On the Interaction of Test Washback and Teacher Assessment Literacy: The Case of Iranian EFL Secondary School Teachers

Razavipour Kiomrs (corresponding author)

Department of Linguistics and Foreign Languages, Shiraz University, Shiraz, Iran

Tel: 98-916-344-3386 E-mail: kioreem@hotmail.com

Riazi Abdolmehdi

Department of Linguistics, Faculty of Human Sciences, Macqarie University, Sydney, Australia

Rashidi, Naser

Department of Linguistics and Foreign Languages, Shiraz University, Shiraz, Iran

Abstract

It is by now well established that teacher characteristics play a major role in the way high stakes tests impact education (Alderson and Hamp-Lyons 1996). What remains an open question, however, is specifying the type of characteristics that have the potential to moderate the backwash effects of tests. This study was designed to isolate the effects of teachers' assessment literacy in moderating the washback effects of summative tests in the EFL context of Iran. A test of assessment literacy and a questionnaire on English language teaching practices were administered to 53 EFL secondary school teachers**.** Results show that teachers are suffering from a poor knowledge base in assessment and no matter how assessment literate they are; they do tailor their English teaching and testing to the demands of external tests**.** However, more assessment literate EFL teachers seem to be more likely to include non-washback practices in their English teaching. The implications for teacher training and teachers' professional development programs are then discussed.

Key words: washback, impact, assessment literacy

1. Introduction

The notion of test washback has been around for a long time. Educators seem to have always been aware of the effects of tests on educational programs, teachers, and learners. Only very recently, however, have scholars been urged to put the notion to serious empirical investigations. Findings now suggest that the washback phenomenon does exist but it is too complex to lend itself to simple experimental designs (Wall and Alderson 1993, Alderson and Hamp-Lyons 1996). It is also evident that numerous personal and contextual factors interact in shaping the impacts of tests on classroom processes (Watanabe 2004, Alderson and Hamp-Lyons 1996,). Among the involved personal factors, teachers' characteristics have been found to go a long way in predicting the way education is aligned with the demands of high stakes tests. This study was designed to highlight the effects of one important teacher characteristics rarely addressed in language assessment in general and in washback studies in particular: teachers' assessment literacy. Unlike test washback, teacher assessment literacy has only recently found its way into the agenda of language assessment community. In this study, we embarked on studying the level of assessment literacy among Iranian EFL teachers and the manner in which it may moderate the negative backwash of summative, external examinations. To this end two major research questions were posed:

1) What is the extent of assessment literacy among Iranian EFL teachers in Iran secondary schools?
2) To what extent the washback effect of final examinations is moderated through teachers' rate of assessment literacy?

2. Literature Review

Testing is an integral part of every educational system. That is why evaluation is one of the necessary modules of each curriculum development program. Tests are originally designed to be at the service of learning and teaching (Davies 1990). Optimally, tests are to evaluate the outcomes of education *a posteriori*. However, tests have come to act beyond the original role they were given. With the advent of external tests, a reversal of roles has occurred in educational programs so as sometimes it is teaching which is at the service of testing. This role reversal is to be better conceived of as a continuum. Washback refers to the extent tests outmaneuver teaching (Hughes 1989, Shohamy, Donitsa-Schmidt, & Ferman 1996). A family of similar terms, with slight differences in shades of meaning, have emerged which all have in common a concern for the undesired or desired influences of tests on learning, teaching, and society. In general education, the terms impact, curriculum alignment, and consequential validity are better known than the terms washback and backwash which are frequently used in language education (Hamp-Lyons 1997). Studies in general education have it that high stakes tests lead teachers to waste the

