Available online at www.elixirpublishers.com (Elixir International Journal)

Social Sciences

Elixir Soc. Sci. 76 (2014) 28326-28335

ISSN: 2229-712X

The washback effects of TKT, DELTA versus the alternative assessment on the teaching reflection of Iranian EFL teachers

Hamid Reza Shahidy[1] and Rana Azarizad[2]

[1]Islamic Azad University of Garmsar, Daneshjoo St., Daneshjoo Sq., Garmsar, Iran.

[2]Iran University of Science and Technology, Hengan St., Resalat Sq., Tehran, Iran.

ARTICLE INFO

Article history:

Received: 7 September 2014;

Received in revised form:

27 October 2014;

Accepted: 5 November 2014;

Keywords

Washback, DELTA, TKT, Alternative assessment, Reflection.

ABSTRACT

This study endeavors to investigate if assessment tools such as TKT, DELTA, and the alternative assessment have any statistically significant washback effects on the reflection of Iranian EFL in- service teachers. To fulfill the requirements of the study, the researchers selected 90 subjects and categorized them into three groups. Three assessment packages which included the actual samples of TKT, DELTA, and alternative assessment tools, along with the instructional and coaching materials related to all these modes of assessment, were randomly presented to the three groups respectively. The researchers adopted a pre-test post-test comparison group design to investigate the washback effect of each assessment tool and compare the three groups in terms of teacher reflection. Having used one-way ANOVA to analyze the collected data, the researchers concluded that the alternative assessment tools, compared to DELTA and TKT, had the strongest washback effect on teachers' reflection.

Introduction

Assessment in second language teacher education is a rich, complex, and shifting enterprise. Added to this complexity is the more general challenge of assessing teaching as whether to document its processes (what the teacher is doing), or its outcomes (what the students appear to have learned). Freeman and Johnson (2004) believe that in second language teacher education, it is important to position the discussion of the individual teachers who are being assessed in context, since those judgments are, at least in part, a function of the individual teacher's position within the broader social setting and workforce. They also argue that distinction between foreign and second language teachers complicates the task of mapping assessments of what these groups of teachers should know, since, in some circumstances, the knowledge needed may shift when one is teaching a language as a foreign language in one context or teaching the same language as a second language in another. Evaluation of language teachers, according to Bailey (1996) is a complex and contentions topic. Writing in the field of general education, Brazer (1991, p.82) has dubbed teacher evaluation "a theatre of the absurd". Popham (1988) says that it is "with few exceptions, an anemic and important enterprise promising much but producing little" (p.269). Nunan and Lamb (1996) say that for many teachers, supervision and evaluation are mandatory aspects of their terms of employment. External evaluation, particularly when it is for purpose of clarification or continued employment, can be extremely threatening and may be the most anxiety creating situation that the teacher is ever likely to face.

Teacher evaluation, in contrast, according to (Mosher and Purpel, 1972) is a major component of teacher supervision, a profession which has been called "managing messes" (Schon, 1983, p.14). Three basic types of evaluation are identified in the program evaluation literature. The two most frequently discussed are formative and summative evaluations. These two terms apply to teacher evaluation as well. Formative evaluation is used to gain intermittent feedback concerning the nature of some activity or practice while it is in progress (Daresh, 2001). While formative evaluation of teachers is related to promoting professional development and helping teachers improve, summative evaluation of teachers is associated with tenure promotion or "terminating" (Hazi, 1994, p.200). Acheson and Gall (1997) discuss the specific steps that should be followed in planning formative and summative evaluation. Teacher evaluation depends on some understanding of teaching, but over the years there has been a great deal of debate about how to define and measure teacher effectiveness. As Stodolky (1984) noted "evaluation of teachers rests on the assumption that the characteristics of good or effective teachers are known and recognizable" (p.11). Nowadays, however, many criteria are used in language teacher evaluation. Pennington (1989) categorizes language teacher evaluation tools as being either fluid-response instruments or fixed-response instruments. The fluid response instruments include "conversation, letters, and open ended questionnaires, rating scales, tests, and different kinds of summative descriptive data" (p.168). There are pros and cons of both types of data. "Fluid responses instruments allow individuals to comment on teachers' work, but they are difficult to interpret, to tally and to score in any reliable manner" (Pennington, 1989, p.169). Fixed response instruments are effective in terms of the initial ratings and subsequent tabulations, but they "discourage reflective thoughtful responses and do not allow respondents to convey detailed specific information" (p.169). Murdoch (2000, pp. 55-6) states that teacher evaluation needs to be founded on five key principles or aims: "1) To encourage reflective practice; 2) To empower and motivate teachers; 3) To assess all aspects of a teacher's professional activity; 4) To take account of students' views; and 5) To promote collaboration."

Tele:

E-mail addresses: rana.azarizad@gmail.com

By the late 1980s in the United States, requirements existed for either full certification or endorsements in teaching most foreign languages. These assessments included tests in the target language, methodology, and cultural knowledge (McFerren, 1988). Presently, standardized tests such as the ETS praxis battery assess language knowledge, metalinguistic knowledge, and pedagogical knowledge (Educational Testing Service, 2005). In 2005, the University of Cambridge ESOL Assessments developed the Teaching Knowledge Test (TKT), which is now offered in 21 countries. The Teaching Knowledge Test (TKT) is a test about teaching English to speakers of other languages. It aims to increase teachers' confidence and enhance job prospects by focusing on the core teaching knowledge needed by teachers of primary, secondary or adult learners, anywhere in the world. This flexible and accessible award will help you to understand: 1) Different methodologies for teaching. 2) The 'language of teaching'. 3) The ways in which resources can be used Cambridge ESOL, a department of the University of Cambridge, has designed and produced DELTA as part of a framework of teaching awards and tests for prospective English teachers. It covers all areas of knowledge at an advanced level and includes teaching practice. DELTA modules can be taken at any stage in a teachers' career. DELTA is suitable for in-service teachers of English in a variety of teaching contexts e.g., adult, primary, or secondary teaching contexts, and are intended for an international audience of non-first language or first language teachers of English. Candidates taking Delta Module One, Two or Three will normally have an initial ELT qualification and will have had at least a year's ELT experience but these are not requirements. The Delta Modules may also be taken by: '1) Teachers who wish to refresh their teaching knowledge. 2) Teachers who wish to review and updates their practice. 3) Teachers who wish to extend their expertise in a specialist area.' Early in the decade of the 1990s, as teachers and students were becoming aware of the short comings of standardized testing and all the problems found with such testing , a novel concept emerged that began to be labeled 'alternative assessment'. A variety of labels has been used to distinguish it from traditional, standardized testing: performance assessment, authentic assessment, portfolio assessment, and assessment by exhibition (Garcia & Pearson, 1994). While the standardized test industry has become a powerful juggernaut of influence on decisions about people's lives, it also has come under severe criticism from the public (Kohn, 2000). This new form of assessment, according to Richards and Renandya, focuses more on measuring one's ability to use language holistically in real-life situations and is typically carried out continuously over a period of time. Alternative assessment is also more multiculturally sensitive and free of linguistic and cultural biases found in traditional testing. The literature (Holt, 1994) presents ample discussions and illustration of a variety of alternative assessment which can be adapted to varying situations. Assessment tools such as portfolios, interview, project work, observation by peers, self-observation, self- or peer -assessment, journals, are among the more authentic forms of alternative assessment. Such procedures provide teachers with useful information that can be the basis for improving their instructional plans and practices. However, alternative assessment is not without its concern since some have doubts about the reliability of the procedures that are use as well as the administrative feasibility and cost effectiveness of alternative assessment.

