Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education) 41: 398-410 [2011]

A STUDY ON THE ENGLISH LANGUAGE TEACHERS' PREPARATION OF TESTS

İNGİLİZCE ÖĞRETMENLERİNİN SINAV HAZIRLAMALARINA İLİŞKİN BİR ÇALIŞMA

Arif SARIÇOBAN*

ABSTRACT: In this article the researcher has examined the current situation in test (a) construction: designing, structuring, developing, (b) administering, and (c) assessing the foreign language tests to see if we are still at the same point (traditional) and has given some suggestions on this indispensable issue. To collect the necessary data the 4th year students doing their practicum studies at a state high school in Ankara under the supervision of the researcher are asked to collect one sample of each test (written or oral form) their mentors have been using to assess their foreign language students. The common characteristics of the test samples are scrutinized in terms of validity and reliability, language skills and areas including spelling, contextualization, time, typing, students' foreign language level (simple or complex structures), instructions, and backwash effect. Relying on the findings of the study some recommendations have been made for foreign language teachers.

Keywords: Testing, designing, administering, assessing, foreign language education.

ÖZET: Bu makalede araştırmacı, sınav hazırlama ve uygulamada (a) yapılandırma: tasarım, yapı, geliştirme, (b) uygulama ve (c) yabancı dil sınavlarını değerlendirme konularını incelemiş ve bazı önerilerde bulunmuştur. Ankara ilinde devlet lisesinde staj yapan ve araştırmacının danışmanlığını yaptığı son sınıf stajyer öğrencilerden kendi rehber öğretmenlerinin öğrencilerini değerlendirmek üzere kullandıkları sınavlardan birer kopya (yazılı ya da sözlü) getirmeleri istenmiştir. Sınav örneklerinin ortak özellikleri geçerlilik ve güvenirlilik, dil becerileri ve hece, bağlam, zaman, yazım, öğrencilerin dil yeterlilikleri (basit ya da karmaşık yapıların kullanılması), yönergeler ve yabancı dil öğrenim etkinliklerinin yansıması gibi dil alanlarından oluşan konular incelenmiş ve tartışılmıştır. Çalışmanın bulgularından hareketle yabancı dil öğretmenlerine konuya ilişkin önerilerde bulunulmuştur.

Anahtar sözcükler*:* Sınav, tasarım, uygulama, değerlendirme, yabancı dil eğitimi

1. INTRODUCTION

For decades, testing has been a neglected area in foreign language teaching (FLT) not only in our country but also other countries in that foreign language (FL) tests lack the outcomes of the language learning process. Messick (1996) points out that "… in the case of language testing, the assessment should include authentic and direct samples of the communicative behaviors of listening, speaking, reading and writing of the language being learnt. Ideally, the move from learning experiences to test exercises should be seamless. As a consequence, for optimal positive washback there should be little if any difference between activities involved in learning the language and activities involved in preparing for the test" (p. 241-242). However, in reality, foreign language tests usually seem to focus on recognition rather than production skills of FL learners. This problem still exists in our context. This assertion was once upon a time approved as I recall the first two years of my undergraduate study in the department of English Language Teaching (ELT) where we used to be given pen-paper tests in our "Speaking" course midterms. It is still odd as it was years ago. For instance, in "speaking" courses at the ELT departments students are still asked to perform in a pen-paper test for their midterms; whereas, they should be tested orally. Dialogue completion tests and/or discussion type of tests on a topic (in written form for speaking) usually favored by the teacher are fashionable for use today. Of course, the lecturers of this speaking course seem to claim that they do not have enough time to administer a speaking test in their midterms since the classes are overcrowded. This is the same case in respect to the state high schools in our country.

On the other hand, written foreign language tests seem to lack such important issues as validity, reliability, washback effect, language skills and areas including spelling, contextualization, time, typing, students' foreign language proficiency level (simple or complex structures), and instructions.

* Assoc. Prof. Dr., Hacettepe University, Faculty of Education, Department of Foreign Language Education, arifs@hacettepe.edu.tr.

Validity

The test must test what it is intended to test. In other words, test items must be representative of what we intend to test (Köksal, 2004). In short "the validity of a test is the extent to which it measures what it is supposed to measure and nothing else" (Heaton, 1990, p.159). Heaton states that there are four types of validity: (1) *Face Validity*: The test looks right to the other testers, teachers, moderators, and testees, (2) *Content Validity*: The test should contain a representative sample of what is learnt. For example, "if we are interested in the acquisition of relative clauses in general and plan to represent learners with an acceptability judgment task, we need to make sure that all relative clause types are included" (Mackey & Gass, 2005, p. 107), (3) *Construct Validity* : Bachman and Palmer (1996) define construct validity as a term "used to refer to the extent to which we can interpret a given test score as an indicator of the ability(ies), or construct(s), we want to measure" (p. 21), (4) *Empirical (Statistical) Validity*: We compare the results of the test with the results of some criterion measure such as: (*a) Concurrent Validity: (1)*an existing test known or believed to be valid and given at the same time, or (2)the teacher's ratings or any other such form of independent assessment given at the same time, or (*b) Predictive Validity:(*1) the subsequent performance of the testees on a certain task measured by some valid tests, or (2) the teacher's ratings or any other such form of independent assessment given later.

Reliability

Reliability is the consistency of the measurement or the degree to which an instrument measures the same way each time it is used under the same condition with the same subjects. That is, a test is considered reliable if we get the same result after administering it twice to the same subject group.

