Pazhuhesh-e Zabanha-ye Khareji, No. 20, Special Issue, English, 2005, pp 81-100

The Effect of Coaching on TOEFL Type and Task Based Tests*

Hossein Farhady

Iran University of Science and Technology

e-mail: hfarhady@iust.ac.ir

Abstract

TOEFL is widely used as a certificating device, and is strongly claimed accountable by the people who utilize it. However, there are indications of the vulnerability of TOEFL to the teaching of test taking strategies. The purpose of the present study is to provide empirical evidence that coaching toward TOEFL may invalidate its results as indicators of test-takers' proficiency levels. To test the hypothesis, the scores of 75 subjects on a TOEFL and on a task-based test of language proficiency were compared. The subjects were selected from among those who attended coaching classes for TOEFL. The findings of this study revealed that coaching leads to spurious scores and consequently to unaccountability of TOEFL results. Theoretically, it is claimed that teaching toward task-based tests would not be harmful to pedagogy and learning, because performance on such tests are manifestation of the performance of test takers on real life tasks that they will be expected to perform. This contends that testing practitioners should move toward the use of more 'authentic' and performance-based assessments to cause the least possible damage to teaching, learning, and decision-making processes. Also the implications of this study concerning the safety of coaching toward a particular task, domain, or subdomain in order to enhance students' achievement on the one hand, and test validity on the other, are discussed.

Key Words: task-based test, performance-based test, language proficeincy, psychometric methods, washback.

* This is the revised version of the paper presented at the LTRC 1999, Tsukuba, Japan.

1. Introduction

The English language has become a lingua franca all over the globe due to political, economical, and technological reasons. People are transacting their academic or occupational businesses mostly through satellite networks via English. Hence, all around the world, people try to obtain knowledge of English for their intended purposes. Also, most of the universities admit candidates who meet the requirements of academic language ability. Therefore, to achieve an acceptable awareness of the candidates' knowledge of English, the use of proficiency estimators is growing faster and getting broader in scope than ever.

In spite of this speedy expansion, there are certain pressing problems such as the nature of language proficiency, the content of the tests, and the techniques of measurement which have made the process of language proficiency vague in theory and shaky in practice. I do not intend to review the literature on these issues because much has been lucidly written on them (Alderson, 1981, 1991; Bachman, 1990; Skehan, 1989). However, to put the issue in an ppropriate context, a brief mention of some of them seems necessary.

First, it is indisputable that we have to understand, as clearly and unambiguously as possible, what we intend to measure (Alderson, 1996; Bachman, 2000). If our purpose is to measure language proficiency, a definition, and probably an operational definition, of the construct is the least we can expect. Once satisfied with simplistic linguistically oriented explanations of the construct, now we are challenged by the complex multidimensional models of the construct. In spite of such progress in theoretical dimension, the operational and practical advancements seem to fall well behind. Of course, this is not to downgrade theoretical developments, but to plea for achieving a match between theory and practice (Farhady, 2000).

Second, it is also important that we find a reasonable way to measure what we want to measure. Again, once satisfied with discrete items of language elements and components, we now face the challenge of complicated performance based assessment. Influenced by the developments in educational measurement, the field of language assessment has witnessed a shift from testing culture, which aims at measuring limited and fixed properties of the individuals through a one-shot-case procedure, to an assessment culture, which focuses on measuring the learners' ability through multiple sources of information. Assessment is believed to provide valid evidence of achievement which would facilitate provision of further learning, or certify that a required level has been reached (Izard, 1998). Those who assess have to ensure that the assessment is as comprehensive as possible to meet accountability requirements. These requirements include fairness to test takers so that they are rewarded in accord with their genuine knowledge and skill, trustworthy certification to reassure the community, and informative evidence to those who will use the outcome of assessment as a part of the selection process for entry to later stages of education.

Izard (1998) points out that in assessment procedures, the choice of what to assess, the strategies of assessment, and modes of reporting are of prime concern and depend upon the audiences needing the information assessment provides. If the abilities to be assessed are not determined, the strategies of assessment are not clarified, and the ways the information are reported are not appropriate, audiences will have a distorted perspective of the conclusions.

