey, which was annulled upon the inception of IBFLE.

1.6 Limitations of the Study

Depending on academic requirements, the target group in this study comprises of full time faculty members with the academic titles ranging between *assistant professor* to *professor*. Just like other professions, becoming a member of this group too entails a

benchmark score in language tests. For this benchmark, assistant professors are to get a minimum score of 65% in any of the three languages tested by IBFLE to achieve the linguistic criterion for becoming *associate professors*. Nonetheless, there is no language requirement for associate professors to promote. In this sense, the main group comprises of assistant professors, who are to pass IBFLE to widen their horizons through full professorship. Those who fail in IBFLE cannot attain the position they desire no matter how good their other requirements are. Both the ones who have passed and the ones who have failed are included in this study so as to assess a two-way appraisal of the test. In addition, other professors – associate professors and (full) professors – are also included in the study in order to assess their experienced post-test evaluations on the washback of IBFLE. Yet, the ones who had taken other tests like SEFLE or FLTAP are not studied. Consisting of an institution-wide research, the present study is to cover Dicle University, Diyarbakir, Turkey.

This chapter included the introductory remarks along with the purpose and significance of this study. A detailed account of relevant studies is presented in the following chapter, Review of Literature.

CHAPTER TWO

REVIEW OF LITERATURE

2.1 Presentation

This chapter reviews relevant literature while providing details about IBFLE along with its renowned counterparts around the world. Further, definitions of the concepts together with their types are explained in this chapter.

2.2 What is a Test?

The word 'test' is basically defined as "a way of discovering, by questions or practical activities, what someone knows, or what someone or something can do or is like" (Cambridge Online Dictionary) and as "a procedure intended to establish the quality, performance, or reliability of something, especially before it is taken into widespread use" (Oxford Online Dictionary). These definitions are further expounded in scientific bases, to the extent that "test" is represented as a fundamental element of language testing. Bachman (1995: 20), for instance, defines it as "a measurement instrument designed to elicit a specific sample of an individual's behaviour", drawing special attention to its power in eliciting behaviours as well. Indeed, in this definition, the scopes of a test are expanded to being able to assess not only theoretical knowledge but also physical behaviours. Additionally, Brown (2003: 3) broadens the term to "a method of measuring a person's ability, knowledge, or performance in a given domain", in which he emphasizes the importance of another notion, *i.e.* in a given domain, which confines the spatial and temporal effect of a test to a specific area, skill, or location. This implies that tests could vary depending on their style, language, purpose, coverage, content, mechanics, targeted skills, test takers, scoring, level of challenging, and so on. From these definitions, it follows that a test is perceived as a strategy which is employed to discover a specific pattern of knowledge, information, skill, or behaviour by using various tactics in a specific realm.

2.2.1 Role of Testing in Language Teaching

According to Hughes (1992: 4), testing is an indispensable element for language teaching:

> Information about people's language ability is often very useful and sometimes necessary. It is difficult to imagine, for example, British and American universities accepting students from overseas without some knowledge of their proficiency in English. The same is true for organizations hiring interpreters or translators. They certainly need dependable measures of language ability. Within teaching systems, too, as long as it is thought appropriate for individuals to be given a statement of what they have achieved in a second or foreign language, then tests of some kind or other will be needed.

Bachman (1995: 54) states that "The fundamental use of testing in an educational program is to provide information for making decisions, that is, for evaluation.". Testing, then, is portrayed as the roadmap for making decisions about the test takers and thus providing evaluations on their qualifications.

Brown (2003: 4) makes a clear distinction between the terms 'testing' and 'assessment', depicting them as one involving the other:

> Tests are prepared administrative procedures that occur at identifiable times in a curriculum when learners muster all their faculties to offer peak performance, knowing that their responses are being measured and evaluated. Assessment, on the other hand, is an on-going process that encompasses a much wider domain.

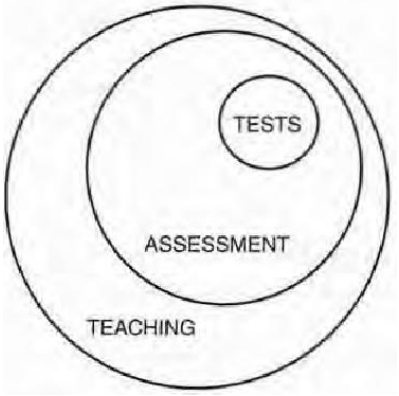

Figure 1. Tests, assessment, and teaching (Brown, 2003: 4)

Bachman and Palmer (1996: 54) argue that the primary concern in language testing is testing the language ability of the individuals, which is followed by "topical knowledge, or knowledge schemata, and affective schemata":

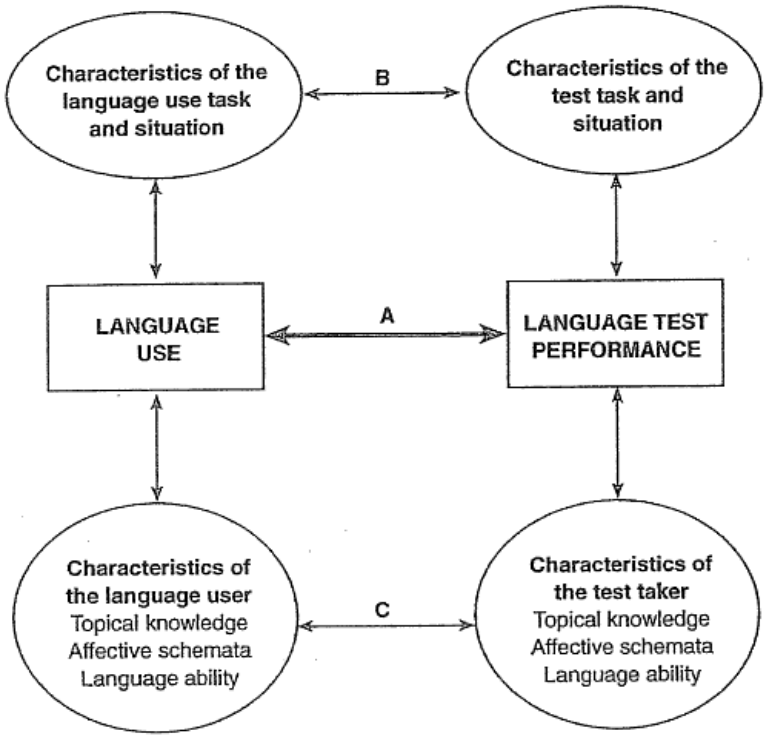

Figure 2. Interrelations between language use and language test performance (Palmer, 1996: 54)

2.2.2 Classification of Tests

There are two main categories for tests that classify them in terms of *content* and *style*.

2.2.2.1 Content-based Classification

According to Hughes (1992: 9), the classification of tests in terms of their contents depends on the information targeted to be obtained:

> We use tests to obtain information. The information that we hope to obtain will of course vary from situation to situation. It is possible, nevertheless, to categorise tests according to a smaller number of kinds of information being sought. This categorisation will prove useful both in deciding whether an existing test is

> suitable for a particular purpose and in writing appropriate new tests where these are necessary.

Depending on the content that is conveyed for assessment, the tests to be analysed in this study include *proficiency tests, achievement tests, diagnostic tests,* and *placement tests*.

2.2.2.1.1 Proficiency Tests

Brown (2003: 44) defines proficiency tests as in the following:

> Proficiency tests are designed to measure people's ability in a language regardless of any training they may have had in that language. The content of a proficiency test, therefore, is not based on the content or objectives of language courses which people taking the test may have followed. Rather, it is based on a specification of what candidates have to be able to do in the language in order to be considered proficient.

From this clarification, it follows that these tests are not limited to a single skill or course, but they are to assess overall ability. The same view is supported by Bachman (1995: 47):

> In developing a language proficiency test, the test developer does not have a specific syllabus and must rely on a theory of language proficiency for providing the theoretical definitions of the abilities to be measured. For example, if an admissions officer needs to determine the extent to which applicants from a wide variety of native language backgrounds and language learning experiences will be able to successfully pursue an academic program in a second language, she may decide to measure their language ability.

These tests, then, are not concerned with how or where the candidates have been educated; rather, they seek to assess the competence, *i.e.* proficiency, depending on a given theory.

2.2.2.1.2 Achievement Tests

Unlike the proficiency tests, which are theory-based, achievement tests are based on a syllabus which has been taught throughout an educational period (Bachman, 1995:

71). According to Hughes (1992: 47, 48), there are two types of achievement tests: *final* achievement tests and *progress* achievements tests:

> *Final* achievement tests are those administered at the end of a course study. They may be written and administered by ministries of education, official examining boards, or by members of teaching institutions. (…) *Progress* achievement tests, as their name suggests, are intended to measure the progress that students are making.

Further, there are a number of advantages provided by these tests (Bachman and Palmer, 1996: 334):

> First, they illustrate the use of classroom teaching and learning tasks as a basis for developing test tasks. (…) Second, they demonstrate the use of syllabus content, in the form of teaching and learning objectives or targets, as a basis for defining the constructs to be measured. Third, they illustrate ways in which our framework of task characteristics can be adapted to suit the needs of a particular situation. Finally, they illustrate the applicability of our approach for developing tests to languages other than English.

2.2.2.1.3 Diagnostic Tests

As opposed to proficiency and achievement tests, diagnostic tests are smaller in terms of coverage, which are purposely designed to diagnose a specific area of knowledge (Brown, 2003: 46). According to Hughes (1992: 13), these tests are used to identify students' strong and weak points:

> They are intended primarily to ascertain what further teaching is necessary. At the level of broad language skills this is reasonably 'straightforward. We can be fairly confident of our ability to create tests that will tell us that a student is particularly weak in, say, speaking as opposed to reading in a language.

Whether these tests rely on a theory or a syllabus is clarified by Bachman (1995: 60):

> When we speak of a diagnostic test, however, we are generally referring to a test that has been designed and developed specifically to provide detailed information about the specific content domains that are covered in a given program or that are

> part of a general theory of language proficiency. Thus, diagnostic tests may be either theory or syllabus-based.

2.2.2.1.4 Placement Tests

Considered as a kind of proficiency test (Brown, 2003: 45), placements tests are defined by Hughes (1992: 14) as follows:

> Placement tests, as their name suggests, are intended to provide information which will help to place students at the stage (or in the part) of the teaching programme most appropriate to their abilities. Typically they are used to assign students to classes at different levels.

For this type of testing, "the test developer may choose to base the test content either on a theory of language proficiency or on the learning objectives of the syllabus to be taken" (Bachman, 1995: 59).

According to Brown (2003:45), there are many varieties for placement tests:

> Placement tests come in many varieties: assessing comprehension, and production, responding through written and oral performance, open-ended and limited responses, selection (e.g., multiple choice) and gap-filling formats, depending on the nature of a program and its needs. Some programs simply use existing standardized proficiency tests because of their obvious advantage in practicality–cost, speed in scoring and efficient reporting of results.

2.2.2.2 Style-based Classification

In style-based categorization, the tests are regarded different depending on two dichotomic features: *direct versus indirect testing* and *norm-referenced versus criterion-referenced*.

2.2.2.2.1 Direct versus Indirect Testing

This dichotomy is concerned with the outcome of an assessment to be either direct or indirect, *i.e.* whether a skill or knowledge can be elicited directly or via another skill. According to Byram (2004: 178) and Hughes (1992: 14-16), the directly tested skills are the productive skills –speaking and writing, and the indirect ones are the receptive skills –

reading and listening. From this distinction, it follows that the assessability of the receptive skills inevitably relies on the conduction of productive skills. Adversely, the productive skills could also be assessed indirectly considering that a dialogue completion exercise would indirectly test speaking (Byram, 2004: 178).

Bachman (1995: 33) argues that this distinction is concerned with whether the outcomes of a test resemble 'real-life' language performance or not, and that "the direct tests are regarded as valid evidence of the presence or absence of the language ability in question".

Another distinction on this dichotomy is drawn by Hughes (1992: 15):

> Direct testing is easier to carry out when it is intended to measure the productive skills of speaking and writing. The very acts of speaking and writing provide us with information about the candidate's ability. With listening and reading, however, it is necessary to get candidates not only to listen or read but also to demonstrate that they have done this successfully. The tester has to devise methods of eliciting such evidence accurately and without the method interfering with the performance of the skills in which he or she is interested.

2.2.2.2.2 Norm-referenced versus Criterion-referenced Testing

In this distinction there are two types of tests which diverge from each other in terms of their evaluation on the test takers' scores. If a test places the test takers in percentage-based sequence or compares a score with the others, the test is called norm-referenced, where *norm* represents the test scores of other test takers that are relied on to enumerate each score. Conversely, a test is called criterion-referenced test when a test does not compare a test score with others but evaluates it depending on a pre-defined score, which is regarded as the criterion to be passed (Hughes, 1992: 17-19).

Bachman and Palmer (1996: 8) assign a crucial advantage for criterion-referenced testing:

> The primary advantage of criterion-referenced scales is that they allow us to make inferences about how much language ability a test taker has, and not merely how well she performs relative to other individuals, including native speakers. We define the lowest level of our scales in terms of no evidence of knowledge or

> ability and the highest level in terms of mastery, or complete knowledge or ability. We will thus always have zero and mastery levels in our scales, irrespective of whether there are any test takers at these levels.

Hughes (1992: 18) also advocates the usefulness of criterion-referenced testing, presenting two 'positive virtues':

> Criterion-referenced tests (…) have two positive virtues: they set standards meaningful in terms of what people can do, which do not change with different groups of candidates; and they motivate students to attain those standards.

2.2.3 Characteristics of a Test

The term 'test' needs a further detailed analysis into what characteristics it should hold to qualify as a proper tool. These characteristics are stated as *reliability, validity, authenticity, and washback* (Bachman and Palmer, 1996: 19-38).

2.2.3.1 Reliability

As the name suggests, a test should be reasonably reliable to qualify as a fair device. Bachman and Palmer (1996: 19) expand on this issue:

> Reliability is often defined as consistency of measurement. A reliable test score will be consistent across different characteristic of the testing situation. Thus, reliability can be considered to be a function of the consistency of scores from one set of test tasks to another.

The point made clear is that reliability is actually the consistency of a test, which would – and is expected to – provide similar, if not the same, scores under equal circumstances.

According to Genesee and Upshur (1996: 60) and Hughes (1992: 36-42), the reliability of a test is both dependent on the test itself and on the individuals taking part in the process of language testing:

Table 1. Types of reliability and ways of enhancing reliability (Genesee and Upshur, 1996: 60)

Type of reliability	*Ways of enhancing reliability*
Rater reliability	Use experienced, trained raters Use more than one rater Raters should carry out their assessments independently
Person-related reliability	Assess on several occasions Assess when person is prepared and best able to perform well Ensure that person understands what is expected (that is, that instructions are clear)
Instrument-related reliability	Use different methods of assessment Use optimal assessment conditions, free from extraneous distractions Keep assessment conditions constant

Similarly, Hughes (1992: 36-42) provided a set of suggestions for the sake of creating more reliable tests:

1. Take enough samples of behaviour.
2. Do not allow candidates too much freedom.
3. Write unambiguous items.
4. Provide clear and explicit instructions.
5. Ensure that tests are well laid out and perfectly legible.
6. Candidates should be familiar with format and testing techniques.
7. Provide uniform and non-distracting conditions of administration.
8. Use items that permit scoring which is as objective as possible.
9. Make comparisons between candidates as direct as possible.
10. Provide a detailed scoring key.
11. Train scorers.
12. Agree acceptable responses and appropriate scores at outset of scoring.
13. Identify candidates by number, not name.
14. Employ multiple, independent scoring.

2.2.3.2 Validity

Validity, in simple terms, seeks to find out whether a test is actually testing what it is intended to (Hughes, 1992: 22). Bachman (1995: 236) further explains validity as follows:

> In examining validity, we look beyond the reliability of the test scores themselves, and consider the relationships between test performance and other types of performance in other contexts. The types of performance and contexts we select for investigation will be determined by the uses or interpretations we wish to make of the test results.

Regarded as "the most important principle for an effective test" (Brown, 2003: 22), validity is explicated by Messick (1996: 7) through a set of questions to be used in deriving the validity of a test:

- Are we looking at the right things in the right balance?
- Has anything important been left out?
- Does our way of looking introduce sources of invalidity or irrelevant variance that bias the scores of judgments?
- Does our way of scoring reflect the manner in which domain processes combine to produce effects and is our score structure consistent with the structure of the domain about which inferences are to be drawn or predictions made?
- What evidence is there that our scores mean what we interpret them to mean, in particular, as reflections of personal attributes or competencies having plausible implications for educational action?
- Are there plausible rival interpretations of score meaning or alternative implications for action and, if so, by what evidence and arguments are they discounted?
- Are the judgments or scores reliable and are their properties and relationships generalizable across the contents and contexts of use as well as across pertinent population groups?
- Are the value implications of score interpretations empirically grounded, especially if pejorative in tone, and are they commensurate with the score's trait implications?

- Do the scores have utility for the proposed purposes in the applied settings?
- Are the scores applied fairly for those purposes, that is, consistently and equitably across individuals and groups?
- Are the short- and long-term consequences of score interpretation and use supportive of the general testing aims and are there any adverse side-effects?

As elaborated in these questions, for a test to be valid, it should cover everything it should assess without leaving any tiny fragment out. Unlike reliability, validity is rather concerned with securing the test material in terms of the correlation between test coverage and purpose. Bachman (1995: 240) further differentiates between these two principles as follows:

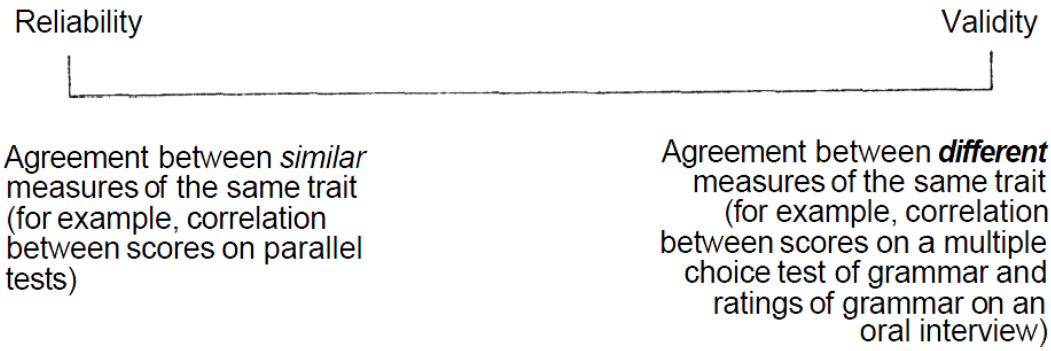

Figure 3. Relationship between reliability and validity (Bachman, 1995: 240)

Depending on analytical perspectives, validity consists of four types, including *content validity, criterion validity, construct validity, and face validity.*

2.2.3.2.1 Content Validity

Labelled as 'content-related evidence' by Brown (2003: 22), this type of validity is generally known as 'content validity' (Bachman, 1995: 244; Alderson, Clapham & Wall, 1995: 173; Hughes, 1992: 22; Carmines & Zeller, 1979: 20). According to Brown (2003: 22), a test can claim content-related evidence of validity if it "samples the subject matter about which conclusions are to be drawn, and if it requires the test-taker to perform the behaviour that is being measured". Kerlinger (1973: 458, cited in Alderson, Clapham & Wall, 1995: 173) defines it as the "representativeness or sampling adequacy of the content", by which content validity is shown as the representation of the test content itself. The same resemblance is presented by Hughes (1992: 22):

> A test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned. It is obvious that a grammar test, for instance, must be made up of items testing knowledge or control of grammar. But this in itself does not ensure content validity. The test would have content validity only if it included a proper sample of the relevant structures. Just what are the relevant structures will depend, of course, upon the purpose of the test.

Bachman (1995: 244-247) evaluates content validity in two different aspects: *content relevance* and *content coverage*. In content relevance, the relation between the examiner and examinees is questioned, *i.e.* the similarities between the two respondents with regards to their ages, genders, statuses and ethnic origins –merged with the notion of 'the specification of the ability domain', which, according to Bachman (1995: 244), is often ignored. In other words, content relevance questions the parallelism between the examiner and examinees while looking for an accurate identification of the behavioural domains in a test. Content coverage, on the other hand, is quite similar to the term "content validity" in that the basic aim here is to find out "the extent to which the tasks required in the test adequately represent the behavioural domain in question" (p. 244).

2.2.3.2.2 Criterion Validity

Criterion validity concerns the parallelism between test scores and the independent standards which are regarded as the criteria for the skills or knowledge tested (Bachman, 1995: 248; Brown, 2003: 24). In this respect, the analysis on this validity is in two ways: *concurrent validity* and *predictive validity* (Alderson, Clapham & Wall, 1995: 177; Hughes, 1992: 23; Litwin, 1995: 37; Carmines & Zeller, 1979: 17).

2.2.3.2.2.1 Concurrent Validity

Litwin (1995: 37) expands on concurrent validity as follows:

> Concurrent validity requires that the survey instrument in question be judged against some other method that is acknowledged as a "gold standard" for assessing the same variable. It may be a published psychometric index, a scientific measurement of some factor, or another generally accepted test. The

> fundamental requirement is that it be regarded as a good way to measure the same concept.

Further, Bachman (1995: 248) analyses this validity in two forms: "(1) examining differences in test performance among groups of individuals at different levels of language ability, or (2) examining correlations among various measures of a given ability", creating a clear-cut dichotomy for the usability of concurrent validity both for individuals with different linguistic competences and for the theoretical correlations among the measures given for a test. Moreover, Carmines & Zeller (1979: 18) regards this type as the correlation of a measure with the criterion that is taking place at the same time of point, *e.g.* a verbal report of behaviour that is officially revealed depending on the participation in the election.