Reliability of a test can be determined both by estimating the *rater reliability* and *instrument reliability*. Rater reliability can be done either by *interrater reliability* which refers to "a measure of whether two or more raters judge the same set of data in the same way" (Mackey & Gass, 2005, p. 129) or by *intrarater reliability* which means that the rater judge the data the same at different times.

Moreover, there are two ways through which instrument reliability can be estimated: test/retest and internal consistency.

a. Test/Retest

We should get the same score on Test 1 as we do on Test 2 on different occasions without any language work between these two occasions. The test instrument is implemented at two separate times to the same subjects. Then, "in order to arrive at a score by which reliability can be established, one determines the correlation coefficient between the two test administrations" (Mackey & Gass, 2005, p. 129).

This type of reliability differs from *mark/re-mark reliability* in the sense that the latter indicates the marking of the same test papers is done by either two or more different testers or by the same tester on different occasions and we still get the same grades or marks.

b. Internal Consistency

Internal consistency estimates reliability by grouping questions in a questionnaire that measure the same concept. For example, you could write two sets of three questions that measure the same concept (say class participation) and after collecting the responses, run a correlation between those two groups of three questions to determine if your instrument is reliably measuring that concept. Split-half, Kuder-Richardson 20 and 21, and Cronbach's are some of the statistical methods to determine reliability.

Quite naturally, there are some factors that might affect the reliability of a test (Heaton, 1990:162). These are: *(1) The Size*: The larger the sample, the greater the reliability, (*2) The Administration*: Is the same test administered to different groups under different conditions or at different times? (*3) Test Instructions*: Are the test instruction simple and clear enough? (*4)Personal Factors* : Motivation, illness, etc., (*5) Scoring the test:*Subjective or objective?

Washback

As is known, washback effect and the effect of tests has been a concern for researchers in the field of education and particularly for those in foreign language education (e.g., Alderson & Hamp-Lyons, 1996; Alderson & Wall, 1993; Andrews, 1995; Bailey, 1996; Chapman & Synder, 2000;

Cheng, 1998, 1999; Cheng, Watanabe, & Curtis, 2004; Haladyna, Nolen, & Haas, 1991; Hamp-Lyons, 1997; Hughes, 1988; Li, 1990; Maddaus, 1988; Messick, 1996; Shohamy, 2001; Shohamy, Donitsa, & Ferman, 1996; Wall, 1997; Watanabe, 1996). Washback is defined by Hughes (1989) as "… the effect of testing on teaching and learning" (p.1). Basically, the tasks that our students are expected to perform in classroom activities must be in line with the tasks they are asked to in tests, which means that they must be familiar with the tasks and the techniques that the foreign language teachers use to assess the learners' language skills (Köksal, 2004:3).

Language Skills and Areas

As we all know language skills are four: listening, speaking, reading, and writing, whereas among the language areas are pronunciation, vocabulary, grammar, and translation. These language skills and language areas are tested either discretely or integratively regarding the goal of the language learning studies (Harmer, 2001). However, whether it is discrete or integrative, we should seek the fact if the tests that will be administered focus on recognition or production skills.

Contextualization

It is placing the target language in a realistic setting, so as to be meaningful to the student (see for example Harmer, 2001). Heaton (1990:52) suggests that the test should include contextually relevant items, i.e. words related to the context but different in meaning to the key word in the sentence (single or ticket in "How much is a ____ to Tokyo, Please). "Single" is correct; however, "ticket" is contextually relevant.

Time

The total time allocated for the test should be stated on the test paper. Additionally, how much time is ideal for a specific task on the test should also be stated for students to complete in the allocated time? Hand writing is a disadvantage for students to be able to read and understand the test items. Therefore, typing on the computer and checking spelling are all important for a healthy test construction.

Typing

Typing of testing materials is highly important because such a method puts down the real appearance of items to test. Thus, the testees can easily see and comprehend the standardized forms of both spoken and written utterances.

Students' Foreign Language Proficiency Level

The language use in the test must be appropriate to the FL proficiency level of the learners. Otherwise, the difficulty level will be higher than learners' comprehension level. Therefore, the language level of learners should be taken into account when constructing the test.

Instructions

Instructions must be simple and clear enough for learners to comprehend what they are required to do in each specific task in the test. They should not get confused by the instructions.

This study depicts the possible problematic cases in testing. To do so, two cases will be examined in detail. Case 1 is Midterm Exam I, whereas Case 2 is the Common Exam. These two exams were administered after the first two units of the main course book (*Breeze 9*). Surprisingly, the English language teachers do not use any oral exam.

2. METHOD

This current descriptive research examines the current situation in testing (a) construction: designing, structuring, and developing, (b) administering, and (c) assessing the foreign language tests to see if we are still at the same point (traditional) and will try to give some suggestions on this indispensable issue. To collect the necessary data the researcher has used "document analysis technique" (Bodgan and Biklen, 1998) usually preferred in qualitative research methodology. "Document analysis is the systematic examination of instructional documents such as syllabi, assignments, lecture notes, and course evaluation results in order to identify instructional needs and challenges and describe an instructional activity. The focus of the analysis should be a critical examination, rather than a mere description of the documents" (Instructional Assessment Resources, 2010). The teacher trainees (n=16) that perform their teaching practice at a state high school are asked to collect the samples of two written exams their mentors (n=5, of whom two are males and the rest

are females) have been using to assess their foreign language students. The English language teachers voluntarily submitted the samples of the test papers. For the document analysis the common characteristics of the written test samples (for common/joint and midterm) have been scrutinized by the researcher in terms of the above mentioned features of testing in 14 stages such as *(1) validity, (2) reliability, (3) backwash effect,(4)language skills and areas including spelling, (5) contextualization, (6)time, (7) typing, (8) language proficiency (simple or complex structures), (9) instructions*, *(10)motivation, (11) scoring, (12) spelling, (13) diagnostic testing* and *(14) homework*.