instructional time, ignore higher-order thinking skills, teach to the test, and limit their focus to those learning areas and tasks that are likely to appear on the tests (Hamp-Lyons 1997, Mehrens and Kaminsky1989, Fredricksen and Collins 1989, Cheng and Curtis 2004). In language education, Alderson and Wall (1993) pioneered the journey of classically studying test washback. They studied the nature of English language teaching classes in a wide range of contexts in Sri Lanka prior to the introduction of an innovative test into the educational system of the country. They also set out to study the same processes after the test had been put into operation for a few years. They put forward 15 hypotheses accounting for all the possible aspects and dimensions of test washback (Watanabe 2004). An important finding of that study was that tests do affect what teachers teach but they are less likely to affect how they teach. They also found that teachers' approach to assessing their students' achievement is to a great extent a function of test washback, that is, teachers' assessment practices are one of those areas subject to the immediate impacts of tests. Shohamy, Donitsa-Schmidt and Ferman (1996) studied the washback of two tests in Israel: a test of Arabic and one of English twice with the interval of a few years. It was found that in addition to features in test design, there are other factors at play in determining the washback of a test. In particular, they found that over time the effects of a test may disappear or increase. Moreover, the prestige of each language in each society contributes to the enhancing or weakening of test consequences. Unlike the test of English, It was found that the test of Arabic failed to generate considerable washback because of the unpopularity of the Arabic language in Israel. Alderson and Hamp-Lyons (1996) investigated the washback of TOEFL and found lots of variations among teachers with regard to the degree and type of test washback. They commented that "our study shows clearly that the TOEFL affects both *what* and *how* [italics original]teachers teach, but the effect is not the same in degree or in kind from teacher to teacher"(p.295). Teachers' educational background, past learning experiences, beliefs about effective teaching and learning, and their attribution orientations have also been found to affect the washback phenomenon (Watanabe 1996, 2004). In summary, teachers' characteristics are of crucial importance in determining the extent and nature of test consequences. Of particular importance among such teacher features is their knowledge base in assessment or what has come to be known as assessment literacy.

Assessment literacy (henceforth AL) is the ability to understand, analyze and apply information on student performance to improve instruction (Falsgraf 2005, p.6). AL is vitally important for good teaching. Eckhout, Davis, Mickelson, and Goodburn (2005, p. 3) argue that good teaching is actually impossible in the absence of good assessment. Despite its crucial role in shaping the quality of teaching there is evidence that teachers universally suffer from poor assessment literacy (Volante and Fazio 2007). Several reasons have been suggested which conspire to deny teachers of an optimal level of AL. A commonly-held belief is that if an individual knows how to teach a language, he or she knows how to assess the product and the process of language learning as well (Spolsky1978, cited in Jafarpour 2003). Such common mistaken beliefs contribute negatively to further neglect of teachers' knowledge base in language assessment. The intimidating appearance of assessment, its being the only branch of applied linguistics inundated with numbers and figures is yet another reason (Bridley 2001). Traditional delivery approaches to teaching assessment courses both in in-service and pre-service programs have also resulted in teachers' alienation from assessment issues (Inbar-Lourie 2008). Given the attested role of teachers' factors mediating the negative and positive washback of examinations, the specific part AL may play in that interaction remains a lacuna in the language assessment literature.

3. Method

3.1 Participants

53 EFL secondary school teachers (13 women and 40 men) sampled from three major provinces of the country participated in the study. As the study was done during summer when schools are closed and teachers can hardly be accessed, a random sampling approach was not feasible. Eight of the participant teachers held M.A.s in TEFL and the rest held B.A.s in either translation, English literature, or TEFL. The relatively smaller sample of female teachers is due to the fact that generally there are fewer female English teachers in the country so that, despite the illegitimacy of male teachers teaching in female schools in Iran, coeducation is against the country's laws, some schools have to employ male English teachers (boys and girls study in separate schools and are taught by the same-gender teachers). The other reason is that women can hardly manage to get engaged in teaching summer courses because of their traditional prime responsibilities at home.

3.2 Instruments

Plake and James (1993) test of assessment literacy was used to measure EFL teachers' knowledge base in assessment. The test consists of a set of 35 multiple choice questions followed by some self-assessment questions relating to the extent teachers see themselves proficient in English, prepared for language teaching, and competent in language assessment. The former section includes items on various aspects of assessment including knowledge of

the terminology of assessment, assessment ethics, assessment results interpretation, and procedures of designing tests according to the local needs. Test items have a problem-solving nature so that to answer them correctly one needs to enjoy a solid, well-integrated knowledge of relevant assessment issues. Using Cronback's alpha, the reliability index of the test was .79. The second instrument was a Likert-scale questionnaire covering the most common or possible language teaching practices in the EFL context of Iran. To develop the questionnaire, the researchers drew upon their own experience of EFL teaching in Iran secondary schools, in-depth interviews with three EFL teachers which were recorded and transcribed, and the related literature. A pool of 50 items was piloted with 10 teachers. Analyzing their responses, several modifications were made to the original version: twenty items were dropped out on the grounds of being of little relevance, the wording of some was modified, and many items were reordered to enhance the validity of responses. The questionnaire in its final format consisted of two categories of items. One group of items related to those language teaching activities which were perceived to be directly targeted at helping students pass the summative tests given at the end of the year. We call this group of items 'washback items' (items 1, 2, 7, 17, 18, 19, 20, 21, 24, 27, 29). The second category of items covered language teaching activities or tasks that were not primarily designed to help learners increase their scores on final tests. Rather, they were teaching practices in line with current perspectives on communicative language teaching that could potentially help students increase their communicative language ability irrespective of the type of examination administered. Teacher-constructed achievement tests were another source of data. Teachers were asked to submit copies of tests that they had recently developed. The requirement was that the submitted tests should be entirely made on their own without copying from a bank of items or borrowing test papers from a colleague.