2.2.3.2.2.2 Predictive Validity

Unlike concurrent validity, predictive validity is concerned with the power of a test in predicting the candidates' performances (Hughes, 1992: 25; Litwin, 1995: 40; Carmines & Zeller, 1979: 18). According to Alderson, Clapham & Wall (1995: 188) proficiency tests are the most common examples for this type:

> Predictive validation is most common with proficiency test: tests which are intended to predict how well somebody will perform in the future. The simplest form of predictive validation is to give students a test, and then at some appropriate point in the future give them another test of the ability the initial test was intended to predict. A common use for a proficiency test like IELTS or the TOEFL is to identify students who might be at risk when studying in an English-medium setting because of weaknesses in their English.

Bachman (1995: 250) argues that the sole use of predictive validity can basically ignore the analysis of the abilities being measured. In other words, an analysis that would solely focus on the predictive validity of a test would inevitably overlook the concurrent – and also the content – validity of the test. As an example, the scores for a mathematics exam could be the predictors of performance in language courses, but the irrelevance of these scores with examinees could be ignored unless an analysis containing either concurrent or content analysis is conducted.

2.2.3.2.3 Construct Validity

Prior to construct validity, there is a need for defining the subject matter – *construct*. Brown (2003: 25) defines it as "any theory, hypothesis, or model that attempts to explain observed phenomena in our universe of perceptions" while Hughes (1992: 26) regard it as "any underlying ability (or trait) which is hypothesized in a theory of language ability". Further, Brown (2000b: 9) makes a psychological definition, quoting: "A construct, or psychological construct as it is also called, is an attribute, proficiency, ability, or skill that happens in the human brain and is defined by established theories". From these definitions, the word "construct" remains an entity – either in the form of a theory or psychological concept – which constitutes the backbone of the structure embedded in a test. Accordingly, construct validity, which is considered as the most difficult type to explain (Alderson, Clapham & Wall, 1995: 183) and as the one with the least role for classroom teachers (Brown, 2003: 25), concerns the measurement of these 'constructs', *i.e.* it regards the extent to which test performance reflects the constructs intended for a test (Bachman, 1995: 255).

Messick (1996: 9-10) defined a number of aspects that are 'a means of addressing central issues implicit in the notion of validity as a unified concept':

- The content aspect of construct validity includes evidence of content relevance and representativeness as well as of technical quality (*e.g.* appropriate reading, level, unambiguous phrasing, and correct keying).
- The substantive aspect refers to theoretical rationales for the observed performance regularities and item correlations, including process models of task performance, along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks.
- The structural aspect appraises the fidelity of the score scales to the structure of the construct domain at issue with respect to both number (*i.e.* appropriate dimensionality) and makeup (e.g. conjunctive vs. disjunctive, trait vs. class).
- The generalizability aspect examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks, including generalizability of test-criterion relationships across settings and time periods, which is known as validity generalization.

- The external aspect includes convergent and discriminant evidence from multitrait-multimethod comparisons, as well as evidence of criterion relevance and applied utility.
- The consequential aspect appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice, as well as to washback.

Genesee and Upshur (1996: 69) compiled a comparative table, adding a detailed evaluation of the three types – content, criterion, and construct validity – along with examples for each of them:

Table 2. Comparison of three approaches to validity by Genesee and Upshur (1996: 69)

	Question	*Method*	*Common use*	*Example*
Content relevance	How well does the content of the instrument sample the kinds of things about which conclusions are to be drawn?	Logically conclude whether the content of the instrument comprises an adequate definition of what it claims to assess.	Achievement tests.	A course examination analyzed to see if the contents relate to the course objectives.
Criterion-relatedness	How does the information from this instrument compare with information from more direct measures of the attribute being assessed?	Administer the instrument and compare results with information from direct measures obtained concurrently or in the future.	Instrument used to select and classify people.	Results of a placement test are compared with performance of students in class.
Construct validity	To what extent do certain explanatory concepts account for the information provided by this instrument?	Set up hypotheses regarding people with much or little of the attribute based upon underlying theory. Test the hypotheses.	Instruments used for description or in scientific research.	A test of learning styles is studied to see how well types account for learning by different teaching methods.

2.2.3.2.4 Face Validity

Unlike the other types, face validity concerns the holistic credibility of a test as perceived by non-experts like lay people, student, or test takers (Bachman, 1995: 285; Brown, 2003: 27; Hughes, 1992: 27). Alderson, Clapham & Wall (1995: 172) argue that this type of validity covers all the other types through an informal perspective which

gives rise to the comment: ‘This test does not *look* valid’. According to Brown (2003: 27), face validity can be raised if the learners are provided with:

- a well-constructed, expected format with familiar tasks,
- a test that is clearly doable within the allotted time limit,
- items that are clear and uncomplicated,
- directions that are crystal clear,
- tasks that relate to their course work (content validity), and
- a difficulty level that presents a reasonable challenge.

2.2.3.3 Authenticity

Authenticity is defined as ‘the quality of being true or real’ (MSN Encarta Dictionary, Oxford Online Dictionary, Cambridge Online Dictionary). However, it is something rather sophisticated, described as the degree of parallelism between the characteristics of the language assessed in a test and the actual language used in the contexts of the target language (Bachman and Cohen, 1998: 23; Bachman and Palmer, 1996: 23).

Brown (2003: 28) defined a set of indicators, signifying the presence of authenticity in a test:

- The language in the test is as natural as possible.
- Items are contextualized rather than isolated.
- Topics are meaningful (relevant, interesting) for the learner.
- Some thematic organization to items is provided, such as through a story line or episode.
- Tasks represent, or closely approximate, real-world tasks.

Bachman (1995: 10) points to a crucial problem with the conduction of authenticity analysis, arguing that the problem underlies with the difficulty of discriminating ‘real-life’ situations from ‘non-real-life’ situations. Since this discrimination would most probably rely on subjective attitudes –rather than scientific

standards– it would be difficult to provide a precise consequence regarding the authenticity of a test. On the other hand, Brown (2003: 28) advocates the provability of validity, claiming that the authenticity of the tests that have been produced in recent years has remarkably risen in that the examiners have growingly been using real-life, or 'episodic' reading texts, and have been enriching their test materials with more speaking and listening materials.

2.2.3.4 Washback

Washback or *backwash* is simply defined as "unpleasant consequences of an event" in dictionaries (MSN Encarta Dictionary, Oxford Online Dictionary, Cambridge Online Dictionary), which is not sufficient for the scientific realm to be handled in this study. *Washback* or *backwash*, in this context, is characterized as "the influence of testing on teaching and learning" (Alderson & Wall, 1992: 2; Brown, 1995: 92; 2002: 11), and as "the extent to which a test influences teachers and learners to do things they would not otherwise necessarily do" (Messick, 1996: 1). In light of these definitions, washback can be hailed as an academic impression on the teachers, learners and materials of a test, in a way that it leads them towards necessary changes and measures that might prove either useful or harmful. Even so, it is not confined to the individuals as such. It is interpreted as something more prevalent, as likened to the notion of *impact* by Taylor (2005: 154), who employed it to denote the *public impact* of washback, which plays a greater part in beyond-classroom processes like employment opportunities, career options, and even some wider notions like "school curriculum planning, immigration policy, or professional registration for doctors". Depending on these qualities, washback subsists as an external factor which dominates not only the classroom tests administered at mainstream education but also, and even more ubiquitously, the public exams which play a more significant role both at individual and public levels (Brown, 2002: 11).

Alderson & Wall (1992: 9) formulated 15 hypotheses towards the clarification of washback, creating an insight into how and what washback will affect:

1. A test will influence teaching.
2. A test will influence learning.
3. A test will influence how teachers teach and
4. A test will influence what teachers teach.

5. A test will influence what learners learn and
6. A test will influence how learners learn.
7. A test will influence the rate and sequence of learning; and
8. A test will influence the rate and sequence of teaching and the associated:
9. A test will influence the degree and depth of learning and
10. A test will influence the degree and depth of teaching.
11. A test will influence attitudes to the content, method, etc. of learning/teaching.
12. Tests that have important consequences will have washback, and conversely
13. Tests that do not have important consequences will have no washback.
14. Tests will have washback on all learners and teachers.
15. Tests will have washback effects for some learners and some teachers, but not for others.

Cited in various studies (Bailey, 1999: 13-14; Djurić, 2008: 17; Reynolds, 2010: 10), these hypotheses point towards two significant assumptions: the first and the most obvious one denotes the centrality of two actors –teacher and learner. Washback is a multi-way effect revolving around these two respondents and characterized accordingly. Secondly, in reference to hypotheses #12 and #13 in particular, washback is expected to exist only with the tests that have important consequences, *i.e.* high-stakes tests, but not with the ones that are not. In other words, washback is applicable only to high-stakes tests like college admission tests or career-gaining tests, but not the ordinary school exams which are relatively less important.

2.2.3.4.1 Positive and Negative Washback

The complexity of washback is deepened through its crucial aspect: *In what ways does washback affect the elements?* Owing to some explanations which are stated for the sake of clarifying the notion of washback, such as "washback is often introduced on language testing courses as a powerful concept that all test designers need to pay attention to, and which most classroom teachers are all too aware of" (Alderson & Wall, 1992: 3) and "tests can affect actual language policy, at educational or societal levels" (Shohamy, 2007: 128), washback is misunderstood as something *all-negative*, which flatly denies the existence of a positive aspect. Moreover, the misconception is emboldened through the diehard connotation between the term *washback* and the words

influence and *impact*. The essential truth is that washback can be both positive and negative depending on the consequence it has brought up (Brown, 2002: 12). Taylor (2005: 154) makes a clear distinction between the two types:

> Negative washback is said to occur when a test's content or format is based on a narrow definition of language ability, and so constrains the teaching/learning context. (…) Positive washback is said to result when a testing procedure encourages 'good' teaching practice; for example, an oral proficiency test is introduced in the expectation that it will promote the teaching of speaking skills.

As simplified by Brown (1995: 92), positive washback exists when a test properly corresponds with the curricular goals, while the negative one prevails when a deviation occurs between these two ends. Clearly then, washback is either a benefit or a danger for language teaching policy in that it can change it either in a positive or negative way. As expected, these directions could lead to different outcomes, as shown in a comparative-style diagram by Brown (2002:11):

Figure 4. Negative and positive washback by Brown (2002: 11)

In addition to these clarifications, both positive and negative washback are explained in different ways depending on their proven outcomes. In their study, Cheng and Curtis (2009: 36) examined a Chinese test, called *The Graduate School Entrance Examination* (GSEE), finding that the test had a positive washback on test takers in that it led them to do more reading and gained them a strong desire to improve, but on the other hand it had a strong negative washback which made them start studying 1 or 2 years earlier and thus prevented them from focussing on their undergraduate studies.

Some washback studies have been devoted to provide strategies for the promotion of positive washback. In this sense, Brown (2000a: 4) outlined a list of useful strategies:

A. Test design strategies

1. Sample widely and unpredictably.
2. Design tests to be criterion-referenced.
3. Design the test to measure what the programs intend to teach.
4. Base the test on sound theoretical principles.
5. Base achievement tests on objectives.
6. Use direct testing.
7. Foster learner autonomy and self-assessment.

B. Test content strategies

1. Test the abilities whose development you want to encourage.
2. Use more open-ended items (as opposed to selected-response items like multiple choice).
3. Make examinations reflect the full curriculum, not merely a limited aspect of it.
4. Assess higher-order cognitive skills to ensure they are taught.
5. Use a variety of examination formats, including written, oral, aural, and practical.
6. Do not limit skills to be tested to academic areas (they should also relate to out-of-school tasks).
7. Use authentic tasks and texts.

C. Logistical strategies

1. Insure that test-takers, teachers, administrators, curriculum designers understand the purpose of the test.
2. Make sure language learning goals are clear.
3. Where necessary, provide assistance to teachers to help them understand the tests.
4. Provide feedback to teachers and others so that meaningful change can be effected.
5. Provide detailed and timely feedback to schools on levels of pupils' performance and areas of difficulty in public examinations.
6. Make sure teachers and administrators are involved in different phases of the testing process because they are the people who will have to make changes.
7. Provide detailed score reporting.

D. Interpretation strategies

1. Make sure exam results are believable, credible, and fair to test takers and score users.
2. Consider factors other than teaching effort in evaluating published examination results and national rankings.
3. Conduct predictive validity studies of public examinations.
4. Improve the professional competence of examination authorities, especially in test design.
5. Insure that each examination board has a research capacity.
6. Have testing authorities work closely with curriculum organizations and with educational administrators.
7. Develop regional professional networks to initiate exchange programs and to share common interests and concerns.

What makes the list remarkable is its all-covering style, where every essential phase has been rendered for a proper test implementation. Further, it achieves in providing the strategies both in a chronological and logical order, enabling the four categories to be handled in three different groups; the first and the second as *pre-testing*, the third as *in-testing*, and the final as *post-testing strategies*.

2.2.3.4.2 Washback Models

Considered as a framework, washback covers a set of elements which are connected to each other and also connected to the notion that exists at the top of the structure –washback. There are several frameworks drawn to depict this rotating relationship. The model by Bailey (1999: 11) poses a good example in illustrating the multi-way connections among the elements of washback. It also achieves depicting the interwoven relationship among respondents and products, which both arise from and return to the ultimate point, “TEST”:

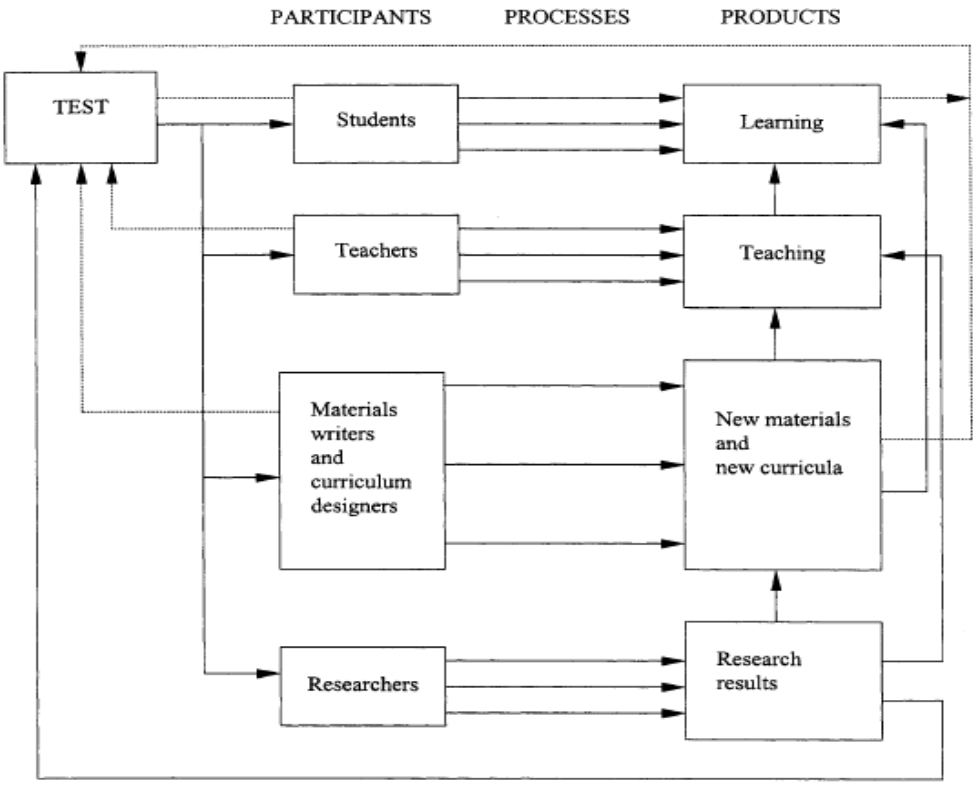

Figure 5. A model of washback by Bailey (1999: 11)

Inspired by the ideas of Hughes (1993) and Bachman and Palmer (1996), Pan (2008: 12) portrayed a holistic model of washback based on a two-axis framework, making a clear distinction between the micro and macro levels. In so doing, both individual and public effects are well-depicted. To assess these levels, according to Pan (2008: 12), "a triangulation of questionnaires, interviews, observations, pre- and-post tests, and document analysis need to be conducted". Pan (2008: 12) also emphasises the importance of the respondents in this complicated mechanism; "This process involves many different stakeholders such as teachers, students, administrators, policymakers, family members and the general public", paying exclusive attention to the role of family members, unlike other models (Fig. 6):

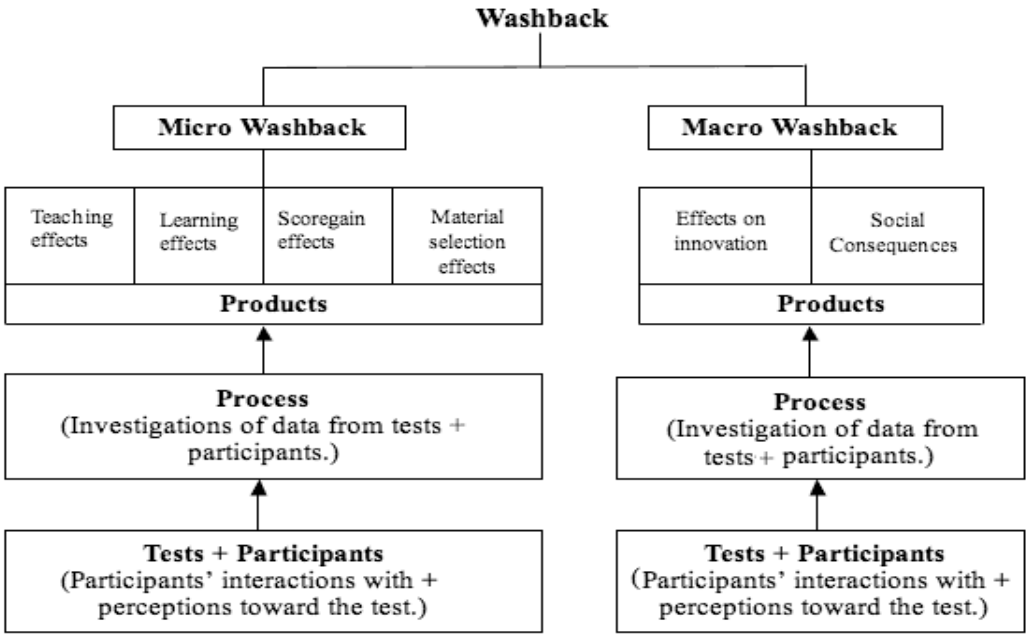

Figure 6. A holistic model of washback by Pan (2008: 12)

A similar yet simplistic model has been suggested by Saif (2006: 5), in which two exceptional elements, *background knowledge* and *motivation*, are shown as crucial elements. Unlike Bailey (1999) and Pan (2008), Saif (2006) handles the whole process in three-phases; *needs*, *means*, and *consequences*, respectively:

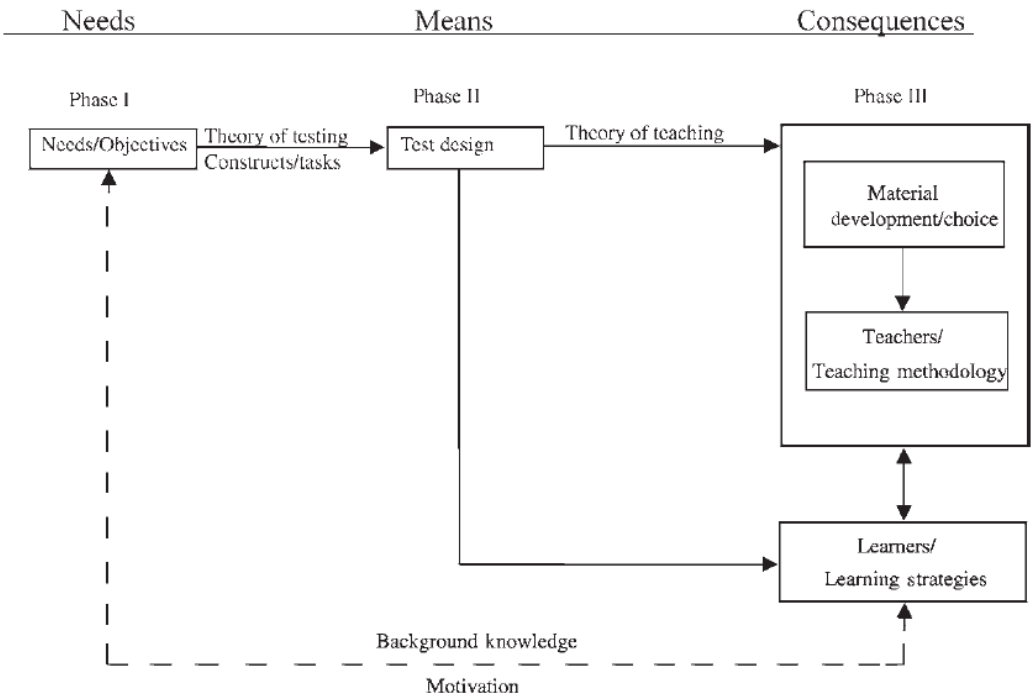

Figure 7. A conceptual framework for washback by Saif (2006: 5)

2.3 High-stakes Tests

High-stakes tests, as the name suggests, refer to the tests whose results are used to make important decisions (Heubert and Hauser, 1999). They could be college admission tests, school exit exams, or career-gaining tests, all of which point to individual benefits. On the other hand, considered at institution/school level, they could be used in certifying schools depending on general test results of students in a school (Lewis, 2000: 1-2).

2.3.1 High-stakes Tests in Turkey

There are numerous high-stakes tests in Turkey but only two of them, IBFLE and SEFLE, are directly concerned with the context of the present study in that they serve as career-gaining tests through the assessment of foreign language proficiency.

2.3.1.1 IBFLE (ÜDS)

IBFLE is a well-known high-stakes test which is required for academic positions and, less commonly, in private sector recruitments. Implemented twice a year –every March and October, IBFLE diverges from other testing systems in that it is not only implemented in three languages, *i.e.* English, French, German, but also served in three different modules –Social Sciences, Natural & Applied Sciences, and Health Sciences, which enables each test taker to be tested in his/her own field of study. All three modules are alike in the number of questions and are served to every test taker in the same booklet, but the answer keys for each module differ.