3. FINDINGS

As to the analysis of the testing materials used in the above mentioned high schools the researcher has examined the two exams (one Midterm and one Common/Joint). Midterm 1 consists of eight sections. In the first section there are five questions to tests *Imperatives*. Students are asked to fill in the blanks with the right *Imperatives* according to the given context. In the second part, a short reading passage is given to the students to answer True/False Questions. The aim of these questions is to test understanding of the reading passage. In Part 3 students are asked to fill in the blanks with "*some, any, a* or *an*," whereas in Part 4 they are asked to fill in the blanks with "*do, does, don't,* or *doesn't*." Part 5 includes the use of "*Simple Present Tense*." Part 6 requires students to rewrite the given sentences by adding the *adverbs of frequency*. In Part 7 they are asked to respond to the *some questions about their own lives*. Lastly, in Part 8 they are asked to complete the sentences by using "*There is, There are, Is there* or *Are there*."

3.1. Validity

The common/joint exam consists of two sections. The reading passage in the first section is intended to test the reading comprehension level of the students with multiple choice questions. On the other hand, the second section of the exam aims to test both vocabulary and grammar which include questions on *Polite Requests, WH-questions, Present simple, Present continuous* and *Comparatives*.

Face Validity

Having examined both of the exams the exam items seem sound except for some minor facts. They are planned and designed by the two English language teachers at this school. These exams can be applied to the other classes which are at the same level as well. Therefore, it can be said that these exams seem not to violate face validity.

Content Validity

We can say that what is asked in the exam is almost parallel with what is taught in the first two units. In addition to this there aren't any unknown words that are not covered during the exam. In the common/joint exam reading, vocabulary and grammar skills are tested; however, writing and listening skills which are enclosed in the course book are not included in this exam.

In the first exam, the grammar topics of the first Unit of the course book are tested. The exam topics are; *Imperatives, Quantifiers, Frequency Adverbs, Simple Present* and a *Short Reading Passage*. However, Quantifiers which include "some and any" are not covered in the first two units in the course book but are tested.

As a rule, exams should include all the topics studied. In Case 2 teachers just present *Present Simple* and *Present Continues* but ignore the topics of Unit 2 such as *Future Tense* and *Should-Must*. However, the exam includes the topics of the third and the fourth units such as *polite requests* and *comparatives*. As a result, it can be said that these two exams seem to lack content validity to a certain extent.

Constructive Validity

The teacher mainly uses grammar translation method. Although the course book is student-centered which integrates all the four skills, the teacher instructs only reading, grammar, and vocabulary. Therefore, the productive skills such as writing and speaking are not tested in the exams stated above. As is asserted by Köksal (2004), "teachers should test learners' writing skills by having them write and their speaking skills by having them speak" (p.4). Therefore, these two exams are said to lack constructive validity to some extent. As a result it can be asserted that the two exams include

recognition type of questions but not production. Only Section 6 in the first exam includes "Rewrite" skills which require students to rewrite the sentences with the adverbs in brackets.

Empirical Validity

In the first midterm exam the highest mark obtained is 98, while the lowest mark obtained is 19. In the common/joint exam the highest mark is 100 and the lowest mark is 8. Another striking point is that in the first midterm exam 21 students out of 33 got 50/100 and in the common exam only 15 of them achieved this. However, only four students (675; 686; 705 and 765) seem to increase their results. By looking at these marks it may be said that both of these tests produce similar results and thus they seem not to lack empirical validity.

Table 1

9/A First and Common Exam Results

Number of Students	Midterm 1	Common Exam
525	28	20
560	37	40
596	86	56
664	59	40
665	38	44
668	25	08
675	87	92
676	34	36
677	85	72
683	39	28
685	98	84
686	96	100
692	96	84
700	61	40
702	40	44
704	64	56
705	86	92
711	45	44
715	85	72
722	45	24
724	59	48
736	19	36
739	67	52
749	67	64
755	47	48
762	50	40
765	79	80
711	56	48
775	88	68
785	72	60
788	42	48
801	57	44
795	64	53

3.2. Reliability

As can be seen in Table 2, 33 students attended both of the exams and the highest score in the first exam is 98 and the lowest is 19. The average score of the students in this exam is 61. When we look at the common/joint exam, the highest score is 100 and the lowest is 8. So the average score (53) in the common exam is lower than the first exam. According to this result, it can be claimed that the test is reliable when it is evaluated for 9-A. Overall, the total average of the exams is 57 which is a

moderate score. As we all know an ideal test should be both reliable and valid. Depending on these results we can easily claim that these two tests have consistency and reliability. Therefore, the findings concerning the reliability and validity of the above exams once again support the fact that a test can be reliable but not valid.

Table 2:*The Scores in Class: 9/A*

	Midterm 1	Common/ Joint Exam
Number of students	**33**	**33**
Highest score	**98**	**100**
Lowest score	**19**	**8**
Average score	**61**	**53**

3.3. Washback Effect

The teacher frequently hands out worksheets to the students. Her students pay great importance to these worksheets because they know that the questions in the worksheets will be parallel to the exam. As the students are familiar with the type of questions by the help of the worksheets their anxiety level is decreased. So this is a positive effect for language learning. In addition, the students are provided with some survival/daily communicative tasks in the classroom activities. However, it has been observed that in the first exam there is no place given for communicative language skills since only grammar and reading are tested.