Alternative assessment is an approach which comprises a range of perspectives that share the same purpose with formative and diagnostic assessment. It is different from traditional testing in that it actually asks test candidates to show what they can do. "The candidates are evaluated on what they integrate and produce rather than on what they are able to recall and reproduce" (Richards & Renandya, 2002). "The main goal of this form of assessment is to gather evidence about how students are approaching, processing, and completing 'real-life' tasks in a particular domain" (Garcia & Pearson, 1994, p. 357). Alternative assessment, in this study, includes instruments such as observation by fellow teachers, self-observation, reflection questionnaire, and evaluation of teachers by their students' questionnaire, teacher portfolio assessment. Alternative assessment, in a nutshell, provides feedback that enables teachers to self-assess and self-adjust their performance.

Test Washback

Examinations have been long used as a means of control. They have been with us for a long time, at least a thousand years or more, if their use in Imperial China to select the highest officials of the land is taken into consideration. Those examinations were probably the first Civil Service Examinations ever developed by the human race. Although the goal of the examination was to select civil servants, its washback effect was to establish and control an educational programme, as prospective Mandarins set out to prepare themselves for the examinations. Even in modern times, the use of examinations to select for education and employment dated back at least 300 years. Examinations are often subject to much criticism. Madaus (1988, p. 85), for instance, pointed out: "The tests can become the ferocious master of the educational process, not the compliant servant they should be. Measurement-driven instruction invariably leads to cramming, narrows the curriculum, and concentrates attention on those skills most amenable to testing. It also constrains the creativity and spontaneity of teachers and students, and finally demeans the professional judgment of teachers". However, in spite of all the criticism leveled at them, examinations continue to occupy a leading place in the educational arrangement of many countries (Eckstein & Noah, 1992). Such use of tests for power and control, as pointed out by Shohamy (1993), is an especially common practice in countries that have centrally controlled educational agencies (Heyneman, 1987; Heyneman & Ransom, 1990; Li, 1990; Workman, 1987). Shohamy (1993) argues that policy-makers in central agencies, aware of the power of tests, use them to manipulate educational systems, to control curricula, and to impose new textbooks and new teaching methods.

Messick (1996) views washback as part of consequential validity. He believes that If the concept of washback is to have any meaning, it is necessary to identify what changes in learning and teaching can be directly attributed to the use of the test in that context. Traditionally, this meant creating an empirical link between a negative consequence and a source of invalidity (Fulcher & Davidson, 2007). Alderson & Wall (1993) argue that other forces exist within society, education, and schools that might prevent washback from appearing, or that might affect the nature of washback despite the communicative quality of a test. They imply that washback is likely to be a complex phenomenon which cannot be related directly to a test's validity. Framing the 'washback hypothesis' Alderson and Wall (1993) made it possible for washback to be studied empirically, and the simplistic nature of the original concept was soon turned into a conceptually rich source of theory and research. Many educationalists have written about the power of examination over what takes place in the classroom. Pearson, for example, says "it is generally accepted that public examinations influence

the attitudes, behavior, and motivation of teachers, learners, and parents" (1988, p. 98).Vernon (1956, p. 166) claims that examinations "distort the curriculum". Morris (1972, p.75); however, considers examinations necessary to ensure that "the curriculum is put into effect ". Swain (1985, pp. 42-4) recommends that test developers "bias for best" and "work for washback". Alderson (1986, p. 104) argues for innovations in the language curriculum through innovations in language testing". Hughes (1993, p. 20) focused on participants, processes and products in this model to illustrate the washback mechanism. Hughes further notes:

"The trichotomy into participants, process and product allows us to construct a basic model of backwash. The nature of a test may first affect the perceptions and attitudes of the participants towards their teaching and learning tasks. These perceptions and attitudes in turn may affect what the participants do in carrying out their work (process), including practicing the kind of items that are to be found in the test, which will affect the learning outcomes, the product of the work."

While Hughes focused on participants, processes, and products in his backwash model, Smith (1991), tried to construct a model showing five components of change: the target system, the management system (consisting of both the members of the system and the structures within the system), the innovation itself, available resources, and the environment in which the change is supposed to take place. Alderson & Wall (1993 p.120-121) came up with 15 hypotheses regarding washback to illustrate areas in teaching and learning that were usually affected by washback. Some of the components of the washback hypothesis, with respect to teachers, are as follows: 1) A test will influence what teachers teach; and 2) A test will influence how teachers teach. 3) A test will influence the rate and sequence of teaching; and 4) A test will influence the degree and depth of teaching. 5) A test will influence attitude towards the content, method, etc., of teaching and learning. 6) Tests that have important consequences will have washback. 7) Tests will have washback effects for some learners and some teachers, but not for others.