This implies that quality assurance seems to be an essential requirement of language assessment procedures. Quality assurance in educational testing relates to both small scale and large scale educational testing. However, while the concerns may be common to large and small scale educational testing, the strategies for assuring the quality of both the process and the products may differ to a marked extent, i.e. quality assurance measures that are practical with small candidatures may not be practical with large candidatures. If some of the quality assurance requirements (e.g. the comparability, dependability, and relevance) are not met, it may well be concluded by some audiences that the assessment process is not fair, failing to give credit where credit was due, or providing a greater reward than some deserved.

Thus, in order for language measurement to be accountable for what it claims, to be defensible against its consequences, and to be comprehensive in what it intends to measure, the field has tended to move away from one single shot case testing practice to a multiple facet assessment procedure. It is an open question, then, to investigate whether the widely known standard proficiency tests have taken such a shift into consideration.

Third, assuming that we know what we are measuring and we know how to measure the construct, an important step is analyzing and interpreting the scores. Again, once satisfied with the somewhat simplistic classical test theory, now we are practicing complex and sophisticated statistical techniques such as G theory and IR theory. Despite the fact that such advances have improved our ability to manipulate the data, not much has been achieved in the meaningful interpretation of scores.

Fourth, the revival of ethical considerations in language testing has introduced new dilemmas. It is believed that many government, public, and private organizations have set codes of ethics for ethical obligations which include professional competency, integrity, honesty, confidentiality, public safety, and fairness, all of which are intended to preserve and safeguard public confidence. Unfortunately, we observe some cases of unethical behavior in the profession of language testing in spite of the recognition of the issue and recommendations for implementing ethical codes (Shohamy, 1990, 1996; Davies, 1997; Hamp-Lyons, 1996).

In spite of the above-mentioned problems, one point cannot be left unnoticed. That is, no matter how unclear our definition of the language construct may be, how imprecise our measurement of this vaguely defined construct may be, how simplistic our application of sophisticated techniques to imprecisely obtained scores of vaguely defined construct may be, and how unethical our decisions on such an outcome may be, we, language testers, must develop and use proficiency tests. And needy people must take such tests. Then it seems quite reasonable for people whose career, educational progress, professional promotion, and simply their lives, can be influenced by such tests to coach and be coached when such tests are to be beaten.

A glance at the status of proficiency tests around the globe indicates that certain tests such as TOEFL, MELAB, IELTS, and CAEL are among the most commonly used tests in the world, among which TOEFL is unquestionably the most popular and IELTS second in rank. On the other hand, from the above-mentioned tests, TOEFL and Michigan can be categorized as traditional multiple choice based tests and others as more performance based tests. The impression is that scores on TOEFL type tests are boosted when the subjects have gone through a coaching process, i.e., have mastered test-taking strategies special for such tests. Therefore, the choice of TOEFL and IELTS was based purely on the basis of their popularity and availability. The positions taken here do not indicate, by any means, any particular bias either for or against these tests.

The purpose of this study, then, is to investigate the dependability and validity of TOEFL results when the candidates have undergone a TOEFL preparation course. That is, if the candidates who obtain high scores after going through preparation classes are not able to score at the same level on a more performance-based language proficiency test, TOEFL certificates may not be claimed to have consequential validity. In order to put the issue in an appropriate perspective, the relationship, similarities, and differences between TOEFL type competence based tests and those of performance based tests will be discussed first. Then the concept of coaching with reference to EFL in the context of third world countries and in relation to the consequential validity of TOEFL will be presented. And finally, the research conducted to test the hypothesis will be reported on, the findings will be presented, and the implications and applications will be discussed.

2. TOEFL vs. IELTS

TOEFL was first developed in 1963 when a national council on "The testing of English as a foreign language" was set up to oversee its development. A number of very famous testing scholars have been and some of them still are in one way or another involved in policy making and in the test construction procedure of the test. It has been jointly administered since 1965 by the Educatıonal Testing Service and the College Board. TOEFL is administered according to policies determined by a 15-member policy council (Spolsky, 1995).