3.4. Language Skills and Areas

When we examine the book we see that *Breeze 9* intends to develop four skills interactively. Each skill is given attention but developing speaking skill is placed in the centre of this course book. Writing is another productive skill to be considered in the book. Listening and reading are both receptive skills and the book encourages students to actively receive and process information. Grammar teaching is acquired through skills and activities. Moreover, the vocabulary teaching is one of the other strengths of this book. New words are introduced in a meaningful context and students are encouraged to guess the meaning of the words through this context. Students are introduced with the skills, language areas and structures which they are supposed to acquire through out the Unit. Each unit ends with a "How much do I know?" section which contains statements to self-evaluate how much students have acquired.

What the teacher applies in the class is to consider reading and writing as major skills to be developed. Although *Breeze 9* contains a CD for developing listening skills because of the lack of the opportunities and of the limited time listening is not paid importance to at all. While looking at language skills grammar is in the centre of learning in foreign language courses but vocabulary is taught when it is needed. However, these two tests seem to lack this issue since speaking and listening skills are not included.

3.5. Contextualization

When we observe the questions in the exams, we can easily urge that all the questions have a context which may be beneficial for the students to understand and find the answer easily. For example, in the third section of the first exam the questions are arranged with regards to dialogues so that students can understand the context easily.

3.6. Time

Even though the allocated time is not stated in the exam papers, the teacher expresses it orally. Generally the allocated time for an exam is 40 minutes. This time is enough for the students to answer all the questions.

3.7. Typing

Two exams are typed on the computer. So, students can easily see, read and understand the questions.

3.8. Language Proficiency

Since the English language proficiency of the students is not appropriate to be instructed in English, the students are occasionally instructed in Turkish which is their native language. However, in the exams the instructions are given in English. Despite this fact, the students seem to comprehend what they are asked to do in the exam. It may be speculated that the teachers might have taught what those instruction require them to do in their lessons.

3.9. Instructions

Both in the first and second exam the instructions are pretty clear. For example, in Section A in the second exam, there is no room for misunderstanding, confusion or ambiguity in regard to the following instructions: (a) Read the text and choose the best alternative, (b) Choose the best alternative, (c) Complete with "Some, any, a or an", and (d) Answer the following questions

3.10. Motivation

As is indicated before the first exam consists of eight sections on grammar and reading comprehension, while in the common/joint exam there are only two sections; the first section includes a long reading passage while the second section only consists of multiple types. However, the both tests do not have separate sections clearly stated such as "Grammar" and/or "Reading" (See Appendix 1). The first exam starts with "Imperatives" (Grammar) in the first section which is followed by "True/False" (Reading). Then, the rest four is about grammar again (Section 3: Some, any, a, or an; Section 4:do, does, don't or doesn't; Section 5: the Simple Present; Section 6:Adverbs). Next is about "General Questions" with Wh- and How-. Lastly, Section 8 ends with again grammar questions on "There is/There are."

The common exam has only two sections: one is a reading passage with multiple choice questions. The other is "Choose the best alternative" type multiple questions which include grammar points, dialogue completion (Item 20). It is generally suggested to start the exam with grammar and vocabulary and proceed with reading comprehension in order not to cause any anxiety problem (from easy to more complex questions). As is seen the structure of the type used in these two exams that is explained above may increase student anxiety and thus may demotivate them.

3.11. Scoring

As we couldn't get the exam papers, we cannot say if evaluation is fair and not biased. However, as far as we have observed in the classroom, the teacher is not biased towards any students. Everyone is equally treated.

3.12. Spelling and Punctuation

There aren't any problems in spelling, punctuation and pronouns with the help of auto correction in computers. The written exams are flawless. In addition to these two written exams, teachers use oral exams and homework to evaluate their students.

3.13. Diagnostic Testing

Both exams have same discrete points. Especially in the first exam we can easily see that there is a section only testing "Simple Present". Further, there are reading parts to test only reading comprehension level in both exams. Regarding questions, it can be said that all the questions are discrete as they test *only one thing at a time*, except for only second section in the common exam. Therefore, it can be easily remarked that the tests are diagnostic. For example, the teacher covered the grammar point "Some/any" in the courses and accordingly she tested these points to check if they can

use them appropriately. Thus, it can be said that the teacher evaluate the weak points of the students in this subject by using a diagnostic part in the exam.

3.14. Homework

The teacher generally gives worksheets as homework to students which are parallel with the subject instructed in the class. Moreover, some parts of the course book are also given as homework when the allocated time isn't enough. The course book *Breeze 9* has a workbook and this is given to reinforce what they have learned in the course. Although some parts of workbook are given as homework this is not done frequently.

3.15. Oral Exams

In practice, the teacher doesn't evaluate students by using oral exams. Although speaking is a vital part of language learning it is not used as an exam type. However, instead of marking speaking, the teacher gives marks according to the worksheets, homework and participation in the lesson by giving plus and minus signs. She considers these marks as oral marks when evaluating their total score.

4. DISCUSSION

At the end of this study it has been observed that there are some issues to be taken into consideration when designing language exams (test). The exam papers collected for the purpose of this study share the following points that need attention. Although quantifiers "*some* and *any*" are not covered in the first two units in the course book, they are tested. Writing and listening skills which are enclosed in the course book are not included in this exam. As a rule, exams should include all the topics studied. In Case 2 teachers just present *The Present Simple* and *Present Continuous* but ignore the topics of Unit 2 such as *Future Tense* and *Should-Must*. However, the exam includes the topics of the third and the fourth units such as *polite requests* and *comparatives*. Then, we should include the topics studied and neglect any.