Teachers are the stakeholders who are often directly affected by assessments. Most teachers, according to Bachman and Palmer (2010) are familiar with the ways in which an externally mandated assessment can influence their instruction. Despite the fact that teachers may personally prefer to teach certain material in a specific way, if they find that they have to use a specified assessment , they may find teaching to the test almost unavoidable (Gipps, 1994). If the content of the assessment is thus aligned with the goals and objectives of instruction and with instructional objectives, then teaching to the test may become an aspect of positive impact on instruction (Bachman & palmer, 1996). Any test is likely to influence the behavior of students and teachers, provided that they know about it in advance. Popham (1987) regards measurement driven instruction as the most cost-effective way of improving the quality of public. Popham refers to measurement-driven instruction (M.D.I) as when a high-stake test influences the instructional programme that prepares students for the tests. Stakes can be high either for the students or for the teachers, and in some cases they are high for both. Teachers tend to spend a significant amount of their teaching time on the knowledge and skills assessed by such a test. Therefore, high -stakes assessment serves as a powerful curricular magnet. Alderson and Wall argue that further research on washback is needed, and that such research must entail increasing specification of the Washback Hypothesis above (1993, p. 127). Shohamy et al. (1996, p.299) stated," the power and authority of tests enable policy-makers to use them as effective tools for controlling educational systems and prescribing the behavior of those who are affected by their results, namely administrators, teachers and students".

Teacher Reflection

If teachers are to be effective in whatever approach they decide to take, it seems reasonable to expect them to act consistently in accordance with their expressed or espoused beliefs (Williams & Burden, 1997). Unfortunately, according to Argyris and Schon (1974) this hardly ever occurs in any professions. In an effort to improve teachers' self – awareness in this respect, some educational theorists have fostered the notion of critical reflection (Bound, Keogh, & Walker, 1985). The intention is to enable teachers to become reflective practitioners (Schon, 1983); thereby they subject their everyday professional practice to ongoing critical reflection and make clear their own particular world view by means of such consideration. Schon (1983, p.49) draws the distinction between "reflection – in – action" and "reflection – on – action". He contends that each individual's knowledge is mainly tacit and implied by the ways in which they act. (Schon, 1983, p.49). The task of the reflective practitioner is to make this tacit or implicit knowledge explicit by reflection on action, by constantly generating questions and checking our emerging theories with both personal past experience and with reflection of others. This is one of the main thrusts of the movement towards teachers as action researchers. Any school supportive of reflective teaching would find it necessary to consider the kind of structure within which learning takes place and the very nature of knowledge itself. Critical reflection, according to Smyth (1991), is not necessarily negative in its orientation, but it does imply that teachers should be aware of their belief systems and constantly monitoring how far their actions reflect those beliefs or are in keeping with them. Ruddock (1984, p.6) points out that "not to examine one's practice is irresponsible; but to regard teaching as an experiment and to monitor one's performance is a responsible professional act".

Williams and Burden (1997) argue that to be an effective teacher, we need to look both inwards and outwards. As reflective teachers, we need to develop our awareness of others' viewpoints, and values. We then need to construct a particular identity of the kind of teacher that we want to be and to seek to reproduce this in our day to day activities, in our actions and in our interaction in the teaching learning areas. A 'reflective' model of teacher education incorporates teachers more actively into the education process. In this model, teachers utilize experiential and received knowledge in their practice, and engage in reflection which allows them to re-examine their practice in light of their decisions, concerns, experiences, and knowledge (Schon, 1983). In fact, Freeman and Johnson (1998) feel that what teachers think and believe about their practices comprise key components in determining what their students do or do not learn. However, he raises concerns about teacher education models which may focus heavily on teacher reflection. He emphasizes the importance of maintaining a balance between the more theoretical and the more experiential forms of teacher education in his call for a 'socioliterate' approach to teacher education. Van Lier (1996), Brown and McIntyre (1993) argue that teacher reflection plays an important role in language teacher education (LTE). Dewey's (1933,1938) distinction between routine and reflective action in teaching highlighted the importance of teachers reflecting systematically upon their working contexts, resources, and actions and applying what they learned from reflection in their everyday and

long-term decision making. He identified three essential teaching qualities: teachers should listen to all points of view (open-mindedness), be alert to all the consequences of their actions (responsibility), and have these qualities at the core of their being and actions (wholeheartedness). Schon's (1983) distinction between reflection-in-action and reflection-on-action was a reminder that teachers make judgment and decisions in the classroom all the time. Schon characterized teachers' theorizing in two ways: drawing upon theories in use when reflecting in action, and drawing upon teaching experience and espoused theories when reflecting on action.

Korthagen and Vasalos (2005) proposed an onion model, which, based on a concept of core reflection, demonstrated how teachers can be helped to progress to deeper levels of reflection by peeling away layers of the 'onion'. Working within the frame of phenomenology suggested that intuition-in-action may be a better description of the judging and decision-making abilities that teachers employ while teaching (Johanson and Kroksmark , 2004) . Research on the nature of teacher education has overall revealed a central but unresolved role for reflective practice in language teacher education. Conceptual research on teacher knowledge (e.g., Freeman and Johnson 1998) tends to draw upon the socially constructed nature of knowledge (e.g., Vygotsky 1962) and personal professional identity. Teachers should acknowledge that being a professional does not mean doing things perfectly all the time. Rather, it means accepting that there are always be a better way of doing things and resolving to do things differently next time around.

Statement of the problem and the research questions

One of the simple facts of life in the present time is that the English language skills of a good proportion of our people are seen as vital if our country is to have access to the information and knowledge that provide the basis for both social and economic development. There is consequently increasing demand for competent English language teachers and for more effective approaches to their preparation and professional development. Central to this enterprise are English teaching and English language teachers. Knowledge for- teaching tends to be defined exclusively as content knowledge among most language teachers in our public education system. Pedagogical knowledge is rarely focused on in general assessments; although it must be recognized as part of what teachers need to teach English as a foreign language. Knowledge- for- teaching is equated to knowing the content knowledge. Teachers at language schools in Tehran are prepared differently, and often with different degrees of exposure to training in the knowledge and practices they need to teach effectively. Considering the varying ways that teachers are deemed qualified internationally, Barduhn and Johnson (2009, p.61) call for "fairer and more rigorous assessments ". Further they note that, in comparison to the standardized assessments of teaching as observable behavior used conventionally, portfolios and other reflective documents may be fairer in documenting the contextual and idiosyncratic aspects that make teaching practice effective (p. 62). The researchers' experience with language teachers at a variety of language schools showed that teachers are not much familiar with the content of TKT, let alone DELTA, TKT, DELTA, ETS Praxis Series tests, and malternatives to assessment such as self-observation and teacher portfolio . Furthermore, formative and summative evaluation of EFL in – service teachers is not commonly undertaken by supervisors in most language schools.

The primary purpose of this study is to investigate if Teaching Knowledge Test (TKT), Delta Modules, and the alternative assessment have any washback effects on the reflection of Iranian EFL in-service teachers. Moreover, the researchers aim to discover if these assessment tools affect the teacher variables differently. Cheng (2005) argues that beliefs about testing reflect beliefs about teaching and learning.