In terms of test classification, IBFLE qualifies for a criterion-referenced proficiency test in that it imposes the passing criterion as 65%. Further, the test is an all-indirect test in that all the language skills but listening, along with two subskills – grammar and vocabulary – are assessed indirectly.

Test takers are given 180 minutes (3 hours) to answer a total of 80 multiple-choice questions which are divided into 9 separate sections. The sections differ in number of questions, question types and required skill(s) or subskill(s):

Table 3. IBFLE sections before 2011 (adapted from the October-2008 issue)

Question Type Explained	Required Skill(s) or Subskill(s)	f	%
1. Choosing a word or phrase that best completes each sentence	Vocabulary Grammar Reading	18	22.5
2. Completing the five missing parts for each of two paragraphs with suitable words or phrases	Vocabulary Grammar Reading Writing	5	6.25
3. Finding the missing half of each sentence	Grammar Vocabulary Reading	12	15
4. Translation (English > Turkish)	Vocabulary Grammar Reading	3	3.75
5. Translation (Turkish > English)	Vocabulary Grammar Reading	3	3.75
6. Answering 4 questions for each of the 6 paragraphs	Reading Vocabulary Grammar	24	30
7. Finding the missing response for each conversation	Reading Vocabulary Grammar Speaking	5	6.25
8. Finding the missing sentence for each paragraph	Reading Vocabulary Grammar	5	6.25
9. Finding the odd sentence in each paragraph	Grammar Vocabulary Reading Writing	5	6.25
	TOTAL	**80**	**100**

An amendment was enacted in 2011, leading to a partial change in IBFLE, which rearranged the number of questions along with the order of some sections:

Table 4. IBFLE sections as of 2011 (adapted from the March-2011 issue)

Question Type Explained	Required Skill(s) or Subskill(s)	f	%
1. Choosing a word or phrase that best completes each sentence	Vocabulary Grammar Reading	17*	21.25
2. Completing the five missing parts for each of two paragraphs with suitable words or phrases	Vocabulary Grammar Reading Writing	10*	12.5
3. Finding the missing half of each sentence	Grammar Vocabulary Reading	10*	12.5
4. Translation (English > Turkish)	Vocabulary Grammar Reading	1*	1.25
5. Translation (Turkish > English)	Vocabulary Grammar Reading	3	3.75
6. Answering 4 questions for each of the 6 paragraphs	Reading Vocabulary Grammar	24	30
7. Finding the missing response for each conversation	Reading Vocabulary Grammar Speaking	5	6.25
8. Finding the missing sentence for each paragraph	Reading Vocabulary Grammar	5	6.25
9. Finding the odd sentence in each paragraph	Grammar Vocabulary Reading Writing	5	6.25
	TOTAL	**80**	**100**

* the sections changed

Originally implemented for *would-be* associate professors alone, it is now an indispensable criterion for any academic position at universities, but the (full) professorship. The candidates of associate professor are required to have a minimum score of 65% while for other academic positions a minimum of 50% is essential. As for private sectors, the minimum score depends on the qualification and type of the position available. Remarkably, IBFLE is growing demandable both for state and private positions, causing a great amount of stress and competitiveness due to its highly eliminative style.

2.3.1.2 SEFLE (KPDS)

SEFLE – short for *State Employees Foreign Language Examination* – basically aims to assess the state employees' foreign language proficiency, marking its crucial breakpoint with IBFLE in terms of implemental goal. Depending on the score achieved, state employees have a certain amount of extra money added to their salary. For this reason, the languages offered in SEFLE depend on the applicants' personal choices, *i.e.* they can ask to be assessed in any language that is available for public assessment, which makes SEFLE richer than IBFLE in terms of the variety of languages served. Formerly serving for this purpose alone, SEFLE commenced to be accepted as an academic criterion, just like IBFLE, upon a statute enacted in 2009, which obscured the monopoly of IBFLE in assessing the language proficiency. Following this change, some test takers tried both tests to raise their chance of achievement, and the others stuck with IBFLE since it imposed 20 questions less than SEFLE in the same amount of time. On the other hand, SEFLE is quite similar to IBFLE in terms of its question types, testing style, implemental frequency, test duration, test sections, topics, number of questions, and also level of challenging, but there are three points of divergence: (I) SEFLE takes place in every May and November while IBFLE takes place in March and October, (II) SEFLE has only one version; however, IBFLE offers three different modules, and (III) the variety of languages offered in SEFLE are more than 20 while IBFLE has only three languages. As for the purpose, SEFLE is mainly taken to prove linguistic proficiency in order to gain some extra income in the salary.

Before 2011, SEFLE covered 20 more questions, which reflected in the raised numbers of questions for some sections and addition of two distinct sections (#7 and #9):

Table 5. SEFLE sections before 2011 (adapted from the May-2009 issue)

Question Type Explained	Required Skill(s) or Subskill(s)	f
1. Choosing a word or phrase that best completes each sentence	Vocabulary Grammar Reading	15
2. Completing the five missing parts for each of two paragraphs with suitable words or phrases	Vocabulary Grammar Reading Writing	10
3. Finding the missing half of each sentence	Grammar Vocabulary Reading	10
4. Translation (English > Turkish)	Vocabulary Grammar Reading	5
5. Translation (Turkish > English)	Vocabulary Grammar Reading	5
6. Finding the missing sentence for each paragraph	Reading Vocabulary Grammar	6
7. Finding the best response according to the context given	Reading Vocabulary Speaking	6
8. Finding the odd sentence in each paragraph	Grammar Vocabulary Reading Writing	6
9. Finding the best paraphrase for the sentence given	Reading Vocabulary Grammar	6
10. Finding the missing response for each conversation	Reading Vocabulary Grammar Speaking	6

11. Answering 5 questions for each of the 5 paragraphs	Reading Vocabulary Grammar	25
	TOTAL	**100**

Just like IBFLE, SEFLE also underwent a radical amendment in 2011, which reduced the number of questions to 80 and altered most sections both in number of questions and order:

Table 6. SEFLE sections as of 2011 (adapted from the May-2011 issue)

Question Type Explained	**Required Skill(s) or Subskill(s)**	**f**	**%**
1. Choosing a word or phrase that best completes each sentence	Vocabulary Grammar Reading	14*	17.5
2. Completing the five missing parts for each of two paragraphs with suitable words or phrases	Vocabulary Grammar Reading Writing	10	8.0
3. Finding the missing half of each sentence	Grammar Vocabulary Reading	10	8.0
4. Answering 4 questions for each of the 5 paragraphs	Vocabulary Grammar Reading	20*	25.0
5. Finding the missing response for each conversation	Vocabulary Grammar Reading	4*	5.0
6. Finding the best paraphrase for the sentence given	Reading Vocabulary Grammar	4*	5.0
7. Finding the best response according to the context given	Reading Vocabulary Speaking	4*	5.0

8. Finding the missing sentence for each paragraph	Grammar Vocabulary Reading Writing	4*	5.0
9. Translation (English > Turkish)	Reading[1] Vocabulary[2] Grammar[2]	3*	3.75
10. Translation (Turkish > English)	Reading[1] Vocabulary[1] Grammar[1] Speaking[2]	3*	3.75
11. Finding the odd sentence in each paragraph	Reading[1] Vocabulary[2] Grammar[2]	4*	5.0
	TOTAL	**80**	**100**

* the sections changed

2.3.2 High-stakes Tests around the World

Around the world, there are innumerable IBFLE-like high-stakes tests, considering that almost every country in the world would have at least one language proficiency test to be used for placements or admissions. Among these, there are two world-famous tests, TOEFL and IELTS, which not only serve for their own countries but also for the entire world.

2.3.2.1 TOEFL

TOEFL – short for *the Test of English as a Foreign Language* – is the most widely respected English-language test in the world, recognised by more than 8,000 colleges, universities and agencies in more than 130 countries (http://www.ets.org/toefl). Initiated in 1963, the TOEFL is administered by Educational Testing Service (ETS), based in the United States. The 'Paper Based TOEFL' (PBT) was replaced by the 'Computer Based Test' (CBT), which was altered to the new 'Internet Based TOEFL' (IBT) in 2005-2006 with the addition of Speaking section and raised subsections for each section. The PBT is a 2.5 hour test of students' reading, listening and writing skills for studying or living in English-speaking countries. It consists of four sections –listening comprehension, structure and written expression, reading comprehension and writing.

The IBT is a 4.5-hour test of English students' reading, listening, speaking and writing skills for university study. It consists of four sections –reading, listening, integrated speaking and integrated writing / listening / reading (Reynolds, 2010: 3; www.ets.org/toefl).

All three versions also differ in terms of score scales. Total scores for PBT, CBT, and IBT are 677, 300, and 120, respectively. Accordingly, each section has equal weight in scoring, all of which are separately evaluated and then summed to form the resultant score:

TOEFL Score Scales

Internet-based Test	
Listening	0 to 30
Reading	0 to 30
Speaking	0 to 30
Writing	0 to 30
Total Score	**0 to 120**

Computer-based Test	
Listening	0 to 30
Structure/Writing	0 to 30
Reading	0 to 30
Total Score	**0 to 300**

Paper-based Test	
Listening Comprehension	31 to 68
Structure/Written Expression	31 to 68
Reading Comprehension	31 to 67
Total Score	**310 to 677**

Figure 8. TOEFL score scales (www.ets.org/toefl)

2.3.2.2 IELTS

IELTS – short for *the International English Language Testing System* – is used as an indicator of a candidate's ability to communicate in English, recognised by more than 6,000 institutions in over 135 countries. IELTS was developed by the British Council in partnership with IELTS Australia and Cambridge ESOL in 1989 (http://www.ielts.org/test_takers_information/what_is_ielts.aspx).

There are two versions of the IELTS –the Academic Version and the General Training Version. The Academic Version is taken by those who want to enrol in universities and other institutions of higher education and for professionals such as medical doctors and engineers who want to study or practice in an English-speaking country. The General Training Version is intended for those planning to undertake non-academic training or to gain work experience, or for immigration purposes.

Diverging in reading and writing sections, the two versions share the same style in listening and speaking sections:

Figure 9. Implemental differences between the two versions of IELTS (www.ielts.org/test_takers_information/what_is_ielts/test_format.aspx)

In both versions, scoring is graded on a band scale which classifies the test scores in nine different levels (Fig. 10):

The IELTS Band Score Scale	
9	Expert user
8	Very good user
7	Good user
6	Competent user
5	Modest user
4	Limited user
3	Extremely limited user
2	Intermittent user
1	Non user
0	Did not attempt the test

Figure 10. IELTS score scale
(www.ielts.org/test_takers_information/what_is_ielts.aspx)

2.4 Review of Relevant Literature

Studies on the effect of high-stakes exams has brought up many ideas that could be employed for a better understanding of washback effect, enabling test takers and also test developers to improve their techniques related to the exams. Arslan (2005) conducted research on the elements of IBFLE, paying special attention to the question types, their implications for test takers, and the necessity of learning English. Regarding the mechanics of the exam, each part of the exam was scrutinized individually –in an attempt to figure out the level of challenge for each part per se. The conclusions suggest that (1) most of the test takers are not aware of test techniques, (2) they are aware of their linguistic competence although they do not know how to improve it, and (3) that their competence is not in parallel with the style required in IBFLE.

It is possible to suggest that the effect of a single exam could mean different things to different test takers, as shown by two distinct studies on *English Component of the Foreign Language University Entrance Exam (ECFLUEE)*, by Sevimli (2007), in which High School students at 10^{th} and 11^{th} grade were examined to find out the washback effect not only on the students but also on the context of their learning and teaching, and by Yildirim (2010), in which the focus was on the effect of the exam on prospective English teachers, *i.e.* the students majoring in ELT. The results by Sevimli

(2007) showed that there is a negative washback on the teaching and learning processes and that the system used puts excessive emphasis on test-oriented syllabus, discarding the use of communicative skills. However, Yildirim (2010) revealed that the exam has some negative effects on students' linguistic competence and on their performance in their first year education at university.

A washback study does not necessarily have to cover an exam as a whole; it can simply focus on a single section of an exam, as studied on the Writing section of *College English Test Band Four* (CET-4) on ELT in Chinese colleges and universities by Chu and Gao (2006). The study, explicitly aiming to foster the positive effects and to minimize the negative ones, questioned the existence of the washback effect in essence. The study revealed that the writing subtest has both positive and negative effects and that the students do not pay attention to this subtest due to its low weight compared with the other sections.

Unlike the studies that question the existence of washback, the study conducted by Saif (2006) sought to find possible ways of creating positive washback by analysing the factors in the background of the test development process and reckoning the conditions which would best lead to positive washback. The study concludes with the notion that the test as such is not sufficient to change the education system in that there is "an intricate web of different yet related factors that could enhance or interfere with a test's effects being realized as educational change" (Saif, 2006: 30).

The washback on TOEFL was studied by Reynolds (2010), where she studied the washback through the student perspectives in three TOEFL preparation classrooms in the U.S.A. As the implications of the study, the results came up with two suggestions: (I) TOEFL teachers should avoid excessive explanations, granting more time for their students to communicate, and (II) TOEFL preparation courses should be more intense so as to involve as few students of similar English competencies as possible.

Inspired by Alderson & Wall (1992), Lewkowicz and Zawadowska-Kittel (2008) conducted a research on the impact of the new school-leaving exam of English in Poland, analysing teachers' views on the exam. The exam, called Nowa Matura, was introduced in Poland in 2005. According to the results, the respondents indicated that they focus on

teaching communicative skills and tend to stress the need for fluency thanks to the exam. It was also revealed that the task types introduced in the exam constitute the primary content the teachers use in their lessons and that the teachers teach test-taking strategies to help students perform well on the exam. On the other hand, the teachers criticized the test, claiming that the test is too easy in that it is basic level and it merely imposes a 30% pass mark.

In this chapter, the test in question, IBFLE, was introduced and further detailed through the presentation of two world-famous tests –TOEFL and IELTS. Also, relevant studies devoted to explore the washback effect in different tests were expounded. The methodology employed in the conduction of the study is presented in the following section.

CHAPTER THREE

METHODOLOGY

3.1 Presentation

The methods employed by the researcher are presented in this chapter. In addition to research design, data on the respondents and instruments of the study are presented here.

3.2 Research Design

The present study focuses on the washback of a well-known high-stakes test in Turkey, IBFLE, in an attempt to find out the intricacies and challenges behind the implemental scheme of the test. As the target group to be studied, full-time faculty members –professors, associate professors, and assistant professors, working at Dicle University, were selected. Of these, only the ones who took IBFLE to gain academic achievement were studied. The respondents were excluded if they had taken other tests like SEFLE or FLTAP. In deciding on the faculty members, the study aimed to find out their long-standing attempts in passing the IBFLE to attain higher titles. This being the case, the study elicited both quantitative and qualitative data to obtain a comprehensive picture of the problem that underlies with the essence of the challenges of the test in question. The quantitative data was gathered through the respondents' responses to multiple-choice and Likert-scale items while the qualitative data were obtained from the responses to the five open-ended questions in the interviews.

The research questions were arranged up to the items in the questionnaire and the interview:

Table 7. Correspondence of research questions with the items in the questionnaire and the interview

Research Question	Questionnaire Items	Interview Item(s)
Question 1: Is there a relationship between IBFLE success and factors such as gender, IBFLE languages, IBFLE modules, private tuition, private courses / classes, internet usage, overseas education, and undergraduate foreign language education?	1-5 & Open-ended Items	
Question 2: What language (sub)skills are emphasized in IBFLE? **a.** Differences by IBFLE languages **b.** Differences by IBFLE modules	6-10	1
Question 3: What is the effect of IBFLE test organization on the preparation of the students? **a.** Differences by genders **b.** Differences by IBFLE achievement **c.** Differences by IBFLE modules	11-16	2
Question 4: According to test takers, what changes does IBFLE need? **a.** Differences by academic titles **b.** Differences by IBFLE achievement **c.** Differences by IBFLE modules	17-23	3-4
Question 5: What competences are gained through IBFLE? **a.** Differences by genders **b.** Differences by IBFLE modules	24-29	

Prior to actual study, a pilot study was undertaken by the instructors working at the School of Foreign Languages. The colleagues were language instructors in English, French, or German. Their professionally-driven engagement in IBFLE and SEFLE presented expert opinion. They took both the questionnaire and the interview. The written

reflections along with the verbal comments were used to revise both the questionnaire and the interview. Further, the pilot study led to a reduction both in the questionnaire and the interview, eliminating the items with little or no relevance.

3.3 Respondents

The questionnaire was aimed to cover the best of 575 faculty members (163 professors, 127 associate professors, and 285 assistant professors) while only 161 of them were either available or volunteering to participate. They were all visited in person and asked for participation in the study. Further, the ones that were unavailable at the time of survey took the questionnaire via e-mail (n=17). Despite knowing that the great majority of the (full) professors had not taken IBFLE and thus would not take the questionnaire, they were double checked for participation so as to ensure the exceptional respondents. As a result, only 4 (3%) professors were confirmed and included in the study. The majority of the respondents was represented by assistant professors (n=115, 71%), who were relatively younger and thus took IBFLE later than the others. With a total of 42 (26%), associate professors came the second.

3.4 Instruments

The present study targeted a group of faculty members who had taken IBFLE for academic purposes. Depending on the classification of research methods by Creswell (2009: 285), a ‘mixed-method research’ was found appropriate for the present study. In this case, the process to be assessed only relied on personal statements or experiences, thus outlawing the feasibility of methodological triangulation (Pan, 2008: 12), and thus reducing the instruments to a questionnaire and an interview.

3.4.1 The Questionnaire

Data were collected by means of a questionnaire containing 29 items. The questionnaire was partially adapted from Arslan (2005: 127) and Elaldı (2005: 96) and was further detailed with the items related to washback. The questionnaire consisted of a total of 29 items in five sections:

a) *the first section* –involving multiple-choice questions– finds out personal details such as age, gender, and profession;

b) *the second section* –in YES/NO-type scale– examines the preparation styles as well as the details of background education;

c) *the third section* –in 3-point Likert scale including (1) "Not at all", (2) "Somehow", and (3) "Very"– aims to find out the emphasis on the language (sub)skills in IBFLE;

d) *the fourth part* –constituting the largest section and constructed in 5-point Likert scale including (1) "Strongly Disagree", (2) "Disagree", (3) "Undecided" (4) "Agree", and (5) "Strongly Agree"– both assesses the effect of IBFLE test organization on the preparation of the students and presents possible amendments that test takers might consider suitable for the future of the test;

e) *the last section* –the very part related to washback– is the exclusive part that solely refers to those who have passed the test. In this section, the respondents are asked to declare what skills and profits, and to what extent, they have gained through IBFLE.

For 3-point Likert scale, the interval scale was assessed as 1.00 – 1.67; "Not at all", 1.68 – 2.33; "Somehow", and 2.34 – 3.00 "Very". As for the 5-point scale, the interval was assessed as 1.00 – 1.80; "Strongly Disagree", 1.81 – 2.60; "Disagree", 2.61 – 3.40; "Undecided", 3.41 – 4.20; "Agree", and 4.21 – 5.00; "Strongly Agree".

The reliability of the questionnaire was assessed via Cronbach's Alpha, which found the reliability of the 29-item Likert-type scale as .744.

3.4.2 The Interview

To gain more personal evaluations, the questionnaire was supported by an interview of five open-ended questions, which sought to obtain further evaluations of the ideas formulated in the questionnaire. Unlike the questionnaire, the interview was conducted for a limited number of respondents (n=10) –only for the ones who accepted to be interviewed. The interview consisted of five questions, all of which were aimed to provide deeper insight for the outcomes elicited in the questionnaire:

The first question read "How important is foreign language in your academic life? In what circumstances (listening, speaking, reading, writing) do you use a foreign

language?", which would both figure out the importance of foreign language with regards to its context in IBFLE and narrow down this importance to the levels of language skills –in an attempt to gain more specific responses regarding the role of IBFLE in the learnability of each skill per se.

The second question mainly sought to classify the preparation methods employed by the respondents while trying to pick out the special techniques (if any) so as to present them to future test takers of IBFLE.

The third and fourth questions queried the opinions about the usefulness of IBFLE along with modification proposals for the amendment of IBFLE.

The final question, reading "Do you have any other remarks?", was specifically added to the interview in order to enable the respondents to express their feelings that have not been queried either via the questionnaire or the interview.

3.5 Data Collection

For the present study both quantitative (via the questionnaire) and qualitative (via the interview) data were collected. The data on the numbers and titles of the faculty members were elicited from the online academic database of the university. Depending on these data, there existed 575 faculty members (163 professors, 127 associate professors, and 285 assistant professors) at the time of the survey. At first hand, all the offices were visited in person and the questionnaire was offered to anyone available during the visits, and a total of 144 (89.4%) took the questionnaire. As for the second stage, 10 (6.2%) of them volunteered to take the interview beside the questionnaire. The interviews were conducted in the native language (*i.e.* Turkish) so as to enable a comforting atmosphere for the respondents. The ones that were not available during the visits were asked to participate via e-mail, among whom 17 (10.6%) responded.

3.6 Data Analysis

The quantitative data were analysed using Statistical Package for Social Sciences 17.0 for Windows (SPSS). The analysis included means, standard deviations, percentages, frequency rates, and the values showing reliability. For the analysis of variants, two tests were used, where *t*-test was used for the cases with two variants and

One-Way ANOVA was used for the cases with more than two variants. A post hoc LSD test was also performed to find out the significant differences among the groups.

The qualitative data were handled by content analysis. Both the written and recorded interviews were transcribed after each interview. The translations were carried out by the researcher and were double checked by a colleague proficient in English. For the qualitative data analysis, the responses were grouped depending on their similarities. A category for each grouping was made and similar patterns and arguments were listed in the same category. By this way, it was considered that the qualitative data could be presented in a sensible order and in an intelligible way.