The effects of the test format on teaching and learning practices are documented (Cheng, 1997; Frederiksen, 1984; Shohamy et al., 1996) and discussed (Bailey, 1996; Madaus, 1988; Messick, 1996) in previous studies. In this study, although the course book is student- centered which integrates all the four skills, the teacher instructs only reading, grammar, and vocabulary. Therefore, the productive skills such as writing and speaking are not tested in the exams stated above. The first exam includes some questions in a dialogue form for grammar items (Section 3: some, any, a, or an; Section 4: do, does, don't, or doesn't) and so is the case in the common exam (Section 2: Item 20). Of course, as we all know "dialogues are part of daily communication" (Köksal, 2004:4). However, in this question type only grammar is tested but not speaking which include speakers' ideas or feelings. We should test our learners' writing skills by having them write and their speaking skills by having them speak.

The students are provided with some survival/daily communicative tasks in the classroom activities. However, it has been observed that in the first exam there is no place for communicative language skills since only grammar and reading are tested. It is more meaningful when we have communicative tasks in tests by considering the type of instructional activities our students are exposed to in the classroom. Otherwise, in our country where English is not the medium of communication how will our students have the chance to communicate in English? While looking at language skills grammar is in the centre of learning in foreign language courses but vocabulary is taught when it is needed.

The structure used in these two exams may increase student anxiety and thus demotivates since section headings are not clearly stated and some sections include questions for different language areas. In recent years it seems that the quality of language instruction has not increased to the desired level and neither does testing. There still are some points that need to be paid attention indicated above. Instruction (teaching) and testing should overlap. We should test what we have taught and not ignore any point. The test must be representative of what has been studied in the class. "Needless to say, it is hard to separate teaching and testing from each other as they are so interrelated" (Kabadayı, 1998:50). The test sections should be clearly indicated such as "Grammar" and/or "Reading" not to

cause any motivational and anxiety problems. The tests should include communicative tasks for daily communication since they are practiced in class.

5. IMPLICATIONS AND CONCLUSION

Teachers' beliefs about effective teaching methods and their background are documented as reasons for how English is taught in classrooms (Alderson and Hamp-Lyons, 1996; Cheng, 1999; Watanabe, 1996) but their experience and competence in testing English has not been thoroughly mentioned before in literature although it is a fact that most of the teachers of English all over the world are not native speakers of the language they are teaching.

In centralized education systems it is common practice to impose tests to trigger and facilitate innovation in teaching and learning practices as "in comparison to introducing reforms through teacher training, development of new curricula or new textbooks, changing the test is a substantially cheaper venture" (Shohamy, 2001, p. 40). Thus, it is important to accept the fact that more has to be done rather than only imposing or changing the test to achieve the intended washback effect. Previous research shows that it is crucial to take teachers' beliefs and background into consideration and emphasizes the importance of teacher training in order to facilitate intended reforms in teaching and learning (Alderson and Hamp-Lyons, 1996; Andrews, 1995; Cheng, 1999; Watanabe, 1996).

It is also important to bring all the stakeholders (policy makers, test constructors, test- and text-book publishers, teachers, students, and parents) together, to examine practices and needs, and to make intended washback transparent, expressed explicitly and communicated to all so that responsibility to accomplish reform in teaching and learning can be realized (Chapman and Snyder Jr, 2000; Hughes, 1988; Shohamy, 2001; Watanabe, 1996).

As asserted by Qi (2005) and other scholars as well (Bailey, 1996; Frederiksen, 1984; Madaus, 1988; Messick, 1996; Shohamy et al., 1996), the format of the test may undermine the intended washback. The multiple-choice format may be preferred as it is more reliable and easier to score especially in large-scale testing conditions. However, it is emphasized that they fail to assess higher order cognitive skills (Frederiksen, 1984) and cause drilling of strategies (Madaus, 1988). Thus, another important suggestion could be changing the format of the test if the intend is to reform ELT practices in classrooms. Using the foreign language for communicative purposes in academic settings is a growing need at universities worldwide (Hughes, 1988; Zareva, 2005) which cannot be reached with tests in only multiple choice format at high schools. This need for a change in format has also been realized by the administrators of one of the most well-known test in the world, the Test of English as a Foreign Language (TOEFL). With the new TOEFL-iBT, the test administrators proposed an integrated approach to teaching, learning and assessing all four skills not only through multiple choice questions but also with speaking and integrated writing tasks. Although TOEFL-iBT is very new and thus no research has been done to reveal whether intended washback effects have been realized in language classrooms, it still stands as a promising example for the future of large-scale language testing.

It is also important to include other means of assessment (see Shohamy, 2001), such as portfolios, in order to come up with a valid picture of what a student can do rather than putting "weight on a single test" (Linn, 2000). This also emphasizes the importance of *learner autonomy* (see also Bailey, 1996) in foreign language learning which has been recognized by the Council of Europe. A language portfolio is proposed by the Council of Europe in the Common European Framework of Reference for Languages (CEFRL) in 2000. CEFRL is a starting point and basic reference which is "open and flexible so that it can be used in different situations with necessary adaptations" (Kohonen, 2001) across Europe. It is a valuable source and an indispensable guideline for learners, teachers, curriculum developers, and policy makers as Little (2006) states that CEFRL "is offered as a basis for sustained international cooperation in the development of language education policy, the construction of language curricula, the implementation of language learning and teaching, and the assessment of language learning outcomes" (p. 169). Although there are some concerns and questions on how CEFRL can be successfully applied in real practice of assessment due to limited feedback received from new experiences from different European countries (Eckes et al., 2005; Figueras et al., 2005; Weir, 2005), its value is still highly appreciated.