This study aims to answer the following questions.

Q1. Does TKT have a washback effect on the teacher reflection of Iranian EFL teachers?

Q2. Does DELTA have a washback effect on the teacher reflection of Iranian EFL teachers?

Q3. Does the alternative assessment have a washback effect on the teacher reflection of the Iranian EFL teachers?

Q4. Do TKT, DELTA, and the alternative assessment have any significantly different washback effects on the teacher reflection of Iranian EFL teachers?

Participants

To achieve the goals of this study, 102 research participants were selected from 150 language teachers, and then were incorporated into three groups in a random fashion. The 102 research participants were all teaching English in different branches of Simin Language Institute, and scored between one standard deviation below and one standard deviation above the mean on an actual paper – based test of English as a foreign language (TOEFL-PBT). Almost all had college degrees such as bachelor's or Master's Degrees in English. They all started teaching English after they had passed a standard entrance exam (actual TOEFL - PBT) and a teacher training course held by the institute. Our subjects had, on average, five years of teaching experience at the institute and other language schools in Tehran. They taught English at different levels of language proficiency ranging from beginning to advanced levels. Our subjects were both male and female. They were non-native teachers who were not statistically different in terms of language proficiency.

Instrumentation

To collect reliable data for the purpose of testing the null hypotheses, the researchers applied the following assessment tools in this study.

Teacher Reflection: Questionnaire (Akabari, Behzadpoor; and Dadvand, 2009)

The instruments used in the treatment phase.

1. Teaching Knowledge Test (TKT University of Cambridge ESOL Examinations - , 2005)
2. DELTA (University of Cambridge ESOL Examinations, 2008)
3. Alternative Assessment

a. Teacher self-observation form (Christison & Bassano, 1984)
b. Teacher observation by others (Brown, 2007)
c . Teacher portfolio assessment (Doolittle, 1994)
d. 'Evaluation of teachers by their students' questionnaire (National Schools of Character : Award Winning Practices, 2005).

Procedure

This study was carried out in three phases, To begin with, the researchers did a survey on a number language schools in Tehran to find teachers who were willing to cooperate in this study. Unfortunately, almost all of them except for Simin language institute were disinclined to have their teachers exposed to such teacher washback project. Since the teachers were on tight schedules, the researchers inquired about their most convenient time at which they were able to take the test. Almost all the teachers who participated in this study were on tight schedules, which made the researchers to administer the pre-test to one or two or 5 teachers at a time. The researchers selected 102 subjects who scored one standard deviation above and below the mean obtained from the performance of the

teachers on an actual paper- based TOEFL administered by Simin Language Institute. Afterwards, the researchers divided them into three groups in a random fashion.

In the second phase of the study, Researchers did a survey on the teacher reflection of all the 102 teachers who were randomly incorporated into three groups. To do so, the researchers gave all the subjects a teacher reflection questionnaire. Later, the actual samples of TKT, DELTA, and alternative assessment tools along with the instructional and coaching materials related to all these modes of assessment were randomly assigned to the three groups respectively. Packages came in three different colors (orange, yellow, violet) .The purpose of giving these packages to teachers was to familiarize them with the new assessment tools and to use these tools as leverage to make teachers study the resources related to the content of these assessment packages and reflect on them. The orange package contained an actual version of DELTA and its relevant coaching materials. The yellow package contained an actual sample of TKT and its relevant coaching materials. The violet package had alternative assessment forms such as Teacher self-observation form , teacher observation by others form, teacher portfolio format , evaluation of teachers by their students questionnaire , and the instructions on how to live up to the requirements of the alternative assessment. In order to brief the research participants (the teachers in the three groups), the researchers met each teacher in person to coach him or her on how to prepare for the teaching assessment tools .The assignment of the assessment tools to the three groups in question is described below.

Group A

To assign the TKT package and coach on the content of TKT, the researchers met the teachers in group A one at a time. While coaching, the researchers informed the research participants that an actual sample of TKT was to be administered at the end of the semester. He also asked them to study the TKT sample items as well as the coaching materials in order to prepare for it. The teachers, in this study, were also asked to give a written report to their head teachers on their study with respect to the assigned TKT package.

Group B

With the teachers in this group, the researchers followed the same procedure which he did for the TKT group except that they were assigned the DELTA package which contained actual DELTA items and the relevant coaching materials. They were all coached on the content of the package and told that they were required to prepare for an actual version of DELTA which was to be administered at the end of the course.

Group C

To achieve the requirements of the alternative assessment which lasted for five months, the researchers submitted an alternative assessment package to the teachers in group C. The package included 'Teacher self-observation' form, 'Teacher observation by others' form, teacher portfolio format, 'Evaluation of teachers by their students' questionnaire '. Later the research participants were asked to study the content of the alternative assessment package and try to adapt their reflective teaching to meet the requirements of these. The researchers notified them that the supervisors were planning to observe their classes on the basis of the observation forms in the package. Moreover, the researchers had the teachers in group C to write a teacher portfolio and submit it to their supervisor at the end of each month. The research participants were informed that the results of self -observations and the teacher portfolios were expected to be in line with 'teacher observation by others', and 'evaluation of the teachers by their students'. Thus, they had to fill out the self-observation form, and prepare the teacher portfolio as honestly as possible. The instruction on how to write a teacher portfolio accompanied the alternative assessment package. The researchers observed the teachers in group C twice a month and filled out the 'teacher observation by others form'. At end of each observation, he asked the teachers to fill out the self-observation form. Furthermore, at the end of each month, the researchers distributed the teacher evaluation questionnaires among the teachers' students and, later, gave teachers feedback on the students ' opinion poll in order to persuade teachers to make necessary adjustments so that they fit in with the needs and expectations of their students.

In the last phase of the study, a reflection questionnaire was administered to the three groups to determine if the assigned assessment packages (TKT, DELTA, and the alternative assessment tools) had any significant impact on the teacher reflection of the teachers in groups A, B, and C respectively, and to identify which assessment had a stronger washback effect on the reflection of our research participants.