In this chapter, the methodology used in the present study, along with the instruments employed and their collection and analysis procedures were presented. The results elicited through the instruments are expounded in Chapter Four.

CHAPTER FOUR

RESULTS

4.1 Presentation

This chapter is based on the results elicited from the data in the questionnaires and interviews, along with their aspects discussed through the washback perspective.

4.2 Profiles of the Respondents

The first section of the questionnaire was aimed to find out the demographic profiles of the respondents. These profiles were handled under four categories: (I) demographic data (age, gender, and academic title), (II) faculty / school (data on the workplaces of the respondents), (III) IBFLE experiences (number of IBFLE attempts, highest scores achieved, and the language and module taken), and (IV) educational backgrounds (private courses taken, overseas education, etc.). The first three categories existed as multiple-choice or open-ended items while the final one consisted of Yes/No-type questions.

4.2.1 Demographic Data for the Respondents

The questionnaire commenced with the items related to the respondents' demographic data including their gender, age, and academic title. The 161 respondents comprised of 134 (83.2%) males and 27 (16.8%) females, which conveyed a male preponderance over females. The majority of the respondents were within the range of 31-40 years of age (51.5%). As for their titles, the assistant professors (71.4%) had a landslide majority over the other two titles:

Table 8. Demographic profiles of the respondents

Variable	Category	f	%
Gender	Male	134	83.2
	Female	27	16.8
Age	21-30	4	2.5
	31-40	83	51.5
	41 and over	74	46.0
Title	Assistant Professor	115	71.4
	Associate Professor	42	26.1
	Professor	4	2.5

A total of 575 professors (163 professors, 127 associate professors, and 285 assistant professors) were reported at the time of survey and 161 of them partook in the study, which revealed that 28.0% of the population (2.4% of professors [4/163], 33.1% of associate professors [42/127], and 40.4% of assistant professors [115/285]) partook in the study. The number of full professors was much smaller than the other two since they had taken FLTAP instead of IBFLE. The rate of associate professors, too, could have been higher if all of them had taken IBFLE instead of FLTAP or SEFLE. As for the ages, almost all the respondents were older than 31 years of age (97.5%).

4.2.2 Faculty / School

As IBFLE is imposed for any faculty member –regardless of their majors, the respondents were chosen from all the faculties and schools that were located at the main campus in the main province of Diyarbakir. All the faculty members were visited in person, but only 144 (89.4%) of them were either available or volunteering to participate. The remaining 17 (10.6%) received the questionnaire via e-mail.

Table 9. Respondents by faculty / school

Faculty / School	f	%
Faculty of Agriculture	12	7.4
Faculty of Architecture	11	6.8
Faculty of Arts	11	6.8
Faculty of Dentistry	11	6.8
Faculty of Economics & Administrative Sciences	4	2.5
Faculty of Engineering	14	8.7
Faculty of Education	27	16.8
Faculty of Law	8	5.0
Faculty of Medicine	18	11.2
Faculty of Natural & Applied Sciences	23	14.3
Faculty of Theology	11	6.8
Faculty of Veterinary	8	5.0
Vocational School	3	1.9
TOTAL	**161**	**100.0**

The overall participation rate for these faculties/schools was almost parallel with the total number of respondents in each institution. Among these, the Faculty of Education had the highest rate of participation (16.8%) and the Vocational School the smallest (1.9%).

4.2.3 IBFLE Experiences

The aim of this category was to specify the details regarding the respondents' experiences with IBFLE. These details consisted of how many times they had tried IBFLE, the highest score they had achieved, along with the language and module they had taken:

Table 10. Details on the respondents' IBFLE experiences

Variable	Category	f	%
IBFLE Attempts	Once	43	26.7
	Twice	18	11.2
	Three or more times	100	62.1
Highest Score achieved in IBFLE	0-64	29	18.0
	65-79	120	74.5
	80-100	12	7.5
IBFLE Achievement*	*Failed*	29	18.0
	Passed	132	82.0
IBFLE Language	English	149	92.5
	German	7	4.3
	French	5	3.1
IBFLE Module	Natural & Applied Sciences	78	48.4
	Health Sciences	37	23.0
	Social Sciences	46	28.6

*Failed; <65%; Passed; >65%

It was revealed that most of the respondents had taken the test three or more times (62.1%), which entails that IBFLE is a challenging test which is not easily passed at the first trial. The respondents were also asked to write the highest score they had achieved in their IBFLE attempts. The scores elicited were scaled depending on three notions:

a) **0-64%:** Considering that a minimum of 65% in IBFLE is required for achieving the test, the scores in this group (below 65%) were accepted as "failed scores" since they are not sufficient for passing the test;

b) 65-79%: The scores between 65 and 79% were regarded as the "satisfactory scores" in that they are sufficient for the achievement of 65% but not so high to be hailed as high scores.

c) 80-100%: The scores between 80 and 100% were classified as the "high-level scores" in that they are the highest scores which require more preparation than the ones that are merely known as "to-the-test" studies.

It was proven that an outright majority of the respondents had received satisfactory scores (74.5%), which endorses that IBFLE has an overwhelming emphasis on the threshold score of 65%. Form this results, it follows that the higher scores are not so significant in that the difference between 65% and 95%, for example, would mean nothing for a candidate for attaining a higher rank.

IBFLE is only administered in three languages: English, German, and French. As expected, English was revealed as the most common one (92.5%). The other two languages had probably been taken by those who had been taught by one of them during their undergraduate education or earlier.

Unlike SEFLE, IBFLE offers three different modules (Social Sciences, Applied & Natural Sciences, and Health Sciences), whose testing content (*i.e.* vocabulary and scientific data that are used to mould each version) is unique to its target group. Among these, the module for Natural & Applied Sciences, which is preferred by departments like Physics, Chemistry, Biology, Engineering, and Mathematics, was revealed as the most common one (48.4%).

4.2.4 Educational Backgrounds

In the first part of the questionnaire, the respondents were asked to declare the details of their IBFLE-aimed education by answering the Yes/No-type items. The items sought to answer whether the respondents received/receive any private tuition, private courses, overseas education, whether they made/make use of the internet while preparing for IBFLE, and whether their undergraduate foreign language education promoted their success in IBFLE –all handled *per se*.

4.2.4.1 Private Tuition

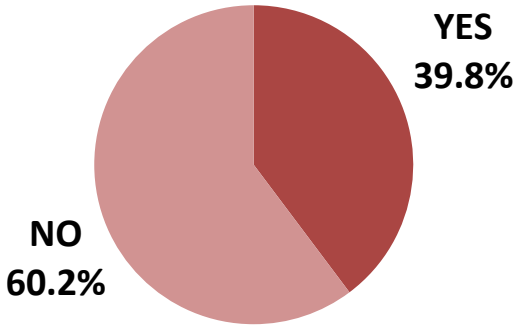

Figure 11. Statistical result for the item related to private tuition

It was found that the rate of negative answers (60.2%) was greater than the rate of positive answers (39.8%), which revealed that the results did not correspond with the rate of respondents that passed the IBFLE (82.0%).

4.2.4.2 Private Course / Class

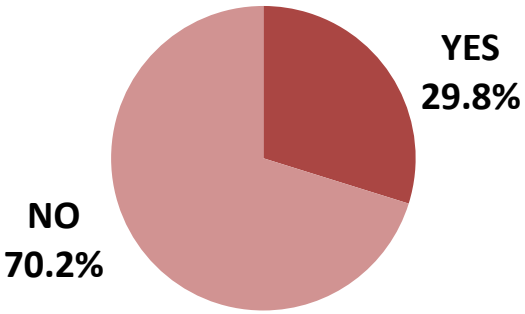

Figure 12. Statistical result for the item related to private course/class

As the results suggested, the number of respondents that took/take private tuition (39.8%) is higher than the ones who took/take courses / classes (29.8%). Also, the fact that an outright majority of the respondents (70.2%) did not appeal to private courses / classes and a great majority of them (60.2%) did not take private tuition embolden the potential existence of the notion called “self-study”. In other words, the respondents that have answered the final two questions as “Yes” may have prepared for IBFLE either through private tuition or courses rather than studying on their own.

4.2.4.3 Internet Usage

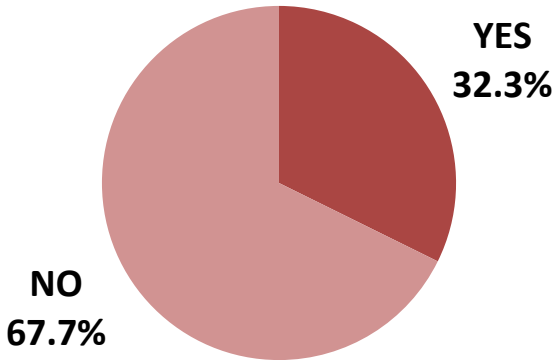

Figure 13. Statistical result for the item related to internet usage

IBFLE is a well-known test in Turkey; thus, it has a nation-wide popularity which is commonly cited and handled in online materials such as pre-IBFLE sources. Nevertheless, the results proved that only a small number of respondents (32.3%) made/make use of these materials, which is quite predictable in that the online materials can only supply linguistic content, *i.e.* vocabulary and grammar, but not the data on the practical tactics that are unique to IBFLE.

4.2.4.4 Overseas Education

It is a common belief that learning a foreign language is best achieved in a country where that language is spoken as a primary means of communication. To see how valid the stereotype is, the respondents were asked to state whether they had been abroad to study for IBFLE, in an attempt to see whether IBFLE is somehow considered as a test that requires an experience of overseas education:

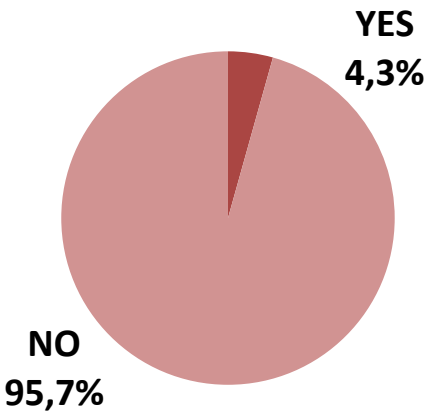

Figure 14. Statistical result for the item related to overseas education

The stereotype was annulled by the striking preponderance of the "No" answers (95.7%). Adversely, it is quite reasonable to presuppose that the number of the "Yes" answers would probably have been higher if IBFLE were a test that covered all four skills so that the overseas education would gain the test takers a distinguished experience in skills like listening and speaking. This is to prove that the predominance of two subskills (grammar and vocabulary) and one skill (reading) in IBFLE do not require a special initiation such as overseas education in that all these (sub)skills are perceptive, which do not necessarily need an authentic or native setting to improve.

4.2.4.5 The Effect of Undergraduate Foreign Language Education on IBFLE Success

The only foreign language that is taught at all undergraduate levels in Turkish universities is English. Weekly hours depend on the department, pending from two to eight hours. However, the greatest part of these syllabi is occupied by grammar teaching while all other (sub)skills are either handled as minor courses or totally neglected. The aim of this item was to obtain the respondents' views concerning the effect of this education on their IBFLE success:

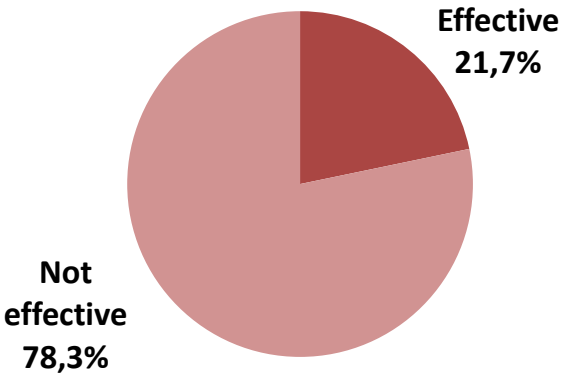

Figure 15. Statistical result for the item related to undergraduate education

The results suggested that the majority of the respondents (78.3%) consider that their undergraduate education had no effect on their IBFLE success. This could translate that (they consider that) they did not receive a systematic education at that level which naturally could not have had such an effect.

4.3 Quantitative Data Analysis

The present study is based on both quantitative and qualitative data. The quantitative data was obtained through the questionnaire while the qualitative data was elicited from semi-structured interviews. The quantitative data was handled under five distinct categories: (I) the relationship between IBFLE success and the factors denoted in the questionnaire (gender, IBFLE attempts, etc.), (II) the emphasis on the linguistic (sub)skills in IBFLE, (III) the effect of IBFLE test organization on the preparation of the students, (IV) the changes needed in IBFLE, and (V) the competences that are gained through IBFLE. For each category, descriptive data were elicited from the results obtained through the questionnaire, and the categories were analysed together with the associated factors.

4.3.1 The Relationship between IBFLE Success and the Factors (gender, IBFLE attempts, etc.)

In the questionnaire the respondents were asked to reflect their demographic and academic profiles as well as their background information, which were formulated as (I) the items that define personal details such as gender, age, title, and (II) and the items

including background information on their IBFLE preparations, such as private tuition, private courses / classes, etc. For each factor, descriptive tables were elicited to distinguish their relationship with IBFLE success:

Table 11. Frequencies and percentages denoting the relationship between IBFLE success and the factors

Factor	Variable	*Failed*		*Passed*		TOTAL
		n	%	n	%	
Gender	Male	27	20.1	107	79.9	134
	Female	2	7.4	25	92.6	27
Language	English	24	16.1	125	83.9	149
	German	4	57.1	3	42.9	7
	French	1	20.0	4	80.0	5
Module	Nat & App Sciences	8	10.3	70	89.7	78
	Health Sciences	3	8.1	34	91.9	37
	Social Sciences	18	39.1	28	60.9	46
Private Tuition	No	23	23.7	74	76.3	97
	Yes	6	9.4	58	90.6	64
Private Course / Class	No	16	14.2	97	85.8	113
	Yes	13	27.1	35	72.9	48
Internet Usage	No	15	13.8	94	86.2	109
	Yes	14	26.9	38	73.1	52
Overseas Education	No	27	17.5	127	82.5	154
	Yes	2	28.6	5	71.4	7
Undergraduate Foreign Language Education	No	24	19.0	102	81.0	126
	Yes	5	14.3	30	85.7	35
Total		29	18.0	132	82.0	161

As the table revealed, female respondents (92.6%) did better than males (79.9%) in IBFLE. In addition, their small number in the failed group (n=2, 7.4%) substantiate their higher success over males (n=27, 20.1%).

Of the languages offered in IBFLE, English (83.9%) presented the highest success, followed by French (80.0%) and German (42.9%). Nevertheless, the small

numbers of respondents in the latter groups are not enough to compare them with the larger group of English under equal circumstances.

The data showed an outright difference between the Social Sciences (60.9%) and the other two groups –Natural & Applied Sciences (89.7%) and Health Sciences (91.9%).

The respondents that received private tuition (90.6%) seem to have done better than the ones that did not receive it (76.3%). However, it does not necessarily mean that IBFLE is not achievable without taking private tuition since the success rate for those who passed the test without receiving private tuition remained at 76.3%. Compared with the overall rate of those who receive/d private tuition (39.8%), the rate of success for the passed group (90.6%) is remarkably important.

In addition to private tuition, there is another criterion that represents a differentiation –private courses / classes. Unlike private tuition, the candidates that received private courses / classes presented a lower success (72.9%) than the ones that did not receive them (85.8%).

The internet is perceived as a useful, easy, and cost-effective tool that is used for many purposes including entertainment, education, and communication. However, the results made it clear that the internet is not an *all-effective* device for IBFLE candidates in that the success rate for the candidates who made use of the internet in their preparation period (73.1%) was quite lower than the ones who did not use it (86.2%).

IBFLE is a nationwide test whose operations are confined to Turkey. Yet, there have been a very small number of candidates (n=7, 4.3%) who resorted to overseas education for IBFLE preparations. In addition to proving the scarcity of the people who receive overseas education for IBFLE, the data also disclosed that this group (71.4%) is not as successful as the one that has not tried overseas education (82.5%). This begs the question as to whether these respondents went abroad to specifically study for IBFLE or to do some other studies that may have little or no effect on their IBFLE success.

Foreign language education is provided through obligatory courses at all undergraduate levels in Turkey. On the strength of this data, the respondents were questioned as to whether (they consider) this education was any helpful in their IBFLE

success. Although a small number of them (n=35, 21.7%) replied that (they consider) this education was helpful in their success, a positive correlation was found between taking English courses at the undergraduate level and getting higher scores in the IBFLE.

4.3.2 The Emphasis on the Linguistic (Sub)skills in IBFLE

The assessment of the skills and subskills in IBFLE are collectively conducted under five categories of questions: (I) vocabulary, (II) grammar, (III) translation, (IV) dialogue, and (V) reading comprehension. The answers below were elicited to specify the ones that are most emphasized, *i.e.* the most challenging ones:

Table 12. Perceptions of the respondents about the difficulty level of the questions in IBFLE (n=161)

	Very		Somehow		Not at all				
Questionnaire Item	f	%	f	%	f	%	n	M	SD
6. The vocabulary questions in IBFLE are easy for me.	27	16.8	105	65.2	29	18.0	161	1.99	0.59
7. The grammar questions in IBFLE are easy for me.	46	28.6	100	62.1	15	9.3	161	2.19	0.59
8. The translation questions in IBFLE are easy for me.	134	83.2	27	16.8	0	0.0	161	2.83	0.38
9. The dialogue questions in IBFLE are easy for me.	84	52.2	67	41.6	10	6.2	161	2.46	0.61
10. The reading comprehension questions in IBFLE are easy for me.	50	31.1	84	52.2	27	16.8	161	2.14	0.68

The vocabulary questions (M=1.99) were proven as the most difficult type, followed by reading comprehension (M=2.14), grammar (M=2.19), dialogue (M=2.46), and translation questions (M=2.83). Depending on the intervals, the vocabulary, reading comprehension, and grammar questions are proven as "somehow easy" while the remaining two (*i.e.* dialogue and translation) as "very easy". This sequence is also meaningful in grading these questions in terms of level of challenging, where the translation questions (M=2.83) were proven as the least challenging and the vocabulary questions (M=1.99) as the most.

In order to see what factors are associated with the emphasis on these (sub)skills, both *t*-test (for cases with two variants) and One-way ANOVA (for cases with more than two variants) were conducted:

4.3.2.1 Differences by IBFLE Languages

IBFLE is a unique test in Turkey in that it has three versions differing in languages –English, German, and French. The notion that they are different from each other could lead to a difference in their emphasis on the subskills:

Table 13. Results of ANOVA for the relationship between IBFLE languages and the skills emphasized in IBFLE

	n	M	SD	f	Sig.
English	149	2.33	.31	2.871	.060
German	7	2.06	.61		
French	5	2.48	.30		
Total	161	2.32	.33		

$p > 0.05$

The data revealed that all three versions differ in terms of their emphasis on the subs-skills: English (M=2.33), German (M=2.06), and French (M=2.48). Accordingly, the French version (M=2.48) is perceived as the most challenging while the German version (M=2.06) is revealed as the least challenging. Indeed, only the French version is confirmed as "very easy" while the other two as "somehow easy". Nonetheless, it is possible to foresee that the results could have changed if there had been more respondents for German and French. As for the relationship, no significant correlation was detected (p=.060).

4.3.2.2 Differences by IBFLE Modules

In addition to three languages, IBFLE versions differ in terms of the modules that cover different vocabulary or topics for their target groups, including (I) Natural & Applied Sciences, (II) Health Sciences, and (III) Social Sciences. Among these, the respondents that took the first version constitute the largest group (48.4%). Module-based

differentiation among the respondents was taken as a potential factor for the emphasis on the (sub)skills, which is available with all three versions:

Table 14. Results of ANOVA for the relationship between IBFLE modules and the skills emphasized in IBFLE

	n	M	SD	f	Sig.
Natural & Applied Sciences	78	2.37	.27	2.679	.072
Health Sciences	37	2.34	.29		
Social Sciences	46	2.23	.43		
Total	161	2.32	.33		

$p > 0.05$

Although no correlation was found in terms of emphasis (p=.072), the version for the Social Sciences (M=2.23) yielded a relatively lower rate compared with the Natural & Applied Sciences (M=2.37) and Health Sciences (M=2.34). Depending on these rates, the Social Sciences is proven as "somehow easy" while the other two as "very easy". This difference had also been verified in the section related to IBFLE success, where the success rate for the test takers of Social Sciences (60.9%) was much lower than the ones for Natural & Applied Sciences (89.7%) and Health Sciences (91.9%). Collectively, the module for Social Sciences is revealed as much more challenging than the other two modules.

4.3.3 The Effect of IBFLE Test Organization on the Preparation of the Students

In this section, the respondents were asked to state to what extent they have been studying each language skill and subskill, in an attempt to question whether the test organization in IBFLE has any effect on the preparation of the students:

Table 15. Descriptives for the extent to which each (sub)skill has been studied

Questionnaire Item	N	M	SD
11. I tried/am trying to improve my listening skills to do better in IBFLE.	161	2.39	1.30
12. I tried/am trying to improve my speaking skills to do better in IBFLE.	161	2.22	1.23
13. I tried/am trying to improve my reading skills to do better in IBFLE.	161	3.96	1.20
14. I tried/am trying to improve my writing skills to do better in IBFLE.	161	3.14	1.36
15. I tried/am trying to improve my vocabulary to do better in IBFLE.	161	4.40	.98
16. I tried/am trying to improve my grammar to do better in IBFLE.	161	4.25	1.07

The results revealed that vocabulary (M=4.40) and grammar (M=4.25) had the highest rates, followed by reading (M=3.96), writing (3.14), listening (2.39), and speaking (M=2.22). Accordingly, vocabulary, grammar, and reading were proven as the indispensable skills in that the rates for these skills presented with "strongly agree". Nonetheless, the writing skill alone was proven as "undecided" meaning that the respondents are not decided whether the studies they have been doing actually pertain to the writing skill. As for the rest, listening and speaking were verified as the least emphasized skills in IBFLE preparations in that they were revealed as "strongly disagree".