3.3 Procedures

Teachers were met in the private language schools where they were teaching or the colleges where they were studying during summer. They were asked to respond to the questionnaire items before they start doing the assessment literacy test. One of the researchers stayed with the participants during the time they were answering the test and the questionnaire to answer their possible questions. As to the teacher-made achievement tests, only 21 were collected and the rest of teachers failed to submit tests for reasons that will be discussed in the next sections. The AL test papers and the likert-scale questionnaires were scored and the teacher-made tests were analyzed and compared against the framework of final exams. The two categories of questions were scored separately so that each teacher had two scores on the questionnaire: a score for bad language teaching practices or negative washback practices and another score for good language teaching practices or positive and non-washback influences. As it is usually the case with backwash studies, descriptive statistics and correlations were used to analyze the data.

4. Results

Table 1 shows the descriptive statistics for teachers' self-assessments of their overall preparation for teaching English, for assessing their students' performance, and their general proficiency in English**.** As the table illustrates teachers had a higher evaluation of their own preparedness for teaching English (Mean = 3.14) compared to their self-assessment of their English language proficiency or their evaluation of their own readiness to do proper language assessments. It seems that Iranian EFL teachers are, to some extent, aware of their own weak assessment background as their self-assessment of their own AL yielded the minimum score compared to the other two scores (Mean = 2.93). Generally, however, teachers seem to believe that they are well 'prepared' for English teaching as well as its assessment requirements given that the maximum possible rating was four on these two scales (preparedness for teaching and preparedness for doing assessment).

Insert Table 1 Here

Table 2 illustrates the descriptive statistics for participants' scores on the AL measure. The maximum possible score is 35. As the table demonstrates, the maximum score obtained is 20 which shows that the most assessment literate teacher had barely managed to answer more than half the items right. The mean is ten out of 35 which is indicative of a very poor performance on the AL measure. Overall, then, it is obvious that teachers do suffer from a poor assessment knowledge base. This is counter to the way they rated their own competence in language assessment, a mean score of 2.93 out of four.

Insert Table 2Here

It was indicated in the previous section that less than half the participant teachers agreed to submit a copy of their self-made tests. When asked about the reasons for their decline to submit a copy of their self-made tests, some responded that they never took their time to construct new tests and the rest pointed that they administered borrowed test papers from colleagues. By itself, it is an alarming finding because given the importance currently assigned to the role of assessment in promoting language learning, it is unfortunate to find such an indifferent attitude towards assessment among EFL teachers.

Normally, the English final examinations in Iranian secondary schools start with a spelling test followed by a set of fill-in-the-blanks vocabulary items. A few matching synonym or antonym items then follow. Other sections include multiple choice discrete-point grammatical items, a few matching items related to conversations included in textbooks-they are called language functions- and the test ends with a passage for reading comprehension followed by two true/false items and a few open- ended or multiple choice items. The analysis of the achievement tests made by teachers showed severe washback of final examinations as even minor deviations from the format and content of final tests were rare. This was a confirmation of what teachers indicated in the interviews. One of them stated that: "I may sometimes increase the number of vocabulary items and reduce the grammatical ones because I personally do not believe in the usefulness of grammar, however, I am careful to observe the overall framework of the final examinations". The obvious washback effect on teacher-made tests may be, at least partially, induced by the fact that teachers are externally monitored for their conformity with the external test formats. Another teacher said, "I know of no other method of testing, I think final examinations are perfect and standard ways of assessing English knowledge of students". The above quote clearly shows that Iranian EFL teachers are not aware of the many options they have at their disposal for assessing the communicative language ability of their students. This lack of knowledge results in failure to practice which in turn contributes to the further deterioration of teachers' competence in language assessment.