The descriptors in the CEFRL highly promote *lifelong language learning* and *self-assessment* which is central to the development of the European Language Portfolio (ELP). ELP was developed and piloted by the Modern Languages Division of the Council of Europe, Strasbourg, from 1998 to 2000. During the European Year of Languages 2001, ELP was introduced as a tool to promote plurilingualism and pluriculturalism, *life long learning* and *learner autonomy*. The portfolio enables the language learners to keep record of their language learning and cultural experiences either at school or outside school. The 'can do' descriptors of the CEFRL is fundamental to ELP as without them the language learner would not be able to keep track of his/her own progress in a detailed and constructed way as it is possible with the CEFRL.

The ELP has three parts: (1)*Language Passport:* provides an overview of the individual's proficiency in different languages. The proficiency is defined in terms of skills and levels as described in the CEFRL. It includes information about partial or specific competence of language/s, intercultural experiences, and assessments (self-assessment, teacher assessment and/or assessment by educational institutions or examinations boards), (2) *Language Biography:* encourages the language learner to record his/her personal development of language/s, his/her learning process, what he/she can do in each language, linguistic and cultural experiences gained in and/or outside formal educational contexts, and (3) *Dossier:* enables the language learner to choose materials to report and to document achievements.

The introduction of the ELP is recommended by the Council of Europe to the Ministry of Education of all member states. It has been the part of the CEFRL which is most widely adapted so far as Little (2006) suggests that "the Council of Europe's ELP website lists 75 accredited models from 26 member states and three international non-governmental organizations" (p. 182). The recognition and application of such a portfolio may replace tests so that a more rational and reasonable way of assessing foreign language learners' success could be possible.

Moreover, to enable foreign language teachers acquire and develop their testing skills in-service teacher training courses should be designed and offered in their institutions since a considerable number of foreign language teachers are graduates of Linguistics, English Language and Literature, Translation and Interpretations and so on.

REFERENCES

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

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

Andrews, S. (1995). Washback or washout? The relationship between examination reform and curriculum innovation. In D. Nunan, V. Berry & R. Berry (Eds.) *Bringing about Change in Language Education. (pp. 67-81).* Hong Kong: University of Hong Kong.

Bachman, L.F. & Palmer, A.S. (1996). Test Usefulness: Qualities of Language Tests. In Bachman, L.F. and Palmer, A.S. *Language Testing in Practice: Designing and Developing Useful Language Tests*. (pp.17-42). Oxford, New York: Oxford University Press.

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

Bodgan, R., & S. K. Biklen. (1998). *Qualitative Research for Education: An Introduction to Theory and Methods*. Boston: Allyn and Bacon.

Chapman, D.W. & Synder, C,W, Jr. (2000). Can high-stakes testing improve instruction: Reexamining conventional wisdom. *International Journal of Educational Development 20*, 457-474.

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

Cheng, L. (1998). Impact of a public examination change on students' perceptions and attitudes toward their English learning. *Studies in Educational Evaluation, 24,* 3, 279-301.

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

Cheng, L., Watanabe, Y., & Curtis, A. (2004). *Washback in Language Testing: Research Contexts and Methods.* Mahwah, NJ: Lawrence Earlbaum.

Eckes T., Ellis, M., Kalnberzina, V., Pizorn, K., Springer, C., Szollás, K., Tsagari, C., et al. (2005). Progress and problems in reforming public language examinations in Europe: cameos from the Baltic States, Greece, Hungary, Poland, Slovenia, France and Germany. *Language Testing, 22,* 3, 355-377.

Figueras, N., North, B., Takala, S., Verhelst, N., & Van Avermaet, P. (2005). Relating examinations to the Common European Framework: A Manual. *Language Testing, 22,* 3, 261-279.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39,* 3, 193-202.

Haladyna,T.M., Nolen, S.B., & Haas, N.S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20,* 5, 2-7.

Hamp-Lyons,L. (1997). Washback, impact, validity: Ethical concerns. *Language Testing 14*, 3, 295-303.

Harmer, J. (2001). *The Practice of English Language Teaching.* Essex: Pearson Education Limited.

Heaton, J. B. (1990). *Writing English Language Tests (New Edition)*. Longman Group UK Limited.

Hughes, A. (1988). Introducing a needs based test of English language proficiency into an English medium university in Turkey. In A. Hughes (Ed.) *Testing English for university study.* (pp. 134-153) London: Modern English Publications.

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

Instructional Assessment Resources. (2010). Assess Teaching: Document Analyses. Retrieved March 28, 2011, from http://www.utexas.edu/academic/ctl/assessment/iar/teaching/plan/method/doc-analysis.php.

Kabadayı, A. (1998). Testing the Tests: A Case Study on the Determination of the Learners' Performance by Conducting C-Test and Cloze in In-class Procedure. *Dil Dergisi*, *66,* 50-57.

Kohonen, V. (2001). Developing the European Language Portfolio as a pedagogical tool for advancing student autonomy. In L. Karlsson, F. Kjisik and J. Nordlund (Eds.), *All together now: Papers from the Nordic Conference on autonomous language learning*. (pp. 20-44). Helsinki: University of Helsinki Language Centre.

Köksal, D. (2004). Assessing teachers' testing skills in ELT and enhancing their professional development in testing and assessment through distance learning on the NET. *Turkish Online Journal of Distance Education-TOJDE*, *5,* 1.