For statistical differences, three divergence points were employed: (I) gender, (II) IBFLE achievement, the ones that have passed versus the ones that have not, and (III) IBFLE modules, differentiation of the test takers in terms of their majors:

4.3.3.1 Differences by Genders

Gender-based differentiation was assessed via a *t*-test:

Table 16. Results of *t*-test for the relationship between gender and the skills emphasized in IBFLE preparations

	n	M	SD	t	df	Sig.
Male	134	3.37	.86	-.893	159	.588
Female	27	3.52	.73			

$p > 0.05$

According to the test, there exists no significant difference in terms of gender (p=.588). However, it was revealed that the females presented a higher value (M=3.52), signifying that they exert slightly more effort than the males while studying for IBFLE. Nevertheless, both mean values were reflected in the same interval, –"agree", meaning that both groups put almost the same amount of effort in studying the IBFLE skills.

4.3.3.2 Differences by IBFLE Achievement

A skill-based analysis, assessing the amount of study paid to each skill *per se*, could also reveal the overall amount of the effort a test taker has exerted in his/her IBFLE preparations. For the overall scores, IBFLE achievement was taken as a potential factor:

Table 17. Results of *t*-test for the relationship between IBFLE achievement and the skills emphasized in IBFLE preparations

	n	M	SD	t	df	Sig.
Failed	29	3.26	.85	-.957	159	.340
Passed	132	3.42	.83			

$p > 0.05$

Predictably, the passed group (M=3.42) had a higher score than the failed group (M=3.26), suggesting that there is a positive correlation between studying the language skills and passing the IBFLE. Indeed, the interval for the failed group presented with "Undecided" while the other as "Agree", which denotes that the passed group studied more than the failed group did. Both two outcomes point towards the notion that the more attention one pays to IBFLE skills the higher scores one could get in the test. Yet, there was no significant difference between the groups (p=.340).

A descriptive figure was obtained to see to what extent each (sub)skill is studied by each group (the passed and failed groups):

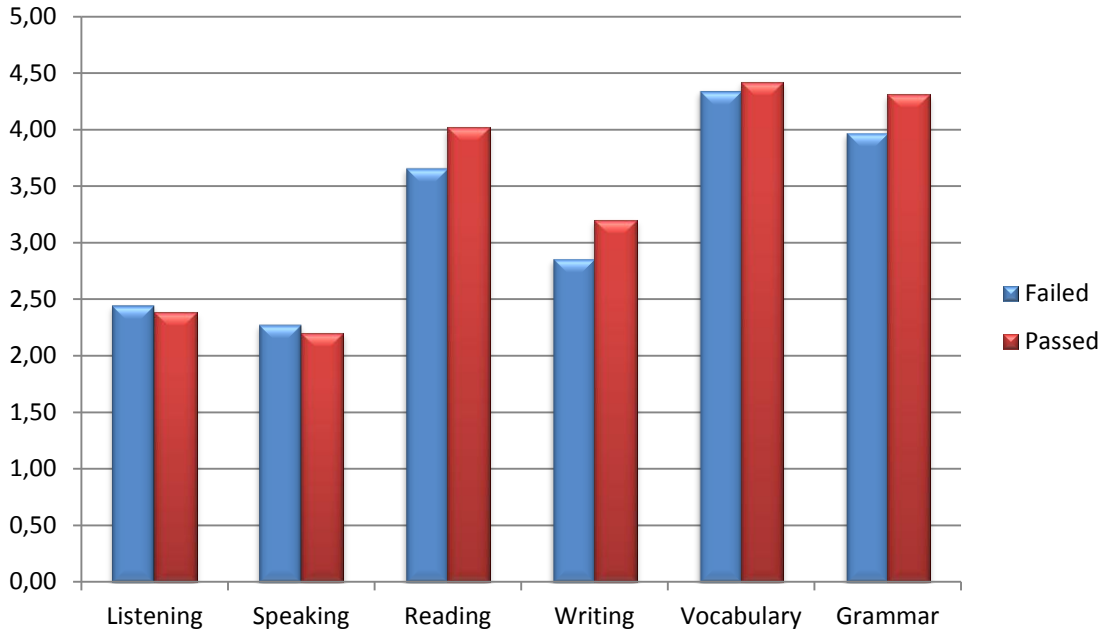

Figure 16. Achievement-based comparison of the skills emphasized in IBFLE preparations

The results signify that listening (M=2.45) and speaking (M=2.28) have been paid more attention to by the failed group while the other four (sub)skills (reading [M=4.02], writing [M=3.20], vocabulary [M=4.42], and grammar [M=4.31]) have been studied more actively by the passed group. Further, the only subskill where two groups presented different intervals is grammar, in which the passed group had "strongly agree" (M=4.31) and the other "agree" (M=3.97).

4.3.3.3 Differences by IBFLE Modules

IBFLE modules categorize the test takers in terms of their majors as well. The effect of this classification was taken as a factor on the language skills emphasized in IBFLE preparations:

Table 18. Results of ANOVA for the relationship between IBFLE modules and the skills emphasized in IBFLE preparations

	n	M	SD	f	Sig.
Natural & Applied Sciences	78	3.56	.81	3.392	.036
Health Sciences	37	3.29	.92		
Social Sciences	46	3.19	.78		
Total	161	3.39	.84		

$p < 0.05$

The statistics denoted a statistical significance among the three groups (p=.036), meaning that there is a remarkable difference among the three groups in terms of the amount of study devoted to IBFLE (sub)skills. The group of Natural & Applied Sciences produced the highest rate (M=3.56), which was followed by Health Sciences (M=3.29) and Social Sciences (M=3.19), where only the Natural & Applied Sciences presented with "Agree" and the other two as "Undecided".

In order to see where the significant difference exists (*i.e.* between/among which groups), a post hoc LSD test was performed (Table 19):

Table 19. Results of Post Hoc LSD test for the relationship between IBFLE modules and the skills emphasized in IBFLE preparations

Section (I)	Section (J)	Mean (I-J)	Sig.
Nat & App Sciences	Health Sciences	.276	.096
	Social Sciences	.376	.015*
Health Sciences	Nat & App Sciences	-.276	.096
	Social Sciences	.100	.584
Social Sciences	Nat & App Sciences	-.376	.015*
	Health Sciences	-.100	.584

$p < 0.05$

The test revealed that the difference exists between the Natural & Applied Sciences and Social Sciences (p=.015), *i.e.* between the groups which exert the greatest and the least effort in IBFLE skills.

Further, the mean value for each module was compared in each skill in order to get a skill-based insight into the differentiation denoted by ANOVA:

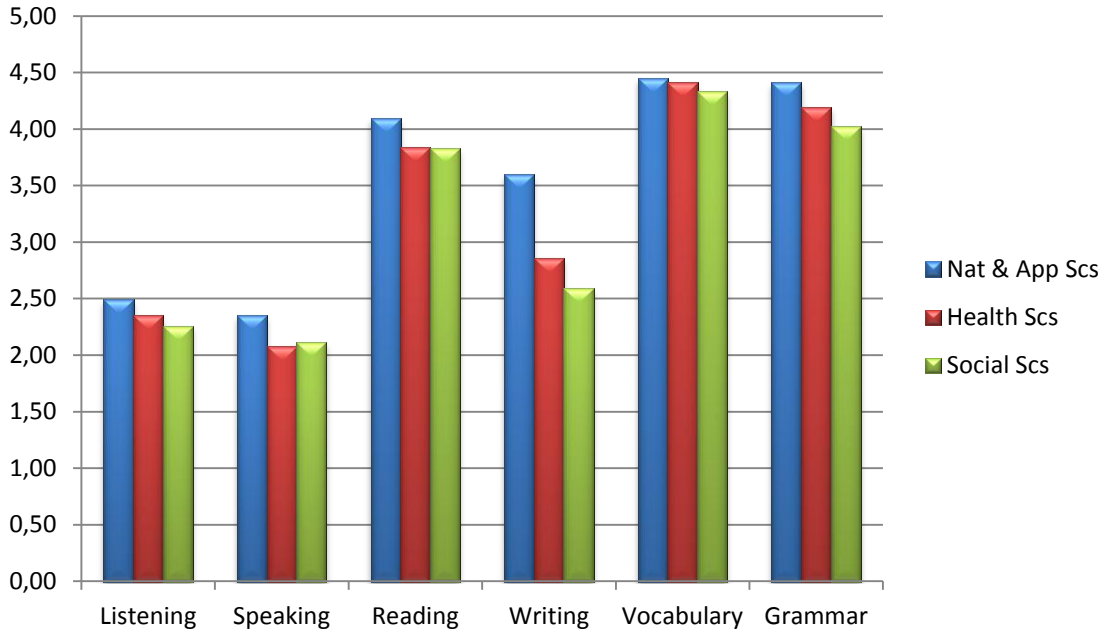

Figure 17. Module-based comparison of the skills emphasized in IBFLE preparations

It is quite obvious from the statistics that in all the skills the Natural & Applied Sciences had the highest and the Social Sciences had the lowest rates. On the basis of the intervals, there are two (sub)skills, grammar and writing, where the mean values presented different intervals: in grammar, the Social Sciences (M=4.02) and Health Sciences (M=4.19) were proven as "Agree" while the Natural & Applied Sciences (M=4.41) yielded "Strongly agree"; as for writing, the differentiation got deeper, in that the Social Sciences (M=2.59) remained at "Disagree", the Health Sciences (M=2.86) at "Undecided" and the Natural & Applied Sciences (M=3.60) at "Agree". Contrary to these results, the group with the leading rates, the Natural & Applied Sciences, had not been proven as the most successful group in IBFLE scores, but the second most successful, following the Health Sciences.

4.3.4 The Appropriateness of IBFLE

On the strength of interviews in the pilot study, a list of seven items was formed to elicit the respondents' views on the necessities for IBFLE. The items were related to the level of challenge in IBFLE, the test duration of IBFLE, the appropriateness of IBFLE in assessing foreign language competence, the variety of languages in IBFLE, IBFLE's coverage in four skills (*i.e.* listening, speaking, reading, writing), the implemental frequency of IBFLE, and the mergence of IBFLE with SEFLE. The results revealed that the item related to four skills (M=3.65) yielded the highest rate while the item relating the appropriateness (M=2.34) was revealed as the lowest.

Table 20. Descriptives for the changes proposed for IBFLE

Questionnaire Item	N	M	SD
17. The level of challenge in IBFLE is approvable.	161	3.27	1.21
18. The test duration for IBFLE is adequate for answering all the questions.	161	3.48	1.26
19. IBFLE is a proper tool for assessing foreign language competence.	161	2.34	1.19
20. The variety of languages in IBFLE should be enriched.	161	3.24	1.28
21. IBFLE should be modified so as to cover all four skills (*i.e.* listening, speaking, reading, writing).	161	3.65	1.24
22. IBFLE should be administered more frequently.	161	3.34	1.24
23. IBFLE should be merged with SEFLE.	161	2.86	1.27

Regarding the results, the item related to the level of challenge was revealed as M=3.27, signifying that the respondents are undecided in this sense. A total of 180 minutes (3 hours) is allowed at the test, which is agreed by the respondents to be quite sufficient (M=3.48). Conveying the lowest rate, the item concerning the appropriateness of IBFLE (M=2.34) implies that the test takers disagree with the idea that IBFLE is a proper tool in the assessment of linguistic competence. Unlike SEFLE, IBFLE is limited to three languages –English, German, and French. Depending on the results, the respondents are undecided as to whether more languages are needed in IBFLE (M=3.24). On the other hand, the need for an all-covering testing style (M=3.65) was agreed by the

respondents and it was revealed as the most crucial need for IBFLE. IBFLE is a biannual test, administered every March and October. The results revealed that the respondents are undecided about an elevation in the implemental frequency of IBFLE (M=3.34). As mentioned earlier, IBFLE and SEFLE are quite alike. The respondents were questioned so as to see the degree of similarity perceived by test takers. The results indicated that the respondents are undecided about such a mergence (M=2.86).

4.3.4.1 Differences by Academic Titles

The academic titles of the faculty members were considered as a potential factor on the changes needed in IBFLE:

Table 21. Results of ANOVA for the relationship between academic titles and the changes proposed for IBFLE

	n	M	SD	f	Sig.
Professor	4	3.57	.65	1.937	.148
Associate Professor	42	3.27	.53		
Assistant Professor	115	3.12	.61		
Total	161	3.17	.59		

$p > 0.05$

The data indicated that the level of advocacy was parallel with the academic titles, where the professors got the highest rate (M=3.57), followed by the associate professors (M=3.27) and the assistant professors (M=3.12). Meaningfully, while the professors agree with the changes, the remaining two groups remain undecided. Despite this differentiation, there was no correlation established among the variables (p=.148).

4.3.4.2 Differences by IBFLE Achievement

In order to see whether there is a difference between the perceptions of the failed and passed groups in terms of the changes needed in IBFLE, a *t*-test was conducted:

Table 22. Results of *t*-test for the relationship between IBFLE achievement and the changes proposed for IBFLE

	n	M	SD	t	df	Sig.
Failed	29	2.71	.61	-4.872	159	.000
Passed	132	3.27	.54			

p 0.05

The results verified a significant difference between the two groups (P=.000). Additionally, it was revealed that there is a remarkable difference between the mean values, where the failed group (M=2.71) conveyed a relatively smaller support for the changes in IBFLE when compared with the passed group (M=3.27); yet, both values remained at the level of "undecided".

4.3.4.3 Differences by IBFLE Modules

Just like IBFLE languages, IBFLE modules too categorize the test takers in three groups: Natural & Applied Sciences, Health Sciences, and Social Sciences. Knowing that the test takers for each module are of different majors, their views on IBFLE changes were anticipated to differ:

Table 23. Results of ANOVA for the relationship between IBFLE modules and the changes proposed for IBFLE

	n	M	SD	f	Sig.
Natural & Applied Sciences	78	3.16	.51	.052	.950
Health Sciences	37	3.20	.72		
Social Sciences	46	3.16	.63		
Total	161	3.17	.59		

p > 0.05

The results did not provide a statistical significance among groups (p=.950). Indeed, no remarkable difference was revealed among the mean values for each group; all three had quite similar values: Natural & Applied Sciences and Social Sciences both had M=3.16 and Health Sciences had a slightly higher rate of M=3.20, where all three of them are undecided about the changes proposed for IBFLE.

4.3.5 The Competences Gained through IBFLE

The present study made a research on a group of 161 faculty members working at Dicle University to find out the washback of a well-known foreign language test, IBFLE, on their professional activities and studies. In the questionnaire conducted, the final section was specifically designed only for those who had taken and passed the test, in attempt to specify the abilities and gains they had obtained through their IBFLE achievement. Hence, only the passed group (n=132, 82.0%) took this section to declare what they had gained through IBFLE. The IBFLE gains were formulated in six items: the first two were related to general gains while the remaining four pertained to four linguistic skills (speaking, listening, writing, and reading, respectively) (Table 24):

Table 24. Descriptives for the competences gained through IBFLE

Questionnaire Item	N	M	SD
24. Preparing for IBFLE helped me to learn the target language.	132	3.02	1.32
25. I am using my IBFLE gains in my professional studies.	132	3.20	1.31
26. I can easily speak the target language in the settings where foreign people partake.	132	2.55	1.16
27. I can easily understand a person speaking in the target language in academic meetings.	132	2.86	1.21
28. I can write academic articles using the target language.	132	3.08	1.27
29. I can easily understand the academic papers in my field that are written in the target language.	132	3.95	1.08

The first of the two general items was to question the general learnability of the language through IBFLE. The statistics revealed that the respondents are “undecided” about the idea that IBFLE helps learn the target language (M=3.02). The second general item was related to the usability of IBFLE gains in professional settings such as classroom activities, article writing and reading, and presentations. It was also revealed that the respondents are “undecided” about the usability of IBFLE gains (M=3.20).

As for the four skills, the item concerning the reading skill (Item #29) was revealed with the highest value (M=3.95), which proved that the respondents agree that

the reading skill is improvable through IBFLE. Reading was followed by writing (Item #28, M=3.08) and listening (Item #27, M=2.86) – about which the respondents feel undecided – and speaking (Item #26, M=2.55) – whose improvability was disagreed by the respondents.

4.3.5.1 Differences by Genders

In order to see the differences between male and female respondents in terms of IBFLE gains, a *t*-test was conducted:

Table 25. Results of *t*-test for the relationship between gender and the competences gained through IBFLE

	n	M	SD	t	df	Sig.
Male	107	3.13	.84	.690	130	.491
Female	25	3.00	1.00			

$p > 0.05$

The test proved that there exists no significant difference between groups (p=.491) in terms of IBFLE gains. However, it was revealed that the males (M=3.13) had a slightly higher rate than the females, denoting that the males apparently gained more skills and/or knowledge from IBFLE studies than the females did. Nevertheless, the mean values imply that both groups are undecided about the competences gained through IBFLE.

4.3.5.2 Differences by IBFLE Modules

Previously, it was revealed that the group of Social Sciences had the lowest rate in IBFLE success. Depending on this fact, IBFLE modules were taken as a possible factor for devising the outcomes of IBFLE preparations in terms of IBFLE modules:

Table 26. Results of ANOVA for the relationship between IBFLE modules and the competences gained through IBFLE

	n	M	SD	F	Sig.
Natural & Applied Sciences	70	3.20	.91	.742	.478
Health Sciences	34	3.02	.93		
Social Sciences	28	2.99	.68		
Total	132	3.11	.87		

$p > 0.05$

The data in this section bears ample parallelism with the previous results in that they both conclude that the test takers in Social Sciences (M=2.99) are the ones that feel most challenged in IBFLE-related processes. On the one hand, although the group of Health Sciences had a higher rate of IBFLE achievement (91.9%) than the group of Natural & Applied Sciences (89.7%), the latter group (M=3.20) is revealed as the one that gained much more profit than did the former one (M=3.02). Collectively, all three groups were confirmed as undecided about the competences gained through IBFLE.

4.4 Qualitative Data Analysis

In accordance with the mixed style of the study, an interview was conducted as a complementary unit for the quantitative data. The interview consisted of five open-ended questions which asked for a detailed account of the intricacies included in IBFLE. Prior to response analyses, percentages relating to the similarity of the responses were obtained for each question *per se*:

4.4.1 The Importance of Foreign Language for the Test Takers

In the first question in the interview, the respondents were asked to state their opinions relating the importance of IBFLE (*i.e.* in what areas they use it) in their academic lives. The responses elicited were graphed depending on their similarities:

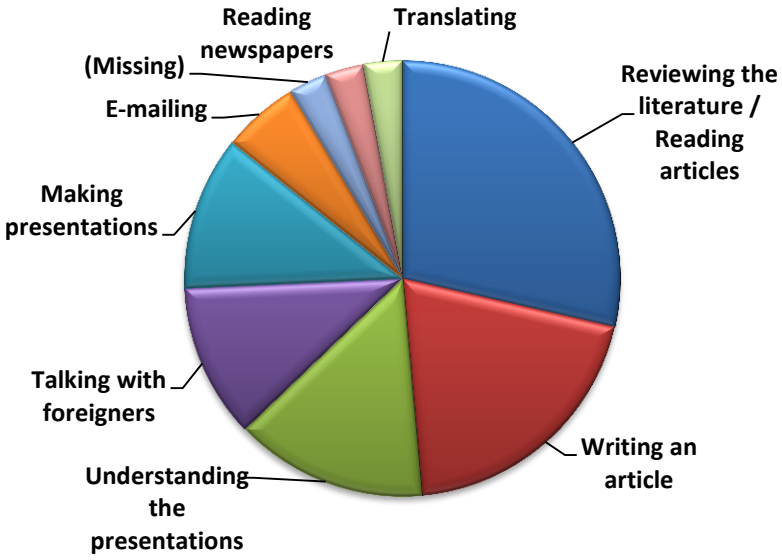

Figure 18. The responses elicited for the Interview Question #1

As revealed, the respondents need the target language mainly for three notions: reviewing the literature (29%), writing an article (20%), and understanding the presentations (14%). These notions are followed in sequence by talking with foreigners (11%), making presentations (11%), e-mailing (6%), reading newspapers (3%) and translating (3%). Depending on these notions, it is possible to confer that the mono-skill structure of the IBFLE (*i.e.* the reading skill being dominant) is not sufficient for university faculty members to fulfil their academic transactions both in their domestic activities and foreign relations –both of which not only include literature reviewing but also article writing, talking with foreigners, creating effective presentations, and e-mailing. These needs were expressed by several respondents as in the following:

> *"I use it for writing articles and e-mails, and for talking to foreign friends and guests." (Respondent 4)*

> *"It is very important. I use it in academic conferences and presentations, for e-mailing with foreigners, and for reading articles." (Respondent 7)*

> *"It is quite important. I use it while reviewing the literature." (Respondent 9)*

There were also some respondents who expressed a detailed account of their needs:

"It is very important for me. I use it in all the skills, but specifically in reading articles and in doing speaking and listening in international conferences. It is also very important in writing, but I am not so good at writing." (Respondent 33)

"It is very important. I use it both for reading and writing articles, and also it is very useful for reading foreign newspapers." (Respondent 49)

"I use it both for making presentations and understanding the ones that are presented at academic settings, and also I use it for talking to foreigners." (Respondent 50)

"It is an indispensable part of my academic life because the majority of the literature in my field is written in English. As for the skills, I am OK in speaking but I do have some difficulties in writing. It is much needed in conferences and presentations such as panels since none of the attendants have to speak my language (Turkish), rather they speak English - some of them may be able to speak Turkish though." (Respondent 58)

"It is an indispensable tool for writing international papers. I use it both for reviewing the literature and for translating. It is something really important. It is also something needed in international symposiums." (Respondent 72)

"It is important in every aspect; I mean, it is something very important for communicating with academicians from foreign universities. Yet, the importance of reading and writing should also be noted." (Respondent 64)

4.4.2 The Techniques Employed in IBFLE Preparations

For the second question in the interview, the respondents were asked to answer "What special techniques have you been using while preparing for IBFLE?":