TOEFL is at present one of the largest public examinations in the world. It has enjoyed a steady increase in the number of candidates applying to take the test, from 50,000 in 1968-69 to 741,000 in 1990-1991 (ETS, 1991-2000;

Spolsky, 1995), approaching a million by now. It is administered at more than 1,300 test centers in 170 countries (ETS, 1991-2000). TOEFL is used by over 2,300 universities in the United States and Canada as well as UK and Australia to determine whether prospective foreign graduates and undergraduates have attained a proficiency level which would enable them to perform educational tasks in an English speaking environment. In addition to this, many US government organizations and private employers accept a TOEFL score as an indicator of an employee's ability to use the English language.

TOEFL is arguably the most research-undergone of all foreign language tests. There are a good number of academic papers not sponsored by its examining bodies which have investigated different aspects of TOEFL and its 'add on' Speaking and Writing sections. The research projects conducted on TOEFL vary both in the nature of questions and in the sophistication of analyses. That is, from questions as simple and straightforward as the relationships among different sections or part-whole correlations of TOEFL utilizing simple correlational analyses to complicated topics such as automated response selection through IRT models, all have been thoroughly explored (Henning, 1992, 1993; Turner, 1992; Angoff, 1988; Perkins, et al. 1988; to name just a few).

Despite extensive research on TOEFL, there does not seem much of implementation of the findings to have taken place. For years, TOEFL has been used with almost similar format, expected content, and fixed quantity. The only observable change has been the inclusion of the vocabulary section in the reading comprehension part rather than having it as a separate section.

IELTS (International Language Testing System), on the other hand, is at present jointly developed and administered by three separate bodies: the

British Council, the University of Cambridge Local Examinations Syndicate (UCLES), and the International Development Program of Australian universities and colleges (IDPA). This cooperation was intended to prevent any perception of Euronotic bias in the development and the use of the test. Although not as intensively researched as TOEFL, there have been four major published studies conducted on IELTS (Criper & Davies, 1988; Hughes, et al. 1988; Alderson & Clapham, 1996). The first two studies led to the ELTS being updated in 1989 to its present format. The British was first proposed in late 1970s by the British Council which asked UCLES to develop a test of English for Special Purposes suitable for foreign students seeking to study in British universities. It was in fact meant to be a replacement for its predecessor, the English Proficiency Test Battery (EPTB or 'Davies Test'), which was based on multiple choice and cloze test type items. ELTS, as Spolsky points out, was meant to be based on Munby's (1978) 'notional functional syllabus', originating from Wilkins (1976), expanding on Hymes' (1972) and later Canale's (1988) models of communicative competence. In 1986-1987, ELTS was taken by almost 14000 applicants in 150 testing centers in the world and was accepted by all the British universities and various professional institutions in Australia and Canada.

However, ELTS was criticized not only for the deficiencies pertained to its underlying model but also for the lack of validation. Skehan (1984) emphasizes that although Munby's model might have had some positive effect on syllabus design in the sense that it includes all the language use related skills advocated in 1970s, it has serious problems for testing. He argues that the sampling of the skills to be accommodated in the test does not have any well-established criteria, as it is not clear how the skills are related to one another and how relatively important each of them is (Farhady, 1983, 2000; Stansfield,1986).

Following Criper and Davies' (1988) validation study which magnified some of the weak points of the test, the British Council and UCLES, together with the International Development Program of Australian universities (IDPA), set up a revision project directed by Alderson and Clapham. The project led to the development of the International ELTS, which according to Alderson and Wall (1993: 12), "probably reflected the best of current teaching and thinking", i.e., the communicative theories widely accepted in applied linguistics. Influenced by research into second language reading conducted by Clapham (1997) and Wall and Claphem, 1994) along with the findings of research recommended by the International Editing Committee, the last modifications were made in IELTS in 1995. The new test had only one Academic Reading Module and one Academic Writing Module instead of three subject specific subtests of the previous versions. At present, the test is released in the two modules of General and Academic, with the same format but different contents geared to the purposes and needs of the candidates for each module.