Results and Discussion

This study has addressed four research questions which can be summarized as follows:

The first questions number 1 to 3 dealt with the impact of TKT, DELTA, and the teaching alternative assessment on the teacher reflection of our research participants respectively. The fourth question is concerned with whether these three types of assessment have significantly different washback effects on the teacher reflection .Theses research questions have been restated as four null hypotheses which have been tested by analyzing the findings obtained through a pre-test post-test comparative research design. The statistical analyses and interpretations of the research data are as follows. The present data are measured on an interval scale. The subjects are independent, i.e. the performance of any of the subjects on the tests is not affected by the performance of other subjects. The assumptions of normality and homogeneity of variances require that the population - not the samples - from which the samples have been selected, should have a normal distribution and should show homogeneous variances (Filed, 2009; Pallant, 2005). The normality of the present data was tested. As displayed in Table 1. The ratios of skewness and kurtosis over their respective standard errors were within the ranges of +/- 1.96 (Filed, 2009; Pallant, 2005).

Table 1. Normality Tests, Reflection

Groups		N	Skewness			Kurtosis		
		Statistic	Statistic	Std. Error	Normality	Statistic	Std. Error	Normality
AA	Pretest	30	-.315	.427	-0.738	-.906	.833	-1.088
	Posttest	30	.724	.427	1.696	.168	.833	0.202
DELTA	Pretest	30	.223	.427	0.522	-.004	.833	-0.005
	Posttest	30	-.639	.427	-1.496	-.291	.833	-0.349
TKT	Pretest	30	.244	.427	0.571	-.005	.833	-0.006
	Posttest	30	.374	.427	0.875	1.366	.833	1.639

The assumption of homogeneity of variances need not be checked either particularly when the sample sizes are equal (Bachman, 2004) as is the case in this study. However, the assumption of homogeneity of variances was also checked through the Levene's tests the results of which will be discussed when reporting the one-way ANOVA results.

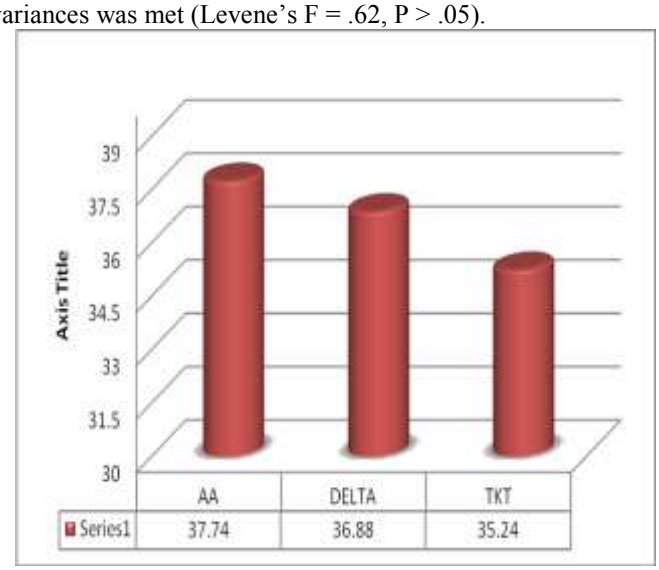
Pretest of Reflection

A one-way ANOVA was run to compare the AA, DELTA and TKT groups on the pretest of reflection in order to prove that they were homogenous in terms of their reflection prior to

the administration of the treatment. As displayed in Table 2, the mean scores for the Alternative assessment, DELTA and TKT groups on the pretest of reflection are 37.74, 36.88 and 35.24 respectively.

The results of the one-way ANOVA (F (2, 87) = 2.41, P > .05; ω^2 = .03) showed an almost weak effect size, which indicated that there were not any significant differences between the means of the three groups on the pretest of reflection.

It should be noted that the assumption of homogeneity of variances was met (Levene's F = .62, P > .05).



Graph 1. Pretest of Reflection

In order to find the answer to this research question: Do TKT, DELTA, and the alternative assessment have any washback effects on the reflection of Iranian foreign language teachers?, a one-way ANOVA was run to compare the TKT, DELTA, and AA groups on the posttest of reflection . As displayed in Table 5, the mean scores for the AA, DELTA and TKT groups on the posttest of reflection are 45.33, 40.33 and 37.17 respectively.

The results of the one-way ANOVA: (F (2, 87) = 25.25, P < .05; ω^2 = .35 showed a large effect size, which indicate that there were significant differences between the means of the three groups on the posttest of reflection. Thus, the fourth null-hypothesis, 'there is no significant difference among TKT, DELTA, and the alternative assessment In terms of the washback effect which they have had on the reflection of Iranian EFL teachers 'is rejected'.

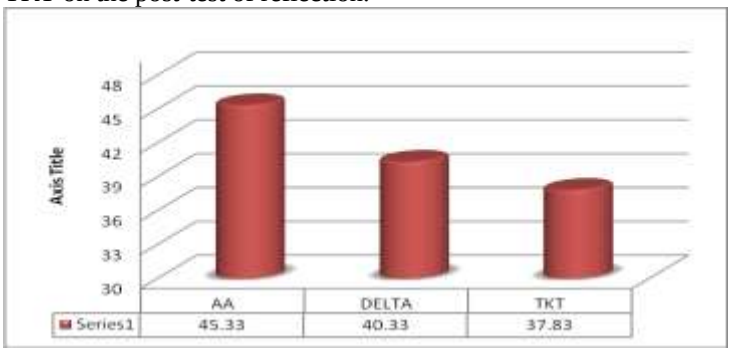
It should be noted that the assumption of homogeneity of variances was met (Levene's F = 2.54, P > .05).

The F-value of 25.25 indicates significant differences between the mean scores of the three groups. Moreover, the post-hoc Scheffe's tests were run to compare the groups two at a time. Based on the results displayed in Table 8, it can be concluded that

A: There is a significant difference between the reflection of the alternative assessment (M = 45.33) and DELTA (M = 40.33) groups on the posttest of reflection (M difference = 5.001, P = .000 < .05).

B: There is a significant difference between the reflection of the alternative assessment (M = 45.33) and TKT (M = 37.83) groups on the posttest of reflection (M difference = 7.499, P = .000 < .05).