Li, X. (1990.) How powerful can a language test be? The MET in China. *Journal of Multilingual and Multicultural Development, 11,* 393-404.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29,* 2, 4-16.

Little, D. (2006). The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact. *Language Teaching, 39*, 167-190.

Mackey, A. & Gass, S. M. (2005). *Second Language Research: Methodology and Design.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Maddaus, G.W. (1988).The influence of testing on curriculum. In L.N. Tanner (Ed.) *Critical issues in curriculum: Eighty-seventh yearbook of the National Society for the Study of Education* (pp. 83-121). Chicago: University of Chicago Press.

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

Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing, 22,* 2, 142-173.

Shohamy, E. (2001) *The Power of Tests: A Critical Perspective on the use of Language Tests.* Harlow, England: Pearson Education.

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

Wall, D. (1997). Impact and washback in language testing. In C.Clapman and D. Dorson, (*Eds.) Encyclopedia of language and education. Vol. 7. Language testing and assessment* (pp. 291-301). Dordrecht: Kluwer Academic.

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

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing, 22,* 3, 281-300.

Zareva, A. (2005). What is new in the new TOEFL-iBT 2006 test format? *Electronic Journal of Foreign Language Teaching, 2,* 2, 45-57.

Genişletilmiş Özet

Son yıllarda ülkemizde yabancı dil öğretimi alanında epeyce bir ilerlemenin olduğu bir gerçektir. Avrupa Dilleri Eğitimi Ortak Çerçeve Programı'na (CEF) ilişkin olarak bu konuda çok çaba sarf edilmiştir. Avrupa Dilleri Eğitimi Ortak Çerçeve Programı "Avrupa'da yabancı dil öğrenimi ve öğretimine ortak temel oluşturmakta ve bugün Avrupa'da dil eğitimiyle ilgilenen büyük çoğunluk tarafından tanınmaktadır. Öğretmenler; müfredat, sınav ve materyal geliştirenler için uygulanabilir bir araç olup öğretim meselelerinde/metodolojide plan yapan ve karar verenlere genel anlamda rehberlik etmektedir.

Bildiğimiz üzere, yabancı dil öğretmenleri iletişimde akıcılığı ve doğru kullanımı sağlamak için iletişimsel etkinlikler oluşturmağa, geliştirmeye ve var olanları uygulamaya çalışmaktadırlar. Tabii ki, bu sınıf içi etkinliklerde veya herhangi biçimdeki ölçme araçlarında mesajın sağlıklı bir şekilde gönderilip alınması için iletişim stratejileri geliştirmek demektir. Öğrencilerimizin gelişimini mi ya da başarısını mı ölçmeliyiz? Yarışmalar arasından seçim, seviye/sıralama kıyaslaması, öğretim metotları üzerinde sınav/değerlendirme, sınav geliştirme izleği, profesyoneller üzerinde baskı vb. işlemler olmalı mı? İlk bakışta cevabımız "EVET" olursa, o zaman sınavlarda nelere bakmalıyız? Öğrencilerimizin gelişimlerini ve başarılarını değerlendirmek için kullanmış olduğumuz sınavların zayıf güvenirliği, zayıf geçerliliği, olumsuz geri etkisi (backwash) (ölçmenin öğretim ve öğrenme üzerindeki etkisi) varmış gibi ve uygulanamaz gibi görünmesi yabancı dil öğretmenleri arasında yaygın bir düşüncedir.

Bu makalede, araştırmacı sınav (a) oluşturmada (tasarımlamak, yapısını oluşturmak, geliştirmek), (b) uygulamada ve (c) yabancı dil sınavını hala aynı noktada olup olmadığımızı görmek için değerlendirmedeki son durumu incelemiş ve bu konu ile ilgili bazı önerilerde bulunmuştur. Gerekli veriyi toplamak için, öğretmenlik uygulamalarını gerçekleştirmek için devlet okullarına giden hizmet öncesi öğretmenlerden yazılı ve sözlü olmak üzere danışmanlarının yabancı dil öğrencilerini değerlendirmek için kullanmış olduğu en az iki sınav örneği toplamaları istenmiştir. Sınav örneklerinin ortak özellikleri geçerlilik, güvenilirlik, dil becerileri ve yazım, bağlam içerisinde sunmak, zaman, bilgisayarda yazım, öğrencilerin yabancı dil seviyesi (basit ve karmaşık yapılar), yönergeler ve geri etkisini içeren konular açısından incelenmiştir.

Belirtilen her iki sınavın (ilk ve ortak sınav) dikkatli incelenmesi bu sınavların görünüş geçerliliğini ihlal etmediği fakat kapsam geçerliliğe ve yapısal geçerliliğe bir dereceye kadar sahip olmadığını göstermiştir. Diğer bir taraftan, sınav sonuçlarından elde edilen bulgulara dayanarak bu iki sınavın tutarlı ve güvenilir olduğunu kolayca savunabiliriz ki öğrencilerin ortalama notları birbirine oldukça yakındır. Geri etkisine gelince, ilk sınavda gramer ve okuma ölçüldüğünden iletişimsel dil becerilerine yer verilmediği gözlemlenmiştir. Buna dayanarak, sınavların geri etkisi içermediğini savunabiliriz. Dil becerileri ve alanları açısından bakarsak, öğrencilere ilk iki ünite boyunca öğrenmeleri gereken dil becerileri, alanları ve yapıları sunulmuştur. Her bir ünite öğrencilerin ne kadar öğrendiklerine ilişkin kendilerini değerlendirmeleri için verilen ifadelerden oluşan "ne kadar biliyorum?" bölümüyle bitmektedir. Okuma ve yazmanın geliştirmek için asıl beceriler olduğunu düşünen öğretmen ona göre uygulamalar yapmaktadır. "Breeze 9" kitabının dinleme becerilerini geliştirmek için CD'si olmasına rağmen, yetersiz fırsat ve zamandan dolayı dinlemeye yer ayrılamamaktadır. Yabacı dil derslerinde, gramer öğrenimin merkezinde yer almakta fakat kelime sadece ihtiyaç duyulduğunda öğretilmektedir. Bununla birlikte, bu iki test bu konuda yetersizdir çünkü konuşma ve dinleme becerileri dâhil edilmemiştir. Ayrılması gereken zaman sınav kâğıtlarında belirtilmemiş, öğretmen tarafında sözlü olarak bildirilmiştir.