The analysis of teachers' responses on the questionnaire shed further light on the way they are impacted by final tests as well as the interaction of assessment literacy and test washback. Descriptive statistics for teachers' responses on the washback items speaks to the heavy influence of final examinations on the way teachers teach English (mean=40, SD=5.25). The maximum possible score was 55 as there were 11 items. This high score clearly shows that Teachers with various levels of AL tailored their teaching to the demands of final examinations. To determine the extent to which teachers with various AL scores adjust their teaching to the requirements of final tests, a correlation was run between teachers' AL scores and their score on the 11 washback items. We expected a reverse correlation between teachers' assessment literacy scores and the degree of their exam-like language teaching. The result did not, however, confirm our expectation since no negative correlation was found between AL scores and the scores of the washback items of the questionnaire. This lack of correlation indicates that regardless of their assessment literacy level, all teachers do experience the washback of final examinations and the amount of the influence is enormous as indicated by descriptive statistics (the mean was 40.11 out of maximally possible 55). We ran a second correlation between teachers' AL scores and their scores on the non-washback items of the questionnaire. Although the correlation coefficient was not very high, it was, however, significant at .05.

Insert Table 3 Here

This correlation shows that better assessment literate teachers, although equally try to meet the demands of final exams- they are to the same extent subject to the negative washback of final tests- they do, however, add variety to their language teaching through employing language teaching strategies which are more supported by current theories of communicative language teaching and learning. Giving more attention to higher-order thinking processes in teaching reading, employing more pair-work and group-work communicative activities, and focusing more on productive tasks in writing and speaking are typical of such non-washback activities. One possible explanation for the association between teachers' AL scores and their scores on the non-washback scale of the questionnaire is that it is teachers' English proficiency not their assessment literacy which explains the correlation or possiblely a combination of both English proficiency and assessment literacy. We ran a correlation between teachers' self-assessment scores of their own English proficiency and their scores on the AL measure. No meaningful relationship was revealed, giving us more trust, but not absolute trust, in the power of assessment literacy in affecting EFL teaching quality. It should also be mentioned that teachers, on average, assessed their own English proficiency as 'good'. The above two points remove part of the doubt for attributing either teachers' assessment literacy scores or communicative language teaching practices to higher English language proficiency. Nevertheless, in the absence of more solid experimental evidence catering for all possible factors it is hard to think of any casual relationships between teachers' competence in language assessment and the nature of their English teaching. Moreover, there seemed to be a negative correlation between teaching experience and knowledge base in assessment; the correlation coefficient was not significant however. It would be plausible to think that under the influence of standardized tests the more experienced teachers have lost a larger portion of their assessment literacy being exposed for a longer time to the paralyzing effects of tests.

5. Discussion and Conclusion

We can posit with a moderate degree of certitude that Iranian EFL teachers' knowledge base in assessment is far below the satisfactory level. On closer inspection of teachers' responses, it was found that more than on-third (19) of teachers could not recognize the appropriate definition of 'reliability' in a multiple choice item (the second item in

the AL measure). There were even a couple of teachers who failed to even answer one single item correctly. The teacher with maximum AL score only slightly answered more than half the items right. All these pieces of evidence show that Iranian EFL teachers of secondary schools have a very poor knowledge base in language assessment.

The way teachers conduct their assessments is in close alignment with final standardized examinations. The large number of teachers who do not construct tests at all is by itself another sad finding. It indicates that they are minimally sensitive to or aware of the importance of assessment in EFL teaching. This lack of practice in communicative assessments and following the fixed, traditional forms of assessment, in the long run, results in further deterioration of their assessment competence.

Teachers with various competencies in assessment model their teaching after the pattern of final examinations in trying to help their students succeed on such tests. Pragmatically, there seems to be nothing wrong with the approach in so far as the success of both students and teachers is judged on the basis of students' performance on such tests. More assessment literate teachers, however, seem to, in addition to trying to help their students achieve success on final tests, invest more in original language learning tasks not targeted directly at raising students' scores on summative tests. An interesting finding though it may seem, we do stay cautious as to passing with any certitude any casual relationship between the quality of EFL teaching and teachers' knowledge base in assessment. Both internal to the study and external factors may account for the association of higher scores on the questionnaire and the AL measure. We may legitimately argue that the significant correlation is likely to be a function of both English proficiency and assessment literacy or the former alone since the test was not translated into Persian(although this was not verified in our analysis of teachers' scores on the self-assessment scale of proficiency and their scores on the AL) . Nevertheless, it is still plausible to think that teachers with poor language proficiency are less likely to employ teaching activities which demand higher levels of communicative language ability. In addition, although self-assessments are praised for their potential merits of inducing positive washback like learner autonomy, their validity as true measures of proficiency is difficult to sustain given the numerous factors that may affect one's judgment of their ability. Had we measured participant teachers' English proficiency through a test like IELTS or TOEFL, rather than through their own self-ratings, we would have been in a better position to comment on their language proficiency as well as their language assessment literacy and teaching practices. As washback studies, by their very nature, do not lend themselves to experimental designs to isolate the effects of separate factors, unless further evidence accrues, such findings should be approached with extreme caution.