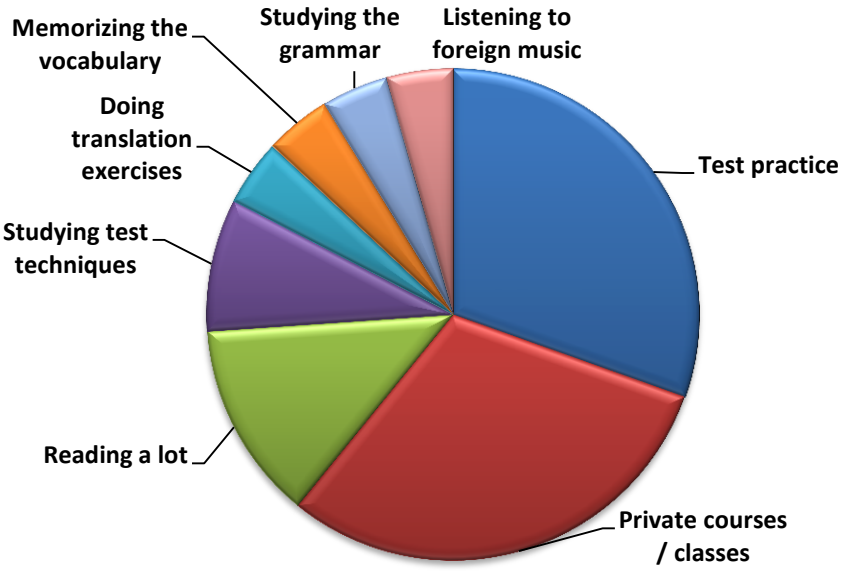

Figure 19. The responses elicited for the Interview Question #2

The results revealed an outright preponderance of the two notions –self-study (30%) and private course / classes (30%). The remaining notions included reading (13%), test techniques (9%), translation (4%), memorization (4%), grammar (4%), and listening to foreign music (4%). From these results, it follows that self-study and private course / classes remain the most popular (and seemingly the most useful) techniques in IBFLE preparations. Yet, they are not proven as the sole source of achievement since the other techniques are also revealed as preferable. The responses advocating this outcome were formulated in the following sentences:

"I did lots of test practice." (Respondent 4)

"I read lots of books and stories and I attended a private course." (Respondent 7)

"I attended a private course which taught me test methods and tactics." (Respondent 9)

"I preferred self-study for IBFLE. There seems to be a problem with our educational system; all the countries but Turkey can achieve foreign language teaching." (Respondent 11)

"I studied myself by doing a lot of test practice." (Respondent 33)

Indeed, there were some respondents who used more than one technique either simultaneously or subsequently:

> *"I did test practice while attending a course and I paid special attention to learning the test technique. I improved myself through intensive studies on reading, vocabulary, and translation." (Respondent 17)*

> *"I began with reading books but then I received private tuition and courses." (Respondent 50)*

> *"I started by studying the grammar and went on with reading. However, what I did most was read newspapers and listen to music, which I think help remember the vocabulary and the other structures more easily." (Respondent 58)*

> *"I tried to study by myself but I gave it up when I saw that I wouldn't have enough time for that. Later on, I took private lessons which provided better results." (Respondent 72)*

> *"I took private courses by the instructors working at our university. But now I am doing test practice and I review the literature relevant to my major as well." (Respondent 85)*

Additionally, there was a respondent who stated his/her obligatory swap between the techniques due to unexpected outcomes:

> *"I was taught by a teacher at first but that education cost me unbelievable mistakes in the language. Then, I started concentrating on the very questions that are solely prepared for IBFLE." (Respondent 64)*

4.4.3 Attitudes towards the Appropriateness of IBFLE

The appropriateness of IBFLE was queried by means of the question that read "Do you agree that IBFLE fulfils its purpose?". The aim of this formulation was to strengthen the mere notion of appropriateness with the idea of purpose, in an attempt to elicit how the test takers perceive the purpose of IBFLE as integrated with the notion of appropriateness:

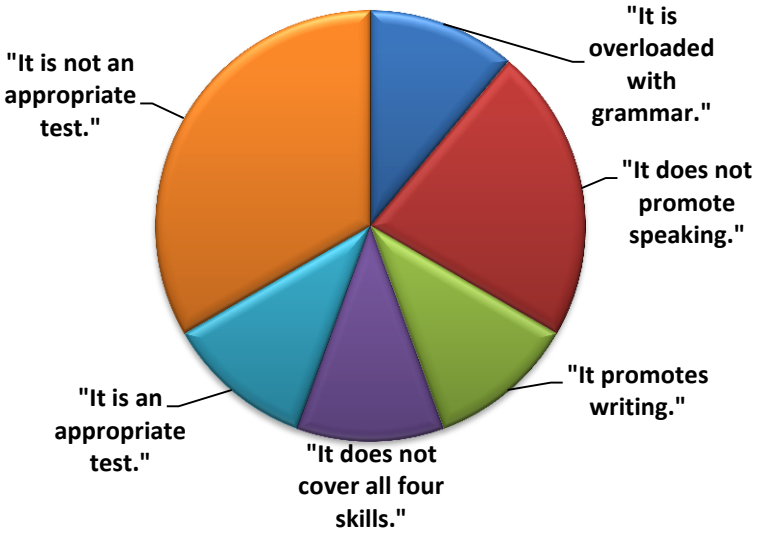

Figure 20. The responses elicited for the Interview Question #3

The responses to this question provided a set of six different views, four of which are classified as negative (66.7%) and the remaining two as positive (33.3%). The positive responses are roughly formulated as "It is an appropriate test." and "It promotes writing" which do not provide any content for the reason of advocacy. Such responses were expressed as follows:

"It is OK." (Respondent 11)

"It only fulfils its purpose when we observe that it functions like an incentive in promoting language learning and improvement." (Respondent 64)

As for the negative responses, the majority of them utterly rejected the appropriateness of the test, adding a reason to the rejection:

"Definitely no! Grammar alone is not sufficient in this realm." (Respondent 4)

"No! Those who pass IBFLE cannot speak but only some of them can write articles." (Respondent 7)

"No. IBFLE is only aimed to assess grammar and reading comprehension. It is not a sufficient instrument in language assessment." (Respondent 49)

"No. It is not sufficient in assessing speaking skills, which are the basic needs of academicians." (Respondent 50)

"I strongly disagree with this idea. It cannot assess linguistic competence at all. Even a person who gets all the questions right in IBFLE cannot prove that he/she has learnt that language very well. What he/she does is to show that he/she has learnt all the structures and vocabulary; however, that person cannot perform what he/she has learnt." (Respondent 58)

"Definitely no! It only imposes a score of 65%, which is not concerned with proving the real language proficiency. As an academician, I would expect it to impose communication skills which we would use in communications and academic sharings." (Respondent 85)

Some of the negative responses not only presented their rejection but also provided a suggestion for the modification of IBFLE:

"Not really! It would be more suitable if it covered all four skills." (Respondent 9)

"IBFLE should aim to improve speaking and communication skills as well." (Respondent 17)

"It should better be called 'OFLE' (Obligatory Foreign Language Examination) since it only imposes us to get the sufficient score and do nothing else." (Respondent 72)

Unlike the others, one of the respondents stated both his/her rejection while acquiescing to the idea that IBFLE is the second best to none:

"I don't think it does but it seems like there is no better exam than IBFLE in terms of language assessment." (Respondent 33)

4.4.4 Attitudes towards the Changes Needed in IBFLE

In accordance with the close-ended items in the fourth section of the questionnaire, the interviewees were asked to indicate their attitudes towards the changes they consider suitable for IBFLE. Regarding the question, "Do you think IBFLE should be modified? In what ways?", a great majority of the interviewees responded with a

straightforward “Yes” (90.0%) to emphasize their enthusiasm for the advocacy of changes in IBFLE:

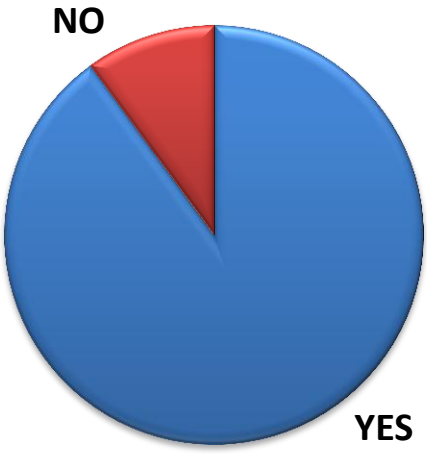

Figure 21. The responses elicited for the Interview Question #4

Most of the responses demanded a fundamental change, specifically suggesting a “TOEFL-like style”:

“Of course it should change. It should be a TOEFL-like test that covers all four skills.” (Respondent 4)

“It should be a TOEFL-like test that covers speaking and listening skills.” (Respondent 7)

“Yes. It should cover all four skills.” (Respondent 9)

“It should cover all four skills and the culture of the language as well.” (Respondent 17)

“The testing system should change; it should not only focus on grammar and vocabulary.” (Respondent 49)

“No, but all four skills could be handled providing that a proper infrastructure is established.” (Respondent 50)

“Definitely yes! It could be a TOEFL-like test; it should cover all four skills so that we can call it a standardized test.” (Respondent 58)

Additionally, there were some respondents who proposed some *more fundamental* actions:

> *"In my opinion, the obligation of linguistic assessment should be avoided for academicians. Instead, some encouraging activities should be installed in an attempt to promote their willingness to speak and write. Also, the achievers should be rewarded. It is a crucial fact that many of our associate professors and Full professors are not able to speak a foreign language." (Respondent 72)*

> *"The system itself should change. Language assessment doesn't necessarily have to be hailed as 'IBFLE'; rather, a more comprehensive test which may be called something different from IBFLE should be enacted." (Respondent 85)*

> *"The module for Social Sciences should be divided into subtests as History, Philosophy, etc. The vocabulary of my major (History) is not sufficiently handled in IBFLE. " (Respondent 11)*

On the other hand, there were some respondents who remained hesitant while supporting their views with various reasons:

> *"It should better remain as it is, because it is not easy to adapt to a new version." (Respondent 33)*

> *"Yes but it would become much more difficult if it included all four skills. Yet, this would be an advantage for freshmen like research assistants in that they would necessarily have to improve all four skills." (Respondent 64)*

4.4.5 Additional Remarks

The final question in the interview was aimed to elicit additional remarks (if any) in an attempt to reveal independent notions that were unmentioned in the questionnaire or in the interview. Among the responses obtained, there was only one response that referred to unmentioned reviews:

> *"IBFLE is an awkward test; while the mean score for the success in the module of Social Sciences does not exceed 47%, the score in other two modules is about 60%." (Respondent 11)*

This response provided a parallel outcome with the statistics that revealed the same result for the success rate among IBFLE modules –the statistics which had proven that the Social Sciences is the most challenged group in IBFLE.

CHAPTER FIVE

DISCUSSION AND CONCLUSION

5.1 Presentation

This chapter is based on the results elicited from the data in the questionnaires and interviews, along with their aspects discussed through the washback perspective.

5.2 Discussion and Conclusion

The present study was aimed to study the effect of a well-known Turkish test on its test takers through a technical perspective called 'washback'. As the population of the study, full-time faculty members at Dicle University, located in the Province of Diyarbakir, Turkey, were selected. Of 575 faculty members, a total of 161 (28.0%) faculty members were studied. There were 115 assistant professors (71.4%), 42 associate professors (26.1%), and 4 professors (2.5%), among whom there were 134 males (83.2%) and 27 (16.8%) females. The respondents were chosen from all the faculties and schools (n=13) present at the time of the survey, among which the Faculty of Education had the highest rate of participation (16.8%) and the Vocational School had the smallest (1.9%).

In addition to demographic profiles, the respondents' IBFLE-related backgrounds were also elicited. The results revealed that 43 (26.7%) respondents had taken the test only once, 18 (11.2%) of them twice, and 100 (62.1%) of them for three or more times. The highest scores achieved were categorized on a three-grade system which classified 0-64% as the first grade (n=29, 18.0%), 65-79% as the second grade (n=120, 74.5%), and 80-100% as the third – the highest – grade (n=12, 7.5%). Of these, the first grade was labelled as the "failed" group and the remaining two as the "passed", depending on the passing criterion of IBFLE –the 65% criterion. Among the languages tested in IBFLE, the outright majority belonged to English (n=149, 92.5%), which was followed by German (n=7, 4.3%) and French (n=5, 3.1%). As for the IBFLE modules, Natural & Applied

Sciences represented the largest group with a number of 78 (48.4%) respondents while Social Sciences had 46 (28.6%) and Health Sciences 37 (23.0%).

As for the IBFLE-related backgrounds, it was found that 64 (39.8%) of the respondents receive/d private tuition. In addition, 48 (29.8%) respondents attend/ed private courses/classes. The results also revealed that 52 (32.3%) of the respondents use/d the internet for IBFLE preparations. Overseas education for IBFLE was also assessed, which proved that only a very small number of the respondents (n=7, 4.8%) resorted to this style as opposed to the remaining 154 (95.2%). In the final item of this section, the respondents were asked to convey whether they think their foreign language education at undergraduate level had any (positive) effect on their IBFLE success. The statistics revealed that only 35 (21.7%) of them confirmed such an effect.

The remaining data on quantitative and qualitative results were handled under the research questions per se:

Research Question 1: Is There a Relationship between IBFLE Success and Factors (gender, IBFLE languages, IBFLE modules, etc.)?

It was revealed that the females did better than the males in IBFLE, though the female/male ratio was 1/4.96. This result became predictable through the results on the skills emphasized in IBFLE, which had proven that females exert more effort than males. In this way, it is revealed that the amount of effort devoted to the IBFLE skills is in a direct proportion with the success to be received in IBFLE.

Of the languages offered in IBFLE, English presented the highest success rate, which was followed by French and German. This may pertain to the overwhelming interest in English as compared with other two. As a prestigious and ubiquitous medium of instruction (Crystal, 1989: 358; Cook, 2003: 25, 26), English is the most appealing language, which is, in this realm, seen as the language of choice in IBFLE processes. Additionally, this ubiquity may be leading to a proliferation in the materials produced and thus widening the spectrum to be used in linguistic processes. Thus, the number of applicants opting for English along with their success rates remains remarkably higher than the ones for any other languages.

Among the modules in IBFLE, the Health Sciences was proven as the most successful one, followed by Natural & Applied Sciences and Social Sciences. The vast difference between the Social Sciences and the other two could best be explained by the notion of acquaintance –which holds that the contexts for Health Sciences and Natural & Applied Sciences, *i.e.* the contents of their IBFLE versions, convey more relevance with the professions of their target groups when compared with the respondents taking the version of Social Sciences. Therefore, the test takers for these versions may be more advantageous than the ones for Social Sciences in that they are intrinsically loaded during their educations with the fundamental knowledge they need in IBFLE.

The data revealed that private tuition is an important tool in passing IBFLE in that the ones that took/take it were more successful than the others. Supposing that this method both provides linguistic content (*i.e.* vocabulary, grammar, and reading comprehension skills) and offers some tips and tactics that are unique to IBFLE, private tuition can be a good way not only for feeling oneself on the right path to success but also for limiting oneself in terms of studying the correct material and content for IBFLE.

Compared with private tuition, private courses / classes presented a lower success in IBFLE. The main distinction between private tuition and private courses / classes is the distinction between *single-way* and *multi-way*; in private tuition the candidate performs his/her studies under the supervision of a tutor who is only charged with one person and is ready to control and check him/her whenever and however needed, on the other hand a classroom setting is a multifaceted atmosphere in which the respondents vary depending on their successes and also their threat levels, which may collectively pose potential challenges for the linguistic improvement of others. Indeed, a classroom setting is more limited than private tuition in that the attention is divided to many people instead of one. All in all, private tuition is proven to be more effective than private courses as regards the IBFLE success.

It was proven that the respondents did not use the internet for IBFLE preparations. This may be a corollary of two notions: (I) the internet does not cover the practical tactics which IBFLE candidates could best use in their preparations, but rather it provides them with fundamental corpus and grammatical knowledge which could be found in any pre-IBFLE book, (II) the internet is a virtual world in which there are not only supportive

materials but also disruptive materials which could either mislead the candidates or prevent them from doing their target studies.

There were only seven (4.3%) candidates who tried overseas education for IBFLE preparations. Indeed, they presented a lower success than the ones who did not try it. This may be a signpost for the confinement of IBFLE to the limits of Turkey in that the *so-called* better education to be received in a native country like the UK or the USA would have less to do with IBFLE when compared with the preparations conducted in Turkey. Moreover, since IBFLE is prepared by Turkish specialists rather than native speakers, the preparations performed in Turkey are likely to provide more fruitful outcomes when compared with the ones conducted outside Turkey.

Mainstream foreign language education at undergraduate degrees was confirmed as an effective factor in IBFLE success. Small though the number was (n=35, 21.7%), the respondents declared that (they consider) this education was helpful in their success. As reported by Amrein & Berliner (2002a: 35), this may be an implication that this education is subliminally performing a *to-the-test* syllabus which intrinsically provide students with test techniques. This theory is overtly supported in the study by Sevimli (2007: 126), which revealed that the high school education that prepared the students for the Foreign Language Component of the University Entrance Examination (FLCUEE) in Turkey implemented a to-the-test syllabus which mainly focused on the (sub)skills assessed in the test, – reading, grammar, and vocabulary – and that it ignored the remaining ones, which averted the students from improving them.

Research Question 2: What Language (Sub)skills are Emphasized in IBFLE?

Both quantitative and qualitative data were employed to provide a detailed answer to this question; a two-way answer was established to merge this question with the interview question #1 such that the quantitative data would provide the (sub)skills emphasized in IBFLE and the qualitative ones would provide an insight into the (sub)skills as to why they are needed. The quantitative data indicated that the vocabulary questions were proven as the ones that are most emphasized and thus the most challenging ones, followed by reading comprehension, grammar, dialogue, and translation questions. Remarkably, the same (sub)skills were revealed to be emphasized

in the speaking exam at Gazi University Preparatory School of English. The study was conducted by Güllüoğlu (2004: 265), which revealed that although the test is supposed to emphasize the speaking skill, it tends to pay more attention to other (sub)skills like grammar and vocabulary. The qualitative data in the present study yielded similar results, proving that academicians need the foreign language mainly in areas like literature reviewing, article writing and understanding the presentations. In particular, there were some other areas noted in the interviews such as talking with foreigners, making presentations, and writing e-mails, which are directly related to speaking and writing skills –neither of which is directly tested in IBFLE. In doing so, the respondents declared that they actually needed more than what is provided to them through IBFLE.

a. Differences by IBFLE Languages

Though no correlation was established in this case, it was found that all three languages in IBFLE differ in terms of their emphasis on the (sub)skills, where the French holds the highest rate, followed by English and German. Meaningfully, English yielded an intermediary level, proving that the larger number of its respondents led to a less diverse level when compared with others –which are far less in number.

b. Differences by IBFLE Modules

IBFLE modules, too, yielded no correlation but all three versions provided different mean values. Significantly, the Social Sciences revealed the smallest rate as compared with others, verifying that this version is highly different from the others in this sense as well. Supposing that this emphasis is reflected as the level of challenging, the Social Sciences is – once again – confirmed as the most challenging module for IBFLE.

Research Question 3: What is the Effect of IBFLE Test Organization on the Preparation of the Students?

It was revealed both in quantitative and qualitative data that the (sub)skills emphasized in IBFLE were not parallel with the ones studied during pre-IBFLE processes. This is a clear evidence to maintain that the IBFLE candidates deal with more than what they are required to, *i.e.* they do not limit their studies to the scopes of the skills required in IBFLE. In sum, both the quantitative and qualitative results converged at the

notion that IBFLE preparations do not only consist the most emphasized (sub)skills (reading, vocabulary, grammar) but also the least emphasized ones (and even the ones not tested at all) such as listening, speaking, and writing. A similar discrepancy was revealed by Cesur (2009: 60), where it was proven that foreign language is not tested in the standardized achievement test administered at Çanakkale Onsekiz Mart University in the way it has been taught in the course.

a. Differences by Genders

The results pointed to a clear parallelism between the skills emphasized in IBFLE preparations and the IBFLE success in terms of genders. It was revealed that females put more effort into their studies than males do, which brings them a higher success than the one for males.

b. Differences by IBFLE Achievement

The results made it clear that the passed group has been dealing with the most emphasized (sub)skills (reading, vocabulary, and grammar) more than the failed group. Indeed, the skill-based analysis also confirmed this theory, proving that listening and speaking have been paid more attention to by the failed group while the other four (sub)skills have been more actively studied by the passed group. This marks that there is a positive correlation between studying "more-related" (sub)skills (*i.e.* reading, writing, vocabulary, and grammar) and gaining higher scores in IBFLE.

c. Differences by IBFLE Modules

Among the modules in IBFLE, the Natural & Applied Sciences presented the highest and the Social Sciences the lowest rate. However, the results were not parallel with the sequence for IBFLE success since the most successful group had been proven as the Health Sciences, followed by the Natural & Applied Sciences. This pertains to the notion that the Health Sciences has spent less effort on the "less relevant" (sub)skills and thus gained a higher success than the Natural & Applied Sciences did. Furthermore, the Social Sciences presented the lowest rates both in overall and skill-based comparisons, both of which correlate with its lowest rate in IBFLE success.

Research Question 4: According to Test Takers, What Changes Does IBFLE Need?

Both in quantitative and qualitative results it was revealed that the respondents do not regard IBFLE as a proper tool in the assessment of foreign language competence. The views are supported with open remarks, mainly complaining about too much grammar, and single-skill style (and thus ignoring speaking, listening, and writing).

The notion related to the grammar was also merged – by some respondents – with the exceeding amount of vocabulary. It is inferred that both of these amounts would inevitably be less – or be at a more reasonable amount – imagining the IBFLE included other skills like listening, speaking, and writing –as denoted in the second notion which was related to the single-skill style of IBFLE. Also, some respondents suggested a TOEFL-like test, which, they think, would radically alleviate the complaints against IBFLE. All considered, there is a remarkable demand of change in the linguistic content in IBFLE.