IELTS as a major international public examination is taken by more than 25,000 candidates each year. There are 210 test centers in 105 countries. The test is accepted for undergraduate or postgraduate entry by Australian and British universities, colleges, and professional and technical institutions.

In addition to differences in appearance, history, and research perspectives between TOEFL and IELTS, there are certain fundamental theoretical differences as well. Theoretically, TOEFL has its origin in the psychometric approach to language testing (Spolsky, 1995). It employs limited response, i.e., closed type and objective test techniques such as multiple choice items. Besides, it greatly emphasizes the statistical features of the test. These two features strongly appealed to language testers of late 1950s who were searching for more principled and scientific methods of evaluating language ability instead of the ones already in use such as written compositions, translation from and into the target language, and so on.

The psychometric methods were seen as having four main advantages to offer to language testing (Ingram, 1991; Spolsky, 1995). Firstly, they would simplify the administration process, as they could be scored reliably with untrained personnel or machines. Secondly, it was believed that these kinds of test items could be used to test separate features of language proficiency in great detail without interference from other features. Thirdly, because candidates could respond very quickly to multiple choice or gap-filling questions, it was possible to include more items in the amount of time available. Fourthly, the test writer could in most cases move the items around on test paper without much affecting the other items.

IELTS, on the other hand, is developed on the basis of new approaches to language teaching and language testing. It may be claimed that IELTS is more content based, task oriented, and authentic than TOEFL. The tasks in IELTS are closer to real life situations than those in TOEFL. A clear example is the listening comprehension parts of the tests. While TOEFL uses meticulous, articulate, and idealistic language, IELTS utilizes real language in real context in this part.

3. Coaching in EFL Context

The idea of test-wiseness (TW) is not new in educational measurement. As early as 1966, Millman, Bishop, and Eble defined TW as “a subject’s capacity to utilize the characteristics and formats of the test or testing situation to receive a high score. p.707" Research indicates that TW leads to higher scores on both achievement and standardized tests. Diamond and Evans (1972) demonstrated that association between stem and alternatives, specific determiners, longest correct alternatives, grammatical clues, and overlapping distractors are five commonly used TW techniques.

Following Millman, et al. (1966) elaborate on TW and claim that a testwise person is able to:

select the option which resembles an aspect of the stem;

eliminate options which are known to be incorrect and choose from among the remaining options;

eliminate similar options, i.e., options which imply the correctness of each other; and eliminate those options which include specific determiners, e.g., *always, never.*

In order to utilize TW categories, test takers appeal to certain test taking strategies. A fairly elaborate list of test taking strategies is offered by Sam Houston University Counseling Center (1995). As Amer (1993) believes, learning test taking strategies cannot happen incidentally; rather, it requires organized and explicit instruction.

For many people, skills in test taking strategies lead to higher scores regardless of the content of the test (Sarnacki, et al. 1979; Bailey, et.al, 1988; Bailey, 1996). Carter (1986), however, makes a distinction between ability in utilizing test taking strategies, i.e., TW, and coaching. According to Bond, TW is independent of knowledge of the subject matter and is applied across a range of content areas. Test coaching, on the other hand, refers to training the test takers within a specified field. **Adams (1992) uses the acronym SCORER for the most important test taking strategies, where S refers to schedule your time, C to clue words, O to omit difficult questions, R**

to read carefully, E to estimate your answer, and R to review you work.

Of course, the point here is not to discuss the psychological or cognitive nature of TW. Nor is it intended to elaborate on the differences between coaching and TW. The main point is that some scholars believe that coaching leads to higher test scores, and thus positive washback. Others believe that coaching would lead to a negative washback. The findings of this study would shed some light on the issue.

4. Method

4.1. Participants

Seventy-five Iranian male and female students at three different language institutes participated in this study. Two of the institutes were private organizations in Iran, namely, Iran Language Institute, and Simin Educational Association, each contributing 52 and 13 participants, respectively. The rest were from Tehran University's paid language classes. The age range of the participants was between 18 and 25. Most participants were planning to take the TOEFL in near future. So, it can be safely assumed that they were quite motivated to learn the materials taught in the preparation classes.