Sınavın yapısına gelince, her iki testte de "gramer" ve "okuma" gibi açıkça belirtilen ayrı bölümler bulunmamaktadır. İlk sınav ilk bölümde verilen emir kipleriyle (gramer) başlamakta ve doğru/yanlış (okuma) aktivitesiyle devam etmektedir. Daha sonraki dört kısım da gramer ile ilgilidir

(Bölüm 3: some, any, a, an; Bölüm 4: do, don't or doesn't; Bölüm 5: Geniş zaman; Bölüm 6: Zarf). Bir sonraki bölüm ise "Wh- ve How-" soru kelimelerini kullanan genel sorular ile ilgilidir. Son olarak, bölüm 8 "there is/there are" kalıbıyla oluşturulan gramer sorularıyla bitmektedir. Diğer bir taraftan, ortak sınav sadece iki bölümden oluşmaktadır: bir bölümde çoktan seçmeli sorularıyla bir okuma metni bulunmaktadır; diğer bölümde ise gramer noktalarını, diyalog tamamlamayı
(Madde 20) içeren çoktan seçmeli soru türünden "en iyi seçeneği seçiniz" soru türü bulunmaktadır. Herhangi bir kaygı sorunu oluşturmamak için sınava önce gramer ve sözcük bilgisinden başlayıp daha sonra okumayı anlamayla (kolaydan daha zor sorulara doğru) devam etmek genel olarak önerilir. Ancak, bu iki sınavda kullanılmış olan bu türde soru yapıları öğrenci kaygısını arttırabilir ve dolayısıyla onların şevkini kırabilir.

Her iki sınavda da bazı ayrık nitelikli sınama maddeleri bulunmaktadır. Özellikle ilk sınavda sadece geniş zamanı ölçen bir bölüm olduğunu kolaylıkla görebiliriz. Ayrıca, her iki sınavda da sadece okumayı anlama seviyelerini ölçen okuma bölümleri mevcuttur. Soruları göz önünde bulundurunca, ortak sınavın ikinci bölümü hariç, bütün soruların bir seferde bir şeyi ölçtüğü için ayrık nitelikte olduğu söylenebilir. Dolayısıyla, sınavların tanılayıcı olduğu kolaylıkla ifade edilebilir.

Son olarak, ders kitabı dört dil becerisini de entegre eden öğrenci merkezli bir ders kitabı olmasına rağmen, öğretmen sadece okuma, gramer ve kelime öğretimi üzerinde durmaktadır. Bu nedenle, yazma ve konuşma gibi üretimsel dil becerileri yukarıda bahsedilen sınavlarda ölçülmemiştir. Ortak sınavda (Bölüm 2: madde 20) olduğu gibi, ilk sınav gramer noktalarını ölçmek için diyalog biçiminde bazı sorular içermektedir (Bölüm 3: some, any, a, an; Bölüm 4: do, does, don't, doesn't). Hepimizin bildiği gibi, diyaloglar günlük iletişimin bir parçasıdır. Fakat bu soru türünde sadece gramer ölçülmüş ve konuşan kişilerin fikir ve hislerini içeren konuşma ölçülmemiştir. Öğrencilerimizin yazma becerilerini onlara yazı yazdırarak, konuşma becerilerini ise onları konuşturarak ölçmeliyiz.

Sınıf içi etkinliklerde öğrencilere bazı günlük iletişim becerilerini gerektiren görevler verilmektedir. Ancak, ilk sınavda yalnız gramer ve okuma ölçüldüğünden iletişimsel dil becerilerine hiç yer verilmediği görülmüştür. Öğrencilere sınıf içinde verilen öğretim etkinliklerin türü göz önünde bulundurularak sınavlarda iletişim becerilerini gerektiren maddeler hazırlamak daha anlamlı olur. Yoksa, İngilizcenin iletişim dili olmadığı ülkemizde öğrencilerimiz iletişim kurmak için İngilizce kullanma fırsatını nasıl yakalayacaklar? Dil becerilerine bakıldığında, gramer yabacı dil derslerinde öğrenmenin merkezinde yer alırken kelime ihtiyaç duyulduğunda öğretilmektedir.

Son yıllarda, ne dil öğretme kalitesi ne de ölçme istenilen seviyeye çıkmamıştır. Yukarıda belirtilen bazı noktalara dikkat edilmesi gerekmektedir. Öğretme ve ölçme iç içe olmalıdır. Öğrettiklerimizi ölçmeli ve diğer noktaları da göz ardı etmemeliyiz. Sınav sınıf içinde öğrenilenleri temsil edici durumda olmalıdır. Tabii ki, öğrenme ve ölçmeyi birbirlerine çok bağlı oldukları için ayırmak çok zordur.