In item-based analysis, the respondents remained undecided about almost all of them. Among these, the notion concerning the merging of IBFLE with SEFLE gained the least support, meaning that the respondents are unsure as to whether IBFLE and SEFLE are two distinct tests which should be merged anyway no matter how more alike they became through the amendments. For the enrichment of the variety of the languages in IBFLE, the respondents also remained undecided –although none of them mentioned the languages to be added to IBFLE. In doing so, they might have reckoned that new languages are likely to bring about new challenges. The respondents were undecided about the level of challenge in IBFLE as well. The reason for this is likely to be concerned with the style of IBFLE itself, which, in effect, focuses on the reading skill alone. The respondents might be hesitant to argue whether the test is rightfully challenging as a single-skill test, or to compare its level of challenge with the ones in an all-inclusive test (*e.g.* TOEFL). The third notion that test takers were undecided about is the implemental frequency of IBFLE. They might have considered that a more frequent implementation, *e.g.* once in three months or every month, might provide the opportunity of taking the exam whenever the preparations end or the test takers feel ready but on the other hand might bring about a sense of disorganization in planning when to start and end the preparations, which would also distract their overall attention. Similarly, it was

proven by Tarhan (1992: 61) that frequent testing has no effect in promoting the students' success on a test. Accordingly, there seems to be no use in raising the implemental frequency of IBFLE.

The only item agreed by the respondents was the one related to the test duration (3 hours). By agreeing that the test duration is sufficient for a total of 80 questions, the respondents probably do not feel crammed for keeping up with the end of the test duration, and thus no change is needed in this sense.

a. Differences by Academic Titles

As reported by Fournier-Kowaleski (2005: 201), who found that the past teaching and testing experience contributes to a change in the attitudes and behaviours towards a test, the level of advocacy in the present study was parallel with the academic titles, where the professors got the highest and the assistant professors the lowest. This provides a clear notion that the ascension of an academic title equally promotes the faculty member's support for the need of change in IBFLE. Further, it is inferred that the respondents have gained more support towards the changes in IBFLE as they got promoted. Meaningfully, the advocacy level towards the need for changes in IBFLE gets higher and higher as one gets more experienced and thus becomes more aware of what points deserve to be changed.

b. Differences by IBFLE Achievement

The passed group presented a higher support for IBFLE changes than the failed group did. The greatness of this difference could be adhered to the idea that the test takers who cannot pass the test are likely to produce some anguish against the potential challenges which might be brought up by the changes in IBFLE. Adversely, the passed group – although they have passed the test and thus they have no further expectations from it – might be considering that the test deserves some changes for the sake of forthcoming candidates.

c. Differences by IBFLE Modules

The results made it clear that although they had differed in various cases (IBFLE success, the (sub)skills emphasized in IBFLE and the ones emphasized in IBFLE preparations), the three modules did not present any significant difference in terms of supporting the IBFLE changes. All three got almost the same rates, proving that they bear the same level of advocacy for the changes in IBFLE.

Research Question 5: What Competences are Gained through IBFLE?

The present study aims to find out the effects of a well-known standardized test in Turkey, IBFLE, which is principally administered to assess the foreign language competence for faculty members or their candidates. This aim was embedded in a procedure called washback effect (Alderson & Wall, 1992: 2; Brown, 1995: 92; 2002: 11), which refers to the effect of the test on its test takers. In the first four questions, the results were elicited to define both test-based and taker-based details in terms of the intricacies and challenges as well as the success rates. The final question seeks to formulate what competences are gained through IBFLE-related processes, *i.e.* the pre-IBFLE processes and the test itself.

Among the four skills, the highest rate pertains to the reading skill, hence justifying the monopoly of the reading skill in IBFLE. The results also correlate with the skills emphasized in IBFLE preparations, where the reading skill had received the highest rate among all four skills. In this way, it became quite apparent that the reading skill is the only linguistic skill which could spontaneously be improved via IBFLE to the extent that the improvement of the writing skill was also proven to be a corollary of IBFLE. This is a key notion for the faculty members in that – they stated that – they use the foreign language mostly in reading their literature, which is directly concerned with the reading skill. This, in a way, justifies the usability of IBFLE among faculty members both for teaching and assessment of the reading skill.

Of the less-relevant skills (*i.e.* writing, speaking, and listening), writing presented a higher value, verifying that this skill is subsequent to the reading skill in that they are alike in terms of the mechanics (organization, structure, punctuation, etc.) and that they complement each other as a combination of perceptive-productive skill –the former being

reading and the latter writing. The remaining skills, speaking and listening, were confirmed as the least relevant skills, though speaking is indirectly yet scarcely assessed in IBFLE. Nonetheless, the number of the respondents studying these skills was quite low, which signifies that these respondents are mistakenly involved in a misleading study without knowing that these skills do not have much to do with IBFLE. Surprisingly, no matter how wrongfully they exert their energy on studying speaking and listening, they state that they cannot speak their target language. In the study by Duran (2011: 75), it was noted that both teachers and students consider that speaking is a crucial part of their tests which should not be ignored in the test so as to have it learned by the test takers. Accordingly, the learnability of speaking (and most probably of listening as well) highly depends on its active participation in IBFLE. Another assumption might hold that IBFLE is stereotypically regarded as a challenging test which subconsciously influences test takers by feigning an all-inclusive test and thus misleading the test takers towards futile studies. On the contrary, the test takers, unable to improve speaking and listening through IBFLE, vocally expressed their need for an all-covering test which would gain them these skills subsequently.

As for the remaining two items, the respondents remained undecided about both of them. For the first item, they presented their uncertainty about the IBFLE's effect on language learning. A similar outcome is reported by Reynolds (2010: 51), who studied a total of 23 respondents to gain their views on the washback on TOEFL. The study revealed that TOEFL students are unsure if preparing for TOEFL contributed to language learning. The reasoning behind this may depend on the respondents' insight into what a language is and what it is used for. They might have reckoned that knowing a language is not confined to being able to read and write academic papers only. Indeed, they might have recalled their highly debated need for an all-inclusive test, –a test that would teach them not only reading and writing but also listening and speaking. As for the second item, the respondents also remained unsure about the usability of IBFLE gains in academic activities, such as classroom instruction, article writing, and reviewing the literature. This could imply that the respondents cannot reach a consensus to assert that the IBFLE gains, which basically consist of the outcomes of reading, vocabulary, and grammar are useful in academic activities in that many of them declared a demand for the changing of IBFLE towards an all-inclusive style.

Concerning the washback effect of IBFLE, it is possible to assert that IBFLE presents both positive and negative washback. The positive washback regards the promotion of reading skill, whereby it was proven that IBFLE takers are enacted to improve their reading skills, no matter consciously or subconsciously, likely because the test is overwhelmingly reading-based. Conversely, the test results in a negative washback in that it not only de-emphasizes the remaining three skills (listening, speaking, writing) but also leads to a confusion among test takers about what and how to study for the test. This, in turn, results in curricular changes both in self-studies and classroom settings like private courses.

a. Differences by Genders

In terms of IBFLE success, females had presented a higher success over males. However, the statistics revealed that males declared a higher rate in terms of IBFLE gains, proving that they have received more profit through IBFLE than the females have. This makes it clear that the females –although they do better in the test– fall behind in performing what they have received through the test in their daily and academic lives.

b. Differences by IBFLE Modules

According to the statistics, the mean values regarding IBFLE gains were quite parallel with the ones related to the skills emphasized in IBFLE and preparations, where the Natural & Applied Sciences got the highest and the Social Sciences the lowest, but they were slightly different from the ones in IBFLE success, where the Natural & Applied Sciences followed the Health Sciences. In brief terms, although the Natural & Applied Sciences exert more effort in IBFLE preparations, the Health Sciences presents a higher success –even the highest. The remaining group, the Social Sciences, however, got the lowest rank in all these areas, substantiating the earlier assumption that this group is the most challenged one in terms of IBFLE success and gains.

5.3 Implications for Test Takers

The implications drawn out to create a scientific basis for future test takers can be outlined as follows:

- Among study styles, private tuition was proven as the most effective one, followed by private course / class and overseas education.
- Among supportive studies, undergraduate foreign language education confirmed quite useful, whereas internet usage and specifically overseas education did not present a significant influence on IBFLE success.
- The vocabulary questions were verified as the most challenging ones in IBFLE, followed by reading, grammar, dialogue and translation questions.
- Studying reading and writing along with vocabulary and grammar were confirmed to promote IBFLE success while speaking and listening remained as the least relevant and thus the least useful skills in IBFLE preparations.
- IBFLE is best functional in promoting the reading skill as opposed to other three skills.

Through these conclusions, future test takers may not only decide how to study but also will be aware of what they can achieve through their studies.

5.4 Implications for Policymakers

The present study came up with some noteworthy issues to be presented to official authorities known as policymakers:

- A great many respondents regard IBFLE as insufficient in the assessment of linguistic proficiency. In asserting this, they clearly state that the test should assume an all-inclusive style which would actively cover not only reading but also the remaining three skills, *i.e.* listening, speaking, and writing. In doing so, also, the complaints regarding overwhelming grammar and vocabulary could be eliminated.
- Among IBFLE modules, the Social Sciences was proven as the most disadvantaged group, presenting the least success rate among the modules. With a view to alleviating this challenge, the module for Social Sciences could be broken into novel modules, *e.g. humanities, arts, statistical sciences,* which would create more focused and thus less comprehensive modules for the test takers in this group.

5.5 Suggestions for Further Research

The present study shed light on the washback effect of a well-known language proficiency test in Turkey, called IBFLE. As the population, the full-time faculty members working at Dicle University, Turkey, were selected. Depending on these limitations, further researchers may focus on (I) other universities in Turkey, (II) full-time instructors (the university teachers that are lower than the faculty members in rank) working either in the same institution or at other universities, or (III) they may choose another test – SEFLE, TOEFL, or IELTS – to conduct washback studies.

REFERENCES

Adams R. J., & McNamara, T. F. (1998, March). *Detecting and evaluating the impact of multidimensionality in language test data*. Paper presented at the Language Testing Research Colloquium, Monterey, CA.

Al-Jamal, D., & Ghadi, N.(2008). English Language General Secondary Certificate Examination Washback in Jordan. *The Asian EFL Journal: 10 (3),* 158-186.

Alderson, J. C. (1998). Testing and teaching: the dream and the reality. *Novelty, 5 (4),* 23-27.

Alderson, J. C., & Banerjee J. (2001). Language testing and assessment (Part 1). *Language teaching, 34,* 213-236. doi: 10.1017/S0261444801001707

Alderson, J. C., & Banerjee J. (2002). Language testing and assessment (Part 2). *Language teaching, 35,* 79-113. doi: 0.1017/S0261444802001751

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

Alderson, J. C., & Wall D. (1992, February). *Does washback exist?*. Paper presented at a Symposium on the Educational and Social Impacts of Language Tests, Language Testing Research Colloquium, Florida.

Alghamdi, G. A. (2006). Backwash: the impact of a core proficiency test on the learning/teaching situation: a case study. *Language and translation: 19,* 1-15.

Allison, D. (1999). *Language testing and evaluation*. Singapore: Singapore University Press.

Amrein, A. L., & Berliner, D. C. (2002a). *An Analysis of Some Unintended and Negative Consequences of High-Stakes Testing* (Report No. EPSL-0211-125-EPRU). Tempe: Arizona State University.

Amrein, A. L., & Berliner, D. C. (2002b). *The Impact of High-Stakes Tests on Student Academic Performance: An Analysis of NAEP Results in States With High-Stakes Tests and ACT, SAT, and AP Test Results in States With High School Graduation Exams* (Report No. EPSL-0211-126-EPRU). Tempe: Arizona State University.

Andrews, S., Fullilove, J., & Wong, Y. (2002). Targeting washback—a case-study. *System, 30*, 207-223.

Arslan, E. (2005). *Analysing ÜDS (English proficiency exam for academicians) exam questions* (Master's thesis). Retrieved from YÖK Database (146080).

Ataman, F. (2004). *A study on reliability and validity of METU School of Foreign Languages Department of Basic English June 1998 and September 1998 proficiency tests.* (Master's thesis). Retrieved from YÖK Database (82246).

Aydın, S. (2004).*The efficiency of computers on testing writing skills.* (Doctoral dissertation). Retrieved from YÖK Database (144645).

Bachman, L. F. (1995). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Cohen A. D. (1998). *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press.

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

Bailey, K. M. (1999). *Washback in language testing*. Princeton: Educational Testing Service.

Banerjee, M., & Shaw, S. F. (2007). High-stakes test accommodations: documentation review by testing agencies in an era of change. *Assessment for Effective Intervention, 32 (3),* 171-180. doi: 10.1177/15345084070320030501.

Brinckerhoff, L. C., & Banerjee, M. (2007). Misconceptions regarding accommodations on high-stakes tests: recommendations for preparing disability documentation for test takers with learning disabilities. *Learning Disabilities Research & Practice, 22 (4),* 246-255.

Brown, H. D. (2003). *Language assessment: Principles and classroom practices*. London: Longman.

Brown, J. D. (2000a). University entrance examinations: strategies for creating positive washback on English language teaching in Japan. *Shiken: JALT Testing & Evaluation SIG Newsletter, 3 (2),* 2-7.

Brown, J. D. (2000b). What is construct validity?. *Shiken: JALT Testing & Evaluation SIG Newsletter. 4 (2),* 8-12.

Brown, J. D. (2002). Extraneous variables and the washback effect. *Shiken: JALT Testing & Evaluation SIG Newsletter, 6 (2),* 10-13.

Brown J. D., & Yamashita S. O. (1995). English language entrance examinations at Japanese universities: 1993 and 1994. In Brown J. D., Yamashita S. O., Griffee, D. T., & Troyer G. V. (Eds.), *Language testing in Japan.* (pp. 86-100). Tokyo: The Japan Association for Language Teaching.

Byram, M. (2004). *Routledge encyclopedia of language teaching and learning*. London: Routledge.

Carmines, E. G. & Zeller R. A. (1979). *Reliability and validity assessment*. California: Sage Publications, Inc.

Cesur, Ç. (2009). *An assessment of the validity of the standardized achievement test administered at Çanakkale Onsekiz Mart University.* (Master's thesis). Retrieved from YÖK Database.

Cesur, K. (2008). *Students' and teachers' perceptions of the test techniques used to assess language skills at university level.* (Master's thesis). Retrieved from Çanakkale Onsekiz Mart University Database (82246).

Chapelle, C. A. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing and research*. Cambridge: Cambridge University Press.

Charles, D. C. (2008). *Teachers' perspectives on the unintended consequences of high stakes testing.* (Doctoral dissertation). Retrieved from Louisiana State University and Agricultural and Mechanical College Database.

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

Cheng, L. (1999). Changing assessment: washback on teacher perceptions and actions. *Teaching and Teacher Education, 15,* 253-271.

Cheng, L. (2000). *Washback or backwash* (Report No. FL 026 284). Kingston: Queen's University.

Cheng, L. & Curtis A. (2009). The realities of English language assessment and the Chinese learner in China and beyond. In Cheng, L. & Curtis A. (Eds.), *English language assessment and the Chinese learner* (pp. 3-12). Routledge: New York.

Cheng L., Watanabe Y., & Curtis A. (2008). *Washback in language testing*. Mahwah: Lawrence Erlbaum Associates, Publishers.

Christopher, N. M. (2008). Social and educational impact of language assessment in Nigeria. *Nordic Journal of African Studies, 17 (3),* 198-210.

Chu, L., & Gao, P. (2006). An empirical study of the washback of CET-4 writing. *Sino-US English Teaching, 3 (5),* 36-38.

Clapham, C. (2000). Assessment and testing. *Annual Review of Applied Linguistics, 20,* 147-161.

Cook, G. (2003). *Applied Linguistics*. New York: Oxford University Press.

Creswell, J. W. (2009). *Research Design.* California: Sage Publications, Inc.

Crystal, D. (1989). *The Cambridge encyclopedia of language.* Cambridge: Cambridge University Press.

Djurić, M. (2008). Dealing with Situations of Positive and Negative Washback. *Scripta Manent, 4 (1),* 14-27.

Dooey, P. (2008). Language testing and technology: problems of transition to a new era. *European Association for Computer Assisted Language Learning, 20 (1),* 21-34. doi: 10.1017/S0958344008000311.

Downey, R., & Suzuki, M., & Moere, A. V. (2010). High-stakes English-language assessments for aviation professionals: supporting the use of a fully automated test of spoken-language proficiency. *IEEE Transactions on Professional Communication, 53 (1),* 18-32.

Duran, Ö. (2011). *Teachers' and students' perceptions about classroom-based speaking tests and their washback.* (Master's thesis). Retrieved from Bilkent University Database.

Elaldı, S. (2005). *Analysing KPDS (English proficiency exam for state employees) exam questions* (Master's thesis). Retrieved from YÖK Database (145287).

Finch, A. E. (2002). Authentic assessment: Implications for EFL performance testing in Korea. *Secondary Education Research, 49,* 89-122.

Fournier-Kowaleski, L. A. (2005). *Depicting washback in the intermediate Spanish language classroom: a descriptive study of teacher's instructional behaviors as they relate to tests.* (Doctoral dissertation). Retrieved from Ohio State University Database.

Genesee, F. and Upshur, J. A. (1996). *Classroom-based Evaluation in Second Language Education*. Cambridge: Cambridge University Press.

Güllüoğlu, Ö. (2004). *Attitudes towards testing speaking at Gazi University preparatory school of English and suggested speaking tests.* (Master's thesis). Retrieved from Gazi University Database.

Hayes, B. M. (2003). *IELTS preparation in New Zealand: an investigation into the nature of the courses and evidence of washback.* (Doctoral dissertation). Retrieved from Victoria University of Wellington Database.

Heubert P., & Hauser R. M. (1999). *High stakes testing for tracking, promotion, and graduation*. Washington D. C.: National Academy Press.

Hughes, A. (1992). *Testing for languages teachers*. Cambridge: Cambridge University Press.

Hughes, A. (1993). *Backwash and TOEFL 2000*. England: University of Reading.

Hsu, H. (2009). *The impact of implementing English proficiency tests as a graduation requirement at Taiwanese universities of technology.* (Doctoral dissertation). Retrieved from University of York Database.

IELTS Website. Retrieved June 18, 2011 from http://www.ielts.org.

Jilani, R. (2009). Problematizing high school certificate exam in Pakistan: a washback perspective. *The Reading Matrix, 9 (2),* 175-183.

Kayri, M. (2009). The multiple comparison (post-hoc) techniques to determine the difference between groups in researches. *Firat University Journal of Social Sciences, 19 (1),* 51-64.

Kerlinger, F. N. (1973). *Foundations of behavioral research.* New York: Holt, Rinehart and Winston.

Kiomrs, R., Andolmehdi, R., & Naser, R. (2011). On the interaction of test washback and teacher assessment literacy: the case of Iranian EFL secondary school teachers. *English Language Teaching, 4 (1),* 156-161.

Lewis, A. (2000). *High-Stakes Testing: Trends and Issues* (Report No. RJ96006101). Aurora: Mid-continent Research for Education and Learning.

Lewkowicz J., & Zawadowska-Kittel E. (2008). Impact of the new school-leaving exam of English in Poland. *Offprint from Research Notes, 34,* 27-31.

Lewthwaite, M. (2006). Teacher and student attitudes to IELTS writing tasks: positive or negative washback?. *UGRU Journal, 5,* 1-16.

Litwin, M. S. (1995). *How to measure survey reliability and validity*. California: Sage Publications, Inc.

Luxia, Q. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing, 22 (2),* 142-173. doi: 10.1191/0265532205lt300oa.

Luxia, Q. (2007). Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in china. *Assessment in Education, 14 (1),* 51-74. doi: 10.1080/09695940701272856.

Madsen, H. S. (1983). *Techniques in testing*. Oxford: Oxford University Press.

Messick, S. (1996). *Validity and washback in language testing* (Report No. ETS-RR-96-17). Princeton: Educational Testing Service.

MSN Encarta Dictionary (Version 2009) [Microsoft Corporation]. Redmond, WA: USA.

Nozawa, Y., Waltman, K., & Lai, E. R. (2007, April). *Discrepancies between test scores from high-stakes and low-stakes contexts: an investigation of the effects of school-level characteristics & test-preparation practices*. Paper presented at the annual meeting of the National Council for Measurement in Education, The University of Iowa, USA.

Pan, Y. (2008). A critical review of five language washback studies from 1995-2007: Methodological considerations. *Shiken: JALT Testing & Evaluation SIG Newsletter, 12 (2),* 2-16.

Pan, Y. (2009). A review of washback and its pedagogical implications. *VNU Journal of Science, Foreign Languages, 25,* 257-263.

Petterway A. L. (2006). A national perspective: a mixed-methods analysis of the impact of high stakes testing on English language learners in major urban high schools in Texas. *Shiken: JALT Testing & Evaluation SIG Newsletter. 1 (1),* 1-6.

Reynolds, J. (2010). *An exploratory study of TOEFL students as evaluators of 'washback to the learners'* (Master's thesis). Retrieved from Eric Database.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-stakes testing in higher education and employment: appraising the evidence for validity and fairness. *The American Psychological Association, 63 (4),* 215-227. doi: 10.1037/0003-066X.63.4.215.

Saif, S. (2006). Aiming for positive washback: a case study of international teaching assistants. *Language testing, 23 (1),* 1-34. doi: 10.1191/0265532206lt322oa

Salmani-Nodoushan, M. A. (2009). Measurement theory in language testing: past traditions and current trends. *I-manager's Journal on Educational Psychology, 3 (2),* 1-12.

Schulz, B. C. (2005). *Teachers' perspectives of how high-stakes testing influences instructional decisions and professionalism.* (Doctoral dissertation). Retrieved from University of Georgia Database.