4.2. Instrumentation

Three instruments were used for data collection purpose. First, a mock TOEFL was given as a pretest to check the students' entry command of English. The purpose of this test was to channel the students into pre-TOEFL or TOEFL courses. Second, an original version of the 1995 TOEFL released by the ETS was used as a posttest. The difference between pre and posttest scores is taken as the gain in test scores due to instruction. Third, a specimen IELTS, released by the IELTS, was given to the students one week before the posttest. This test was to serve as a criterion for a more performance based test.

4.3. Procedures

The procedures regarding instruction here were reported by the teacher of every class and approved by the directors of respective organizations. As a general rule, students participating in these classes are given one of the versions of mock TOEFL. Based on their scores, students are divided into two groups called “pre-TOEFL" and “TOEFL”. Students in the pre-TOEFL class receive instructional materials on fundamentals of the English language. The TOEFL group, however, receive instruction on three distinct but related areas of language including structure, vocabulary and reading, and listening comprehension. Sometimes, vocabulary and reading are taught separately. The materials usually include different kinds of TOEFL preparation books. The emphasis is on test taking strategies regarding TOEFL. They receive between 180 to 200 hours of instruction proportionately allocated to the skills being taught.

It might be interesting to note that instruction in these classes focuses on a) explaining the rules of usage and coaching students how to get the correct scores even with limited command of the language, b) memorizing as many words as students can even with the meanings in their mother tongue, and c) reading as many passages as possible and trying to apply some of the reading techniques to find the answers. In sum, the method of teaching can be said to be pre-scientific intuitive, and grammar translation at best.

5. Analysis and Results

As a pilot research project, the data were analyzed using simple T-tests and correlation statistics. Descriptive values are presented in Table 1.

Table 1. Descriptive Statistics for Study Measures

Variable	Mean	SD	Max. Points
TOEFL Pre	426	48	Scaled
TOEFL Post	499	52	Scaled
LCT	49	4	Scaled
STT	47	9	Scaled
RCT	43	8	Scaled
IELTS	35	17	
LCI	16	4	39
RCI	14	6	40

LC: Listening Comprehension; RC: Reading Comprehension; ST: Structure

It is obvious from Table 1 that students gained a reasonable increase in their TOEFL scores. The mean score of 499 is very close to the required score to be obtained by the applicants. This implies that subjects benefit from coaching classes on TOEFL and take advantage of the instruction. This should also imply that the subjects have achieved enough proficiency in language to be able to pursue their education in a university in the US, Canada, UK, or Australia where the medium of instruction is English. Furthermore, it implies that students have developed abilities to meet the academic requirements of the universities they might be attending.

The argument is raised here regarding the validity, particularly the consequential validity of the TOEFL scores. If the applicants are assumed to be proficient enough, they should be able to perform reasonably well on tests which deal with academic type of activities. As mentioned above, IELTS is such a test. However, the subjects performed well below the score that is equivalent to their TOEFL scores on the IELTS test. Most of the students had a difficult time with the IELTS, while most of them were happy with TOEFL. Second, due to the students' unfamiliarity with IELTS, most of them did not get to the writing section of the test. That's why the scores of writing section are not taken into account in the analysis.

Table 2 presents the t values obtained from comparing the total and component scores of TOEFL and those of IELTS. In all cases of TOEFL, there is a significant difference between the pre and posttest scores, which seems quite logical. However, the subjects performed significantly higher on TOEFL test components than they did on IELTS components. This has a significant implication. That is, TOEFL scores are not a valid indication of students' performance on task based tests. Considering the fact that TOEFL type tests are no longer considered as acceptable means of determining communicative ability of the test takers, it seems that the decisions made on the basis of TOEFL scores for certification or admission purposes may not be as valid as they were thought to be.

Table 2. T-values for Paired Comparisons

Comparison	N	t-observed
1. TOEFL PRE vs. TOEFL POST	56	18
2. TOEFL POST vs. IELTS	56	31
4. LCT vs. LCI	56	8
5. RCT vs. RCI	56	12

However, correlation coefficients among the subtets of both tests reveal another story. As Table 3 demonstrates, there are significant correlations among the component scores as well as total scores of the TOEFL, while the correlations among the subtests of IELTS are not satisfactory either within or across the subtests. This might imply that the difference in the scores may be due to the differences in the form as well as the construct these tests attempt to assess.