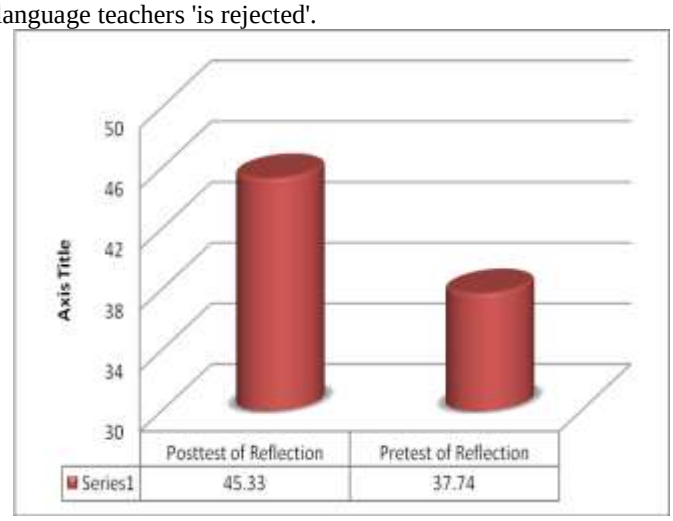
C: There is no statistically significant difference between the reflection of the DELTA (M = 40.33) and TKT (M = 37.83) groups on the posttest of reflection (M difference = 2.49, P = .073 > .05), although DELTA showed a higher mean than that of TKT on the post-test of reflection.



Graph 2. Posttest of Reflection

In order to answer this research question: 'Does the Alternative assessment have a washback effect on the reflection of Iranian EFL teachers? ' , a paired-samples t-test was run to compare the mean scores of the AA group on pretest and posttest of reflection. As displayed in Table 10, the AA group shows a higher mean (45.33) on the posttest of reflection than on the pretest (M = 37.74).

The results of the paired-samples t-test indicate that the difference between the two means is statistically significant (t (29) = 6.13, P < .05, r = .75, it does represent a large effect size). Thus the first null-hypothesis: ' the alternative assessment does not have a washback effect on the reflection of Iranian foreign language teachers 'is rejected'.



Graph 3. Mean Scores on Pretest and Posttest of Reflection (AA Group)

In order to answer this research question: 'Does DELTA have a washback effect on the reflection of Iranian foreign language teachers?' , a paired-samples t-test was run to compare the mean scores of the DELTA group on pretest and posttest of reflection. As displayed in Table 12. , the DELTA group shows a higher mean (40.33) on the posttest of reflection than on the pretest (M = 36.88).

The results of the paired-samples t-test indicate that the difference between the two means is statistically significant (t (29) = 3.14, P < .05, r = .50, it does represent a large effect size). Thus the second null-hypothesis: ' DELTA does not have a washback effect on the reflection of Iranian EFL teachers 'is rejected.

In order to answer this research question: Does TKT have a washback effect on the reflection of Iranian EFL teachers? A paired-samples t-test was run to compare the mean scores of the TKT group on the pretest and posttest of reflection.

Table 2. Descriptive Statistics -Pretest of Reflection

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
Pretest Of Reflection	AA	30	37.74	4.568	.834	36.04	39.45	28	45
	DELTA	30	36.88	4.283	.782	35.28	38.48	28	45
	TKT	30	35.24	4.603	.840	33.52	36.96	26	46
	Total	90	36.62	4.558	.480	35.67	37.58	26	46

Table 3. One-Way ANOVA- Pretest of Reflection by Groups

		Sum of Squares	Df	Mean Square	F	Sig.
Pretest Of Reflection	Between Groups	97.046	2	48.523	2.410	.096
	Within Groups	1751.670	87	20.134		
	Total	1848.716	89			

Table 4. Homogeneity of Variances - Pretest of Reflection

	Levene Statistic	df1	df2	Sig.
Pretest	.622	2	87	.539

Table 5. Descriptive Statistics - Posttest of Reflection

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
Posttest Of Reflection	AA	30	45.33	4.887	.892	43.51	47.16	39	58
	DELTA	30	40.33	3.947	.721	38.86	41.80	31	46
	TKT	30	37.83	3.534	.645	36.51	39.15	29	47
	Total	90	41.17	5.173	.545	40.08	42.25	29	58

Table 6. One-Way ANOVA - Posttest of Reflection by Groups

		Sum of Squares	Df	Mean Square	F	Sig.
Posttest Of Reflection	Between Groups	874.767	2	437.383	25.258	.000
	Within Groups	1506.545	87	17.317		
	Total	2381.311	89			

Table 7. Homogeneity of Variances - Posttest of Reflection

	Levene Statistic	df1	df2	Sig.
Posttest	2.548	2	87	.084

Table 8. Post-Hoc Scheffe's Tests - Posttest of Reflection

(I) Groups	(J) Groups	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
AA	DELTA	5.001*	1.074	.000	2.32	7.68
	TKT	7.499*	1.074	.000	4.82	10.17
DELTA	TKT	2.498	1.074	.073	-.18	5.17
*. The mean difference is significant at the 0.05 level.						

Table 9. Reliability Indices of the Reflection questionnaire

	N of Items	Mean	Variance	K-R21
Pretest145	145	103.80	205.825	.86
Posttest145	145	119.37	225.045	.91

Table 10. Descriptive Statistics Pretest and Posttest of Reflection (AA Group)

	Mean	N	Std. Deviation	Std. Error Mean
Posttest of Reflection	45.33	30	4.887	.892
Pretest of Reflection	37.74	30	4.568	.834

Table 11. Paired-Samples t-test - Pretest and Posttest of Reflection (AA Group)

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Posttest vs. Pretest of Reflection	7.589	6.771	1.236	5.061	10.118	6.139	29	.000

Table 12. Descriptive Statistics - Pretest and Posttest of Reflection (DELTA Group)

	Mean	N	Std. Deviation	Std. Error Mean
Posttest of Reflection	40.33	30	3.947	.721
Pretest of Reflection	36.88	30	4.283	.782

Table 13. Paired-Samples t-test , Pretest and Posttest of Reflection (DELTA Group)

	Paired Differences					t	Df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Posttest vs. Pretest of Reflection	3.448	6.008	1.097	1.205	5.692	3.144	29	.004

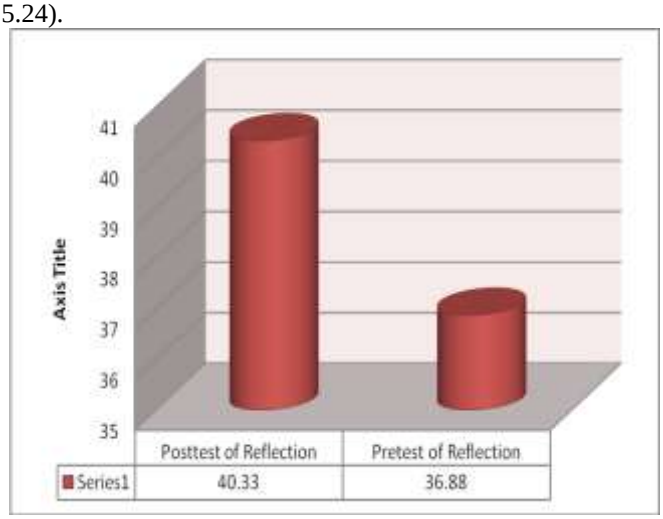
Table 14. Descriptive Statistics Pretest and Posttest of Reflection (TKT Group)