The extremely low assessment literacy of EFL teachers observed calls for a thorough overhauling of both pre-service and in-service assessment training courses. As Inbar-Lourie (2008) asserts the traditional delivery approach to teaching assessment courses have proved to be futile. Assessment courses, both in-service and pre-service, should follow more down-to-earth approaches focusing more on assessment practices than on theoretical issues. Teachers' fear of assessmaent should be allayed through involving them in serious cooperative assessment practices. Iranian EFL teachers are never invited to practice their expertise in any serious large-scale assessment project. Moreover, easing teachers' access to local, national and international testing and assessment journals can contribute to the improvement of teachers' AL. Finally, modifying the structure of traditional, discrete-point final examinations and employing more communicative language testing can benefit the EFL teaching both by accomplishing positive washback and promoting teachers' knowledge base in language assessment.

References

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: a study of washback. Language Testing, 13(3), 280-297

Brindley, J. (2001). Language assessment and professional development. In Elder, C., Brown, A., Grove, E., Hill, K., Iwashita, N., Lumly, T., McNamara, T., O'Loughlin, K. (Eds). *Experimenting with uncertainty: Essays in honor of Alan Davies*. Cambridge: Cambridge University Press.

Cheng, L. & Curtis, A. (2004). Washback or backwash: a review of the impact of testing on teaching and learning. In In L.Cheng,Y.Watanabe, & A.Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp.19–36).Mahwah, NJ: Lawrence Erlbaum Associates, Inc

Falsgraf, C. (2005, April). Why a national assessment summit? New visions in action. National Assessment Summit. Meeting conducted in Alexanderia, Va. Retrieved September 2009 from: http://www. nflrc.iastate.edu /nva/word documents/ assessment_2005/ pdf /nsap_introduction.pdf

Eckhout, T.,Davis,S.,Mickelson, k., & Goodburn, A. (2005). A method for providing assessment training to in-service and pre-service teachers. Paper presented at the annual meeting of the Southwestern Educational Research Association, New Orleans, LA.

Fredricksen, J.R. and Collins, A.1989: A systems approach to educational testing. *Educational Researcher*, 18, 27-32.

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

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

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing*, 25(3), 385-402

Jafarpour, A. (2003). Is the test constructor a facet? *Language Testing*, 20(1), 57-87

Mehrens, W. A., & Kaminsky, J. (1989). Methods for improving standardized test scores: fruitfull, fruitless, or fraudulent? *Educational Measurement: Issues and Practice*, 8(1), 14-22

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

Wall, D., & Alderson, J. C. (1993). Examining washback: the Sri Lankan impact study. *Language Testing*, 10(1), 41-69

Watanabe,Y.(2004).Methodology in washback studies. In L.Cheng., Y.Watanabe, & A.Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp.19–36).Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Watanabe, Y. (2004). Teacher factors mediating washback. In L. Cheng,Y.Watanabe, & A.Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp.19–36).Mahwah, NJ: Lawrence Erlbaum Associates, Inc

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing*, 13(3), 318-333

Table 1. Descriptive statistics for teachers' self-assessment ratings

	N	Min	Max	Mean	Std. Deviation
Preparation for teaching	49	1	4	3.14	.61
Preparation for assessment	47	1	4	2.93	.76
English proficiency	52	1	5	3.03	.79

Table 2. Participants' scores on the AL test.

N	Minimum	Maximum	Mean	Standard Deviation
53	.00	20.00	10.0377	4.80369

Table 3. The correlation between participants' AL scores and their socres on the non-washback items of the questionnaire.

		AL	Non-washback
Assessment literacy (AL)	Pearson Correlation	1	.328*
	Sig. (2-tailed)	.	.0117
	N	53	53