Table 3. Correlation Matrix of the Study Measures

Variables	LCT	RCT	STT	TPost	LCI	RCI	IELTS
LCT	1.00	.72	.90	.90	.76	.45	.44
RCT	1.00	.84	.47	.43	.48	.56	
STT	1.00	.97	.73	.34	.41		
TPost	1.00	.41	.39	.27			
LCI	1.00	.47	.48				
RCI	1.00	.86					
IELTS	1.00						

6. Conclusions and Suggestions

The purpose of this pilot study was to investigate the relationship between the scores of the participants on TOEFL and IELTS type tests after participating in coaching classes. The findings showed a trend in the susceptibility of the TOEFL type tests to overestimate the applicants' performance due to some construct irrelevent factors. This may have undesirable consequences on the applicants' academic career. It might also lead to unfair decisions about the applicants in terms of certification for different purposes. Although the findings are quite tentative and no firm conclusion can be drawn, some suggestions might be helpful.

First, preparation for high stake tests is inevitable. Test takers who feel that a test score would influence their lives in one way or another would and should try any approach that would help them. Therefore, one could assume that coaching is an indispensible part of the test taking process. However, the quality of coaching programs should be investigated. Obtaining high scores on a test without having corresponding knowledge associated with a given score would either underestimate or overestimate the test takers' real ability in using language in communication. This would, in turn, put the consequential validity of test scores under question. Therefore, there should be a close correspondence between the score obtained from a test and its value in terms of the intended performace.

Second, there is nothing wrong with coaching programs that prepare test takers, in any way possible, to obtain high scores in a given test. In fact, this is the very purpose of coaching programs. Then, the problem of high score with low ability level can be attributed to the quality of the test. That is, if a test is designed in such a way that coaching classes can boost the scores without preparing test takers for the intended purpose, the test suffers from validity problem. In other words, it is the test that should not lend itself to spurious preparations.

Third, tests like TOEFL that operate on a fixed content and lends themselves to strategy application, rule memorization, and noncommunicative performance, would be most vulnerable to coaching programs. That is why the TOEFL 2000 project has been in progress and will be implemented in 2005. The new TOEFL would hopefully avoid construct irrelevant factors to influence test takers' performance. This implies that the test would yield more valid and accountable scores.

Last but not least, there is a tendency among test takers to learn and apply test taking strategies in the test taking process. Tests should be developed in such a way that preparing for them would enhance the intended abilities. In this case, coaching would have a positive washback. Otherwise, test takers would ignore the operational value of test scores and would attempt to obtain the required score either with or without the inteded ability. This is the reponsibility of test developing organizations to attempt to free the test from the effect of communicatively inappropriate coaching programs.

Refrences

1- Alderson, J.C., Report of the discussion on communicative language testing. In J.C. Alderson & A., Hughes (eds.), ***Issues in language testing***. ELT Documents 111 (pp. 55-65). London: The British Council, 1981.

2- ———————— Language testing in the 1990s: How far have we come? How much further have we to go? In S. Anivan (ed.), ***Current developments in language testing***. Anthology Series 25. Singapore: SEAMEO Regional Language Centre, 1991.

3- Alderson, C. and L. Hamp-Lyons TOEFL Prepartion Courses: A study of washback. ***Language Testing***, 13. 280-287, 1996.

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

5- Amer, A., Teaching EFL students to use test taking strategy. ***Language Testing***, 10, 71-78, 1993.

6- Angoff, W., Validity: An evolving concept. In H. Wainer and H. Braun (eds.) ***Test validity***, 19-33. Hillsdale, New Jersy, 1988.

7- Anivan, S., (ed.) ***Current developments in language testing***. Anthology Series 25. Singapore: SEAMEO Regional Language Centre, 1991.

8- Bachman, L.F., The trait structure of cloze test scores. ***TESOL Quarterly***, 16, 61-70, 1982.

9- —————— ***Fundamental considerations in language testing***. Oxford: OUP, 1990.