	Mean	N	Std. Deviation	Std. Error Mean
Posttest of Reflection	37.83	30	3.534	.645
Pretest of Reflection	35.24	30	4.603	.840

Table 15. Paired-Samples t-test Pretest and Posttest of Reflection (TKT Group)

	Paired Differences					t	Df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Posttest vs. Pretest of Reflection	2.594	6.012	1.098	.349	4.839	2.363	29	.025

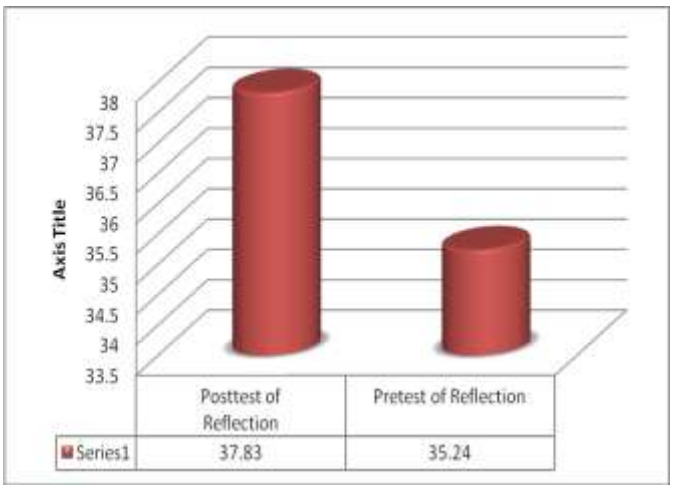
As displayed in Table 3, the TKT group shows a higher mean (37.83) on the posttest of reflection than on the pretest (M = 35.24).



Graph 4. Mean Scores on Pretest and Posttest of Reflection (DELTA Group)

The results of the paired-samples t-test indicate that the difference between the two means is statistically significant (t (29) = 2.36, P < .05, r = .40, it represent a moderate to large effect size). Thus, the third null-hypothesis: ' TKT does not have a washback effects on the reflection of Iranian EFL teachers 'is rejected'.

In a nutshell, we can say that the alternative assessment tools, compared to DELTA and TKT, had a stronger washback effect on the teacher reflection .DELTA had a stronger washback effect than TKT on the teacher reflection. Last but not least, TKT had statistically significant effect on the teachers' reflection, although it had less effect in comparison with TKT and DELTA.



Graph 5. Mean Scores on the Pretest and Posttest of Reflection (TKT Group)

Conclusion and Pedagogical Implications

On the whole, this study examined the washback effects of three teaching assessment tools TKT, DELT, and the alternative assessment (Teacher self-observation, Teacher observation by others, Teacher portfolio assessment, Evaluation of teachers by their students) on the reflection of Iranian EFL in-service teachers. Discussing the results of the study, the researchers rejected the entire four research null hypotheses. The results showed that all three modes of assessment (TKT, DELTA, and the alternative assessment) had statistically significant washback effects on the teacher reflection of all research participants in the three groups. Furthermore, the findings revealed significant differences in the washback effects of the TKT, DELTA, and the alternative assessment on the reflection of our subjects (in-service teachers).

The current study is situated along the above line of discussion of using 'better' assessment to bring about 'better' teaching and learning (more real-life activities and more active learning). The findings of this research will be beneficial to both

pre-service and in-service teachers in some certain ways. This study will give language teachers a real insight into the strengths of test washback in the realm of teacher education. The findings of this research are also beneficial to language school supervisors in that they can use test washback as leverage to persuade in-service teachers to improve reflection and critical thinking skills in the realm of critical pedagogy. Teacher evaluation also assists student teachers in cooperating fully with the head teachers in their teaching training. It also draws their attention to crucial concepts such as self- assessment and self-study to the extent that every teacher student learns to be responsible for his or her professional development. This research raises awareness about the fact that a high level of proficiency in English is a necessary prerequisite but not a sufficient standard for achieving pedagogical expertise. The washback effect of teaching knowledge tests can be used to prevent experienced teachers' knowledge from being fossilized. Another benefit is that teachers and head teachers stop considering teacher evaluation a threat of the absurd"(Brazer, 1991, p.82) and avoid regarding it as an "anemic enterprise promising much but producing little" (Popham, 1988, p. 269). Moreover, language head teachers will come to an understanding that using the teaching knowledge assessment with high consequential validity (washback validity) could drive language teachers to reflect on their success and failure, and familiarize themselves with modern issues in applied linguistics and language assessment. There is no question that reflection on what to teach and how to teach help language teachers to keep up with the latest development in TEFL, and to discover what other people are thinking and doing. (Borg, 2003). Test washback can play an important role in urging in-service language teachers to bring themselves up-to-date with the latest standards of effective teaching. Last but not least, it will certainly motivate researchers, in the realm of teacher education, to investigate the effect of teaching knowledge assessment on other teacher variables, and to identify the factors which possibly interfere with test washback. Teachers, for instance, can be considered to be a crucial factor in mediating test washback. It is well worth mentioning that the washback effect of teaching knowledge assessment tools on teachers is an issue in language testing which, to the researchers' knowledge, seems not to have been attempted by any other researchers.

Acknowledgment

We would like to express our heartfelt thanks to those who made this research possible to conduct, namely our professors, friends and families.

References

[1] K. A. Acheson, M. D. Gall, Techniques in the clinical supervision of teachers: Perspective and in service applications. NY: Longman, 1997.

[2] R., Akbari, , F. Behzadpoor, and B. Dadvand, "Development of English language teaching reflection inventory," System, vol. 38, pp. 211-227, 2009.

[3] J. C. Alderson, Innovations in language testing. London: Nelson, 1986.

[4] J. C. Alderson, Washback in language testing: Research contexts and methods. NJ: Erlbaum, 2004.

[5] J.C. Alderson , D. Wall, "Dose washback exist?" Applied Linguistics, vol. 14, pp. 115- 129, 1993.

[6] C. Argyris, D.A. Schon, Theory in practice. San Francisco: Jossey-Bass, 1974.

[7] L.F. Bachman, Statistical analyses for language assessment. London: Cambridge University Press, 2004.