Sevimli, S. E. (2007). *The washback effects of foreign language component of the university entrance examination on the teaching and learning context of English language groups in secondary education (a case study).* (Master's thesis). Retrieved from Gaziantep University Database.

Shih, C. (2009). How tests change teaching: a model for reference. *English Teaching: Practice and Critique, 8 (2),* 188-206.

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

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing, 18,* 373-391.

Shohamy, E. (2007). Language tests as language policy tools. *Assessment in Education, 14 (1),* 117-130.

Siong, L. K. (2004). Exploring the connection between the testing of reading and literacy: the case of the MUET. *GEMA Online Journal of Language Studies, 4 (1),* 1-10.

Solorzano, R. W. (2008). High stakes testing: issues, implications, and remedies for English language learners. *Review of Educational Research, 78 (2),* 260-329.

Suen, H. K., & Wu, Q. (2006). Psychometric paradox of very high-stakes assessment and solutions. *KJEP, 3 (1),* 113-129.

Suen, L. K., & Yu L. (2006). Chronic consequences of high-stakes testing? Lessons from the Chinese civil service exam. *Comparative Education Review, 50 (1),* 46-65.

Tarhan, M. (1992). *An experimental study on the effects of frequent testing.* (Master's thesis). Retrieved from YÖK Database.

Taylor, L. (2005). Washback and impact. *ELT Journal, 59 (2),* 154-155.

TOEFL Website. Retrieved June 18, 2011 from http://www.ets.org/toefl.

Turner, C. E. (2009). Examining washback in second language education contexts: a high stakes provincial exam and the teacher factor in classroom practice in Quebec secondary schools. *International Journal of Pedagogies and Learning, 5,* 103-123.

Washback. (n.d.) In *Cambridge Online Dictionary*. Retrieved June 15, 2011 from http://dictionary.cambridge.org/washback.

Washback. (n.d.) In *Oxford Online Dictionary*. Retrieved June 15, 2011 from http://oxforddictionaries.com/washback.

Watkins, D., Dahlin, B., & Ekholm, M. (2005). Awareness of the backwash effect of assessment: A phenomenographic study of the views of Hong Kong and Swedish lecturers. *Instructional Science, 33,* 283-309.

Wen-jun, W. (2010). Evaluation on washback effects of College English Test Band 4. *Sino-US English Teaching, 7 (2),* 35-40.

Yıldırım, O. (2010). Washback effects of a high-stakes university entrance exam: effects of the English section of the university entrance exam on future English language teachers in Turkey. *Asian EFL Journal, 12 (1),* 92-116.

APPENDICES

APPENDIX 1

QUESTIONNAIRE – ENGLISH VERSION

Dear Participant,

The aim of this questionnaire is to obtain your knowledge and views on the implementation of IBFLE. The data to be received in this way will only be used for scientific goals, and will utterly be kept confidential. You are kindly invited to share your views by choosing the most appropriate reply that best suits your viewpoint.

Thanks in advance for your invaluable cooperation.

Gender	( ) Male	( ) Female	
Age	( ) 21-30	( ) 31-40	( ) 41 or over
Academic title	( ) Professor	( ) Assc. Prof.	( ) Asst. Prof.
The Faculty / School	:_____		
No. of IBFLE attempts	:_____		
The highest score gained in IBFLE	:_____		
IBFLE language taken	( ) English	( ) German	( ) French
IBFLE module taken	( ) App.& Nat. Scs	( ) Health Scs	( ) Social Scs

	Yes	No
1. Did/Do you take private tuition while preparing for IBFLE?		
2. Did/Do you take private courses / classes while preparing for IBFLE?		
3. Did/Do you make use of the internet while preparing for IBFLE?		
4. Did/Do you receive overseas education while preparing for IBFLE?		
5. Do you think that the foreign language* education at your Bachelor's degree promoted your IBFLE success?		

*foreign language: your target language in IBFLE

	Very	Somehow	Not at all
6. The *vocabulary* questions in IBFLE are easy for me.			
7. The *grammar* questions in IBFLE are easy for me.			
8. The *translation* questions in IBFLE are easy for me.			
9. The *dialogue* questions in IBFLE are easy for me.			
10. The *reading comprehension* questions in IBFLE are easy for me.			

	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
11. I tried/am trying to improve my listening skills to do better in IBFLE.					
12. I tried/am trying to improve my speaking skills to do better in IBFLE.					
13. I tried/am trying to improve my reading skills to do better in IBFLE.					
14. I tried/am trying to improve my writing skills to do better in IBFLE.					
15. I tried/am trying to improve my vocabulary to do better in IBFLE.					
16. I tried/am trying to improve my grammar to do better in IBFLE.					
17. The level of challenge in IBFLE is approvable.					
18. The test duration for IBFLE is adequate for answering all the questions.					
19. IBFLE is a proper tool for assessing foreign language competence.					
20. The variety of languages in IBFLE should be enriched.					
21. IBFLE should be modified so as to cover all four skills (*i.e.* listening, speaking, reading, writing).					
22. IBFLE should be administered more frequently.					
23. IBFLE should be merged with SEFLE.					

THIS SECTION IS UNIQUE TO IBFLE-ACHIEVERS	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
24. Preparing for IBFLE helped me to learn the target language.					
25. I am using my IBFLE gains in my professional studies.					
26. I can easily speak the target language in the settings where foreign people partake.					
27. I can easily understand a person speaking in the target language in academic meetings.					
28. I can write academic articles using the target language.					
29. I can easily understand the academic papers in my field that are written in the target language.					

APPENDIX 2

QUESTIONNAIRE – TURKISH VERSION

Sayın Katılımcı,

Bu anketin amacı, ÜDS ile ilgili tecrübeleriniz ve düşünceleriniz hakkında bilgi almaktır. Anketten elde edilen veriler yalnızca bilimsel amaçla kullanılacak olup, üçüncü kişilerle kesinlikle paylaşılmayacaktır. Lütfen her cümleyi sonuna kadar okuyup her cümle için uygun gördüğünüz yalnızca bir seçeneği işaretleyiniz.

Göstermiş olduğunuz ilgiden dolayı teşekkürler.

Cinsiyetiniz	( ) Bay	( ) Bayan	
Yaşınız	( ) 21-30	( ) 31-40	( ) 41 ve üzeri
Ünvanınız	( ) Profesör	( ) Doçent	( ) Yardımcı Doçent
Çalıştığınız fakülte / bölüm	:_____		
ÜDS'ye kaç kez girdiniz?	:_____		
ÜDS'den aldığınız <u>en yüksek</u> not	:_____		
Başvurduğunuz dil	( ) İngilizce	( ) Almanca	( ) Fransızca
Başvurduğunuz modül	( ) Fen B.	( ) Sağlık B.	( ) Sosyal B.

	Evet	Hayır
1. ÜDS'ye hazırlanırken özel ders aldınız mı/alıyor musunuz?		
2. ÜDS'ye hazırlanırken dershaneye gittiniz mi/gidiyor musunuz?		
3. ÜDS'ye hazırlanırken internetten faydalandınız mı/faydalanıyor musunuz?		
4. ÜDS'de başarılı olmak için yurtdışında eğitim aldınız mı?		
5. Lisans eğitiminizde aldığınız yabancı dil* eğitimi, ÜDS başarınızı arttırdı mı?		

*yabancı dil: ÜDS'deki yabancı diliniz

	Çok	Orta	Hiç
6. ÜDS'deki *kelime* sorularını ne kolaylıkta çözersiniz?			
7. ÜDS'deki *dilbilgisi* sorularını ne kolaylıkta çözersiniz?			
8. ÜDS'deki *tercüme* sorularını ne kolaylıkta çözersiniz?			
9. ÜDS'deki *diyalog* sorularını ne kolaylıkta çözersiniz?			
10. ÜDS'deki *paragraf* sorularını ne kolaylıkta çözersiniz?			

	Kesinlikle Katılıyorum	Katılıyorum	Kararsızım	Katılmıyorum	Kesinlikle Katılmıyorum
11. ÜDS'deki başarımı arttırmak için *dinleme becerimi* geliştirmeye çalıştım/çalışırım.					
12. ÜDS'deki başarımı arttırmak için *konuşma becerimi* geliştirmeye çalıştım/çalışırım.					
13. ÜDS'deki başarımı arttırmak için *okuma becerimi* geliştirmeye çalıştım/çalışırım.					
14. ÜDS'deki başarımı arttırmak için *yazma becerimi* geliştirmeye çalıştım/çalışırım.					
15. ÜDS'deki başarımı arttırmak için *kelime bilgimi* geliştirmeye çalıştım/çalışırım.					
16. ÜDS'deki başarımı arttırmak için *dilbilgisini* geliştirmeye çalıştım/çalışırım.					
17. ÜDS'nin zorluk seviyesi uygundur.					
18. ÜDS için ayrılan süre, soruların tümünü yanıtlamaya yeterlidir.					
19. ÜDS, dil becerisini ölçmek için uygun bir araçtır.					
20. ÜDS'deki dil seçenekleri (*örn: İspanyolca, İtalyanca, Rusça, vb.*) çoğaltılmalıdır.					
21. ÜDS, dört beceriyi (*dinleme, konuşma, okuma, yazma*) kapsayacak şekilde değiştirilmelidir.					
22. ÜDS'nin uygulanma sıklığı arttırılmalıdır.					
23. ÜDS, KPDS ile birleştirilmelidir.					

BU BÖLÜMÜ ÜDS'DEN BAŞARILI OLANLAR CEVAPLAYACAKTIR.	Kesinlikle Katılıyorum	Katılıyorum	Kararsızım	Katılmıyorum	Kesinlikle Katılmıyorum
24. ÜDS'ye çalışmak yabancı dili öğrenmemi sağladı.					
25. ÜDS'den elde ettiğim kazanımları mesleki çalışmalarımda kullanmaktayım.					
26. Yabancıların olduğu akademik ortamlarda hedef dili kullanarak rahatça konuşabilirim.					
27. Bilimsel toplantılarda hedef dilde konuşan bir kişiyi rahatça anlayabilirim.					
28. Hedef dilde akademik makaleler yazabilirim.					
29. Hedef dilde kendi alanımda yazılmış bilimsel yazıları rahatlıkla anlayabilirim.					

APPENDIX 3

INTERVIEW QUESTIONS – ENGLISH VERSION

1. How important is foreign language in your academic life? In what circumstances (listening, speaking, reading, writing) do you use the foreign language?
2. How have you been preparing for IBFLE?
3. Do you agree that IBFLE fulfils its purpose?
4. Do you think IBFLE should be modified? In what ways?
5. Do you have any other remarks?

APPENDIX 4

INTERVIEW QUESTIONS – TURKISH VERSION

1. Yabancı dil, akademik hayatınızda sizin için önemli mi? Yabancı dili hangi durumlarda (dinleme, konuşma, okuma, yazma) kullanırsınız?
2. ÜDS'ye nasıl hazırlandınız?
3. ÜDS'nin amacına uygun olduğunu düşünüyor musunuz?
4. Sizce ÜDS'de değişiklik yapılmalı mı? Nasıl olmalı?
5. Bu konuda eklemek istediğiniz başka bir şey var mı?

ÖZET (SUMMARY)

1. GİRİŞ

Yabancı dil günümüzde akademik kariyer, kişisel gelişim, terfi ve prestij kazanmayı hedefleyen kişilerden talep edilen olmazsa olmaz bir kriter haline gelmiştir. Yabancı dil bilgisinin ölçülmesi İngilizce'de "high-stakes tests" olarak adlandırılan "yüksek riskli sınavlar" yoluyla ölçülmektedir. Bu sınavların Türkiye'deki örnekleri Kamu Personeli Yabancı Dil Sınavı (KPDS) ve Üniversitelerarası Kurul Yabancı Dil Sınavı (ÜDS)'dır. KPDS daha çok yabancı dil tazminatı almak isteyen devlet memurları tarafından tercih edilirken ÜDS temelde akademik kariyer hedefleyen kişilere hitap etmektedir. 2011 yılında çıkarılan düzenlemeden önce soru sayıları KPDS'de 100 ve ÜDS'de 80 iken, düzenlemeden sonra her ikisi için soru sayısı 80 olarak eşitlendi. KPDS, Mayıs ve Kasım aylarında, ÜDS ise Mart ve Ekim aylarında uygulanmaktadır. KPDS'de 20'den fazla dil seçeneği bulunurken ÜDS'deki dil sayısı 3 (İngilizce, Almanca, Fransızca) ile sınırlandırılmıştır. Tek versiyonu bulunan KPDS'nin aksine, ÜDS, Fen Bilimler, Sağlık Bilimleri ve Sosyal Bilimler olmak üzere 3 ayrı modülde uygulanmaktadır.

ÜDS, ülkenin geneline hitap etmesi yönüyle oldukça iyi bilinen bir sınav olmakla beraber zaman zaman eleştirilere maruz bırakılan bir sınavdır. Bu eleştirilerin en yaygın olanı sınavın etkin dil kullanımını değil de mekanik bilgi edinmeyi sağladığı ve dolayısıyla dinleme ve konuşma gibi akıcı dil kullanımını gerektiren becerileri ihmal ettirdiği konusundadır. Bu tür eleştirileri çalışma zemini olarak kabul eden bu çalışmanın hedefi ÜDS'nin meydana getirdiği bu eleştirilerin arkasında yatan sebepleri ortaya çıkarmak ve bu güçlüklerin ortadan kaldırılmasına yardımcı olacak çözümler sunmaktır. Bu amaçla ilgili olarak çalışmanın cevap aradığı sorular şunlardır:

1. ÜDS'de başarılı olmak ile cinsiyet, ÜDS'deki diller, ÜDS'deki modüller, özel ders alma, özel kurslara katılma, ÜDS amaçlı internet kullanımı, yurtdışında eğitim alma ve lisans öğrenimindeki yabancı dil eğitimi gibi faktörler arasında bir ilişki var mıdır?

2. ÜDS'de hangi dil becerilerine vurgu yapılmaktadır?

3. ÜDS'deki sınav düzeninin öğrencilerin çalışma şekli üzerinde bir etkisi var mıdır?

4. Sınava girenlere göre, ÜDS'de ne tür değişiklikler yapılmalıdır?

5. ÜDS hangi dil becerilerinin kazanılmasını sağlamaktadır?

2. YÖNTEM

Araştırmanın evreni olarak Dicle Üniversitesi'nde çalışan öğretim üyeleri (profesör, doçent, yardımcı doçent) seçilmiştir. Anket çalışmasının yapıldığı dönemde öğretim üyelerinin sayısı 575 (163 profesör, 127 doçent ve 285 yardımcı doçent) olarak bildirilmiştir. Rastgele seçim yoluyla 161 (% 28) öğretim üyesine ulaşılmış olup bunların 4'ü profesör (% 3), 42'si doçent (% 26) ve 115'i de (% 71) yardımcı doçentlerden oluşmuştur. Bunlardan 144'ü (% 89,4) ile bizzat görüşülmüş olup geriye kalan 17 (% 10,6) kişi ise çalışmaya e-mail yoluyla katılmıştır. Doçentlerin ve özellikle de profesörlerin katılımının az olmasının nedeni bu unvana sahip öğretim üyelerinin çoğunun ÜDS'den önce yürürlükte olan "Doçentlik Yabancı Dil Sınavı" veya KPDS'ye girmiş olmalarıdır.

Araştırmada, nicel ve nitel olmak üzere iki tür veri toplanmıştır. Nicel veriler, Arslan ve Elaldı tarafından uygulanan anketlerde değişiklikler ve eklemeler yapılarak hazırlanan 29 maddelik bir anket yardımıyla; nitel veriler ise gönüllü olan 10 öğretim üyesine uygulanan yüz yüze görüşmeler yoluyla elde edilmiştir. Nicel verilerin incelemesinde SPSS 17,0 yazılımının Windows için üretilmiş versiyonu kullanılarak ortalama, standart sapma, yüzdelik, frekans oranı ve güvenilirlik oranı konularında incelemeler yapıldı. Değişkenlerin analizinde, iki değişkenli durumlar için *t*-test kullanılırken ikiden fazla değişkeni olan durumlar için Tek Yönlü Varyans Analizi (One-Way ANOVA) kullanılmıştır. Değişkenler arasındaki anlamlı farkın hangi gruplar arasında olduğunu saptamak için post hoc testlerinden LSD testi kullanılmıştır. Nitel veriler ise içerik analizi ile incelenmiştir. Yazılı ve sesli görüşme kayıtları, araştırmacı tarafından İngilizce'ye çevrildikten sonra araştırmacının bu konuda uzman olan bir meslektaşı tarafından kontrol edilmiştir. Görüşme kayıtlarındaki veriler benzerliklerine göre gruplara ayrılarak incelenmiştir.

3. BULGULAR

Yapılan analizlerde, katılımcıların 134 (% 83,2) erkek ve 27 (% 16,8) kadından oluştuğu ve katılımcıların 4'ünün (% 2,5) 21-30 yaş arasında, 83'ünün (% 51,5) 31-40 yaş arasında ve 74'ünün (% 46,0) ise 41 yaş ve üzerinde olduğu görülmüştür. Ayrıca, en yüksek katılımın Eğitim Fakültesi'nden ve en düşük katılımın Meslek Yüksekokulu'ndan olduğu görülmüştür. ÜDS geçmişlerine bakıldığında, katılımcıların büyük kısmının sınava üç veya daha fazla kez girdiği, aldıkları notların çoğunluğunu 65-79 arasındaki notların oluşturduğu, sınav dili olarak en çok İngilizce'nin tercih edildiği ve katılımcıların çoğunun Fen Bilimleri modülünden sınava girdikleri görülmüştür.

Nicel analizlerde, kadınların erkeklere göre, İngilizce'yi tercih edenlerin diğer dilleri tercih edenlere göre; Sağlık Bilimleri modülünden sınava girenlerin diğer modüllerden sınava girenlere göre; özel ders almış olanların almamış olanlara göre ve lisans eğitiminde almış olduğu yabancı dil eğitiminin ÜDS başarısını arttırdığını düşünen kişilerin bu durumun tersini düşünen kişilere göre daha başarılı olduğu ortaya çıkmıştır. ÜDS'deki bölümler arasında en çok vurgulanan becerinin kelime bilgisi ilgili sorular olduğu ve en az vurgulananın ise tercüme soruları olduğu anlaşılmıştır. Kelime bilgisinin, ayrıca öğrencilerin ÜDS'ye hazırlanırken en çok yoğunlaştığı beceri olduğu görülmüştür. ÜDS'yi geçen ile geçemeyenlerin dil becerileri yönünden karşılaştırılmasında, sınavı geçenlerin sınavla doğrudan ilgisi olan okuma, yazma, kelime bilgisi ve dil bilgisi becerilerine daha çok yoğunlaştığı ve geçemeyenlerin ise sınavda test edilmeyen beceriler olan dinleme ve konuşmaya sınavı geçenlere göre daha fazla yoğunlaştığı kanıtlanmıştır. Modüller bazında yapılan karşılaştırmada ise Fen Bilimleri grubunun bütün becerilere daha fazla yoğunlaştığı görülmüştür.

ÜDS'nin uygunluğu konusundaki nicel ve nitel verilerde, ÜDS'nin genel anlamda dil becerilerini ölçmekte yetersiz olduğu ve sınavın dört dil becerisini (dinleme, konuşma, okuma, yazma) ölçecek şekilde değişmesi gerektiği konusunda görüşler ön plana çıkmıştır. Ayrıca, sınavı geçen kişilerin geçmeyenlere göre bu fikirleri daha çok desteklediği görülmüştür.

Araştırmanın "washback" ile ilgili olan son kısmında ise, ÜDS'nin en çok okuma becerisinin geliştirilmesine yardımcı olduğu anlaşılmıştır.

4. SONUÇ ve TARTIŞMA

Bu çalışmada temel olarak ÜDS başarısına etki eden faktörler, ÜDS'nin hangi dil becerilerine vurgu yaptığı, ÜDS'deki sınav düzeninin öğrencilerin çalışma şekillerini nasıl etkilediği, ÜDS'de yapılması gereken değişiklikler ve ÜDS'nin yabancı dil öğrenme sürecindeki etkisi araştırılmıştır. Araştırmadan elde edilen sonuçlar, sınava girecek olan yeni adaylar ve sınavın yöneticileri konumundaki yetkililere yönelik olmak üzere iki farklı kategoride düzenlenmiştir. Buna göre, yeni adaylara yönelik çıkarımlar şunlardır:

- Özel ders alma, ÜDS'ye hazırlanma yöntemleri arasında en etkili yöntemdir.
- ÜDS'ye etki eden faktörler arasında, lisans düzeyinde alınan yabancı dil eğitim ÜDS başarısına olumlu yönde etki etmektedir.
- ÜDS'de en çok vurgulanan (dolayısıyla en zorlayıcı olan) bölüm, kelime bilgisini ölçen sorulardır.
- Okuma, yazma, kelime bilgisi ve dilbilgisi becerilerine yoğunlaşmak ÜDS başarısını arttırırken, konuşma ve dinleme becerilerine çalışmanın ÜDS başarısı üzerinde olumlu bir etkisi bulunmamaktadır.
- Okuma becerisi, ÜDS çalışmalarında en iyi geliştirilebilen dil becerisidir.

Yöneticilere hitaben elde edilen sonuçlar şunlardır:

- ÜDS'ye girenlerin büyük bir kısmı, ÜDS'nin dil becerilerini ölçme konusunda yetersiz olduğunu ve ÜDS'nin dört dil becerisini kapsayacak şekilde değiştirilmesi gerektiğini ifade etmektedirler.
- ÜDS'deki modüller arasında Sosyal Bilimler modülünün ÜDS'deki başarı konusunda diğer iki gruba göre çok düşük seviyede olduğu tespit edilmiştir. Bu durum muhtemelen Fen Bilimleri ile Sağlık Bilimleri modüllerinde sınava giren kişilerin modüllerindeki sorulara Sosyal Bilimler grubundakilere göre daha aşina olmalarından kaynaklanmaktadır.