10- —————— Languague testing in the turn of the century. ***Languager Testing***, 10, 71-78, 2000.

11- Bailey, K.M., Dale, T.L., & Clifford, R.T. (eds.) ***Language testing research: Selected papers from the 1986 colloquium***. Monterey, California: Defense Language Institute, 1987.

12- ———— Working for washback. ***A review of the washback concept in language testing***. ***Language Testing***, 13, 257-279, 1996.

13- Clapham, C., Languague testing and assessment. ***Encyclopedia of language education***, 1997.

14- Canale, M., The measurement of communicative competence. In R.B. Kaplan (ed.), ***Annual Review of Applied Linguistics*** (Vol. 8, 1987, pp. 67-84), 1988.

15- Carter, K., Test wiseness for students and teachers. ***Educational Measurement: Issues and Practice***, 5(4), 20-23, 1986.

16- Criper, C. and Davies, A., ***ELTS validation project***. The British Council/Cambridge University Local Examination Syndicate, London, 1988.

17- Davies, A., The limits of ethics in language tetsing. ***Languauge Testing***, 14. 235-241, 1997.

18- Diamond, J. & Evans, W., ***An investigation ofthecognitive correlates of test wiseness***. Journal of Educational Measurement, 9, 45-50. 1972.

19- Educational Testing Service, ***Bulletin of information for TOEFL /TWE. ETS***, 1991-2000.

20- Farhady, H., New directions for ESL proficiency testing. In J.W. Oller, Jr. (ed.), ***Issues in language testing research*** (pp. 253- 268). Rowley, Mass.: Newbury House Publishers, Inc, 1983.

21- Farhady, H., Inter, intra, and supra interfaces in language assessment. Paper presented at the LTRC 2000, Vancouver, Cananda, 2000.

22- Hamp Lyons, L., Washback, impact and validity in language tsting. ***Language Testing***, 14, 3, 295-303, 1997.

23- Henning, G., Dimensionality of construct validity of language tests. ***Language Testing***, 9(1), pp. 1-11, 1992.

24- Hymes, D., Models of interaction in social life. In J. Gumperzand D. Hymes (Eds.) ***Directions in sociolinguistics***.35-71, 1972.

25- Ingram, D., The international language testing system IELTS: Its nature and development. In Anivan, S. (ed.) ***Current developments in language testing.*** Anthology Series 25. Singapore: SEAMEO Regional Language Centre, 1991.

26- Izard, J., ***Quality assurance in educational testing***. ETS report, 1988.

27- Munby, J., ***Communicative syllabus design***. CUP, 1978.

28- Millman, J., Bishop, H., & Ebel, R., ***An analysis of test wiseness***. Final Report. College Examination Board, 1966.

29- Perkins, D. & Salomon, G., ***Teaching for transfer***. Educational leadership, 46 (1), 22-32, 1988.

30- Sarnacki, R. E., An examination of test wisenessin the cognitive test domain. Review of educational research, 49, 252-279, 1979.

31- Shohamy, E., Language testing priorities: A different perspective. ***Foreign Language Annals***, 23, 385-94, 1990.

32- ——————— Performance assessment in language testing. ***Annual Review of Applied Linguistics***. 15, 188- 211, 1996.

33- Skehan, P., Issues in the testing of English for specific purposes. ***Language Testing***, 1(2), pp. 202-220, 1984.

34- ————— Language testing: Part II. ***Annul Review of Applied Linguistics***, 22 (1)., pp. 1-13, 1989.

35- Spolsky, B., The limits of authenticity in language testing. ***Language Testing***, 2, 31-40, 1995.

36- Stansfield, C.W., (ed.) Toward communicative competence testing: ***Proceedings of the second TOEFL invitational conference***. Princeton, NJ: Educational Testing Service, 1986.

37- Turner, C., Rasch model programs and scalar analysis. ***Language Testing Update***, 12, 62-64, 1992.

38- Wall, D., Clapham, C., & Alderson, J.C., Evaluating a placement test. ***Language Testing***, 11, 321-344, 1994.

39- Wilkins, D. ***Notional syllabus***. OUP, 1976.