[8] L. F. Bachman, "Building and supporting a case for test use," Language Assessment Quarterly, vol. 2, pp. 1-34, 2005.

[9] L. Bachman, A. Palmer, Language testing in practice. Oxford: OUP, 1996.

[10] L. Bachman, A. Palmer, Language assessment in practice. New York: Oxford University Press, 2010.

[11] K. M. Bailey, "Working for washback: A review of the washback concept in language testing," Language Testing, vol. 13, pp. 157-279, 1996.

[12] S.D. Brazer, "The assistant principal: The search for meaning in teacher evaluation," Educational Leadership, vol. 48, pp. 82, 1991.

[13] H.D. Brown, Teaching by principles: An introductive approach to language pedagogy. NY: Pearson Longman, 2007.

[14] S. Brown, D. McIntyre, Making sense of teaching. Buckingham: The Open University Press, 1993).

[15] L. Cheng, "Changing language teaching through language testing: A washback study," Language Testing Journal, vol. 25, pp. 145-149, 2005.

[16] M. A. Christison, S. Bassano, "Teacher self-observation," TESOL, vol. 18, pp. 17- 19, 1984.

[17] A.D. Cohen, "On taking language tests: What the students repot," Language Testing, vol. 1, pp. 70-81, 1984.

[18] J.C. Daresh, Supervision as proactive leadership. IL: Waveland press, 2001.

[19] J. Dewey, How we think. New York: Collier Books, 1933.

[20] J. Dewey, Experience and education. New York: Collier Books, 1938.

[21] M.A. Eckstein, H.J. Noah, Examinations: Comparative and international studies. Oxford: Pergamon Press, 1992.

[22] A. Filed, Discovering statistics using SSPSS. Dubai: Sage Publication, 2009.

[23] J. R. Fredriksen, A. Collins, "A systems approach to educational testing," Educational Researchers, vol. 18, pp. 27-32, 1989.

[24] D. Freeman, K. E. Johnson, "Reconceptualizing the knowledge base of language teacher's education," TESOL Quarterly, vol. 32, pp. 397- 417, 1998.

[25] G. Fulcher, F. Davidson, Language testing and assessment. London: Routledge, 2007.

[26] C. Gipps, Beyond testing: Toward a theory of educational assessment. London: Falmer Press, 1994.

[27] H. M. Hazi, "The teacher evaluation supervision dilemma: A case of entanglements and irreconcilable differences," Journal of curriculum, 1994.

[28] S.P. Heyneman, "Use of examinations in developing countries: Selection, research and education sector management," International Journal of Education Development, vol. 7, pp. 251- 263, 1987.

[29] S.P. Heyneman, A. W. Ransom, "Using examinations and testing to improve education quality," Educational Policy, vol. 4, pp. 177-192, 1990.

[30] D. Holt, "Assessing success in family projects: Alternative approaches to assessment and evaluation," Washington: Center for Applied Linguistics, 1994.

[31] A. Hughes, Backwash and TOEFL 2000, Unpublished manuscript, University of Reading, UK, 1993.

[32] T. Johansson, T. Kroksmark, "Teachers' intuition-in-action: How teachers experience action," Reflective Practice, vol. 5, pp. 357-381, 2004.

[33] F.Korthagen, A. Vasalos, "Levels in reflection as a means to enhance professional growth," Teachers and Teaching, vol. 11, pp. 47-71, 2005.

[34] X.J. Li, "How powerful can a language test be?" The Journal of Multilingual and Multicultural Development, vol. 1, pp. 393-404, 1990.
[35] S. Messick, "Validity and washback in language testing," Language testing journal, vol. 13, pp. 241-255, 1996.
[36] M. M. McFerren, Certification of language educators in the United Sates. Los Angeles: University of California Press, 1988.
[37] R. L. Mosher, D. E. Purpel, Supervision: the reluctant profession. New York: Houghton Mifflin, 1972.
[38] B. Morris, Objectives and perspectives in education: Studies in educational theories. London: Routledge, 1972.
[39] G. Murdoch, " A progressive teacher evaluation system," The English teaching forum, vol. 36, pp. 2-11, 1998.
[40] D. Nunan, C. Lamb, The self directed teacher: Managing the learning process. Cambridge: CUP, 1996.
[41] J. Pallant, SPSS Survival Manual: a step by step guide to data analysis using SPSS. England: Open University Press, 2005.
[42] M. C. Pennigton, "Directions for faculty evaluation in language education," Language, Culture and Curriculum, vol. 2, pp. 167-193, 1989.
[43] W. J. Popham, "The merits of measurement driven instruction," Phi Delta Kappa, vol. 68, pp. 679-682, 1987.
[44] W. J. Popham, "The dysfunctional marriage of formative and summative teacher evaluation," Journal of Personnel Evaluation in Education, vol. 1, pp. 269-73, 1988.
[45] J. C.Richards, W. A. Renandya, "Methodology in language teaching: An anthology of current practice," Cambridge: CUP, 2002.
[46] J. Ruddock, Teaching as an art, teacher research and research based teacher. Anglia: University of East Anglia, 1984.
[47] D.A. Schon, The reflective practitioner: How professionals think in action. New York: Basic Books, 1983.
[48] E. Shohamy, The power of test: The impact of language testing on teaching and learning. Washington: National Foreign Language Center, 1993.
[49] E.Shohamy, S.Donista-Schmidt and I. Ferman, "Test impact revisited: Washback effect over time," Language Testing, vol. 13, pp. 298-317, 1996.
[50] M.L. Smith, "Meanings of test preparation. American Education Research Journal", vol. 28, pp. 521-542, 1991.
[51] B. Spolsky, Measured words. London: Oxford University Press, 1995.
[52] S. S. Stodolsky, "Teacher evaluation: The limits of looking," Educational Researchers, vol. 13, pp. 11-22, 1984.
[53] M. Swain, Language scale communicative testing. Hong Kong: Pergamon Press, 1985.
[54] L. Van Lier, Interaction in the language curriculum: Awareness autonomy and authenticity. New York: Longman, 1996.
[55] P. E. Vernon, The measurement of abilities. London: University of London Press, 1956.
[56] L. Vygotsky, Thought and language. Cambridge: MIT Press, 1962.
[57] M. Williams, R. Burden, Psychology for language teachers: A social constructivist approach. London: CUP, 1997.
[58] G.B. Workman, "Cohesion of conflict?" Institute of language in education journal, vol. 3, pp. 82-102, 1987.