International Journal of Instruction *October 2018 ● Vol.11, No.4*
e-ISSN: 1308-1470 ● www.e-iji.net *p-ISSN: 1694-609X*
pp. 717-732

Received: 25/04/2018
Revision: 16/07/2018
Accepted: 22/07/2018

Construct Ambiguity and Test Difficulty Generate Negative Washback: The Case of Admission Test of English Literature to Graduate Programs in Iran

Kioumars Razavipour
Dr., corresponding author, Shahid Chamran University of Ahvaz, Iran, *razavipur57@gmail.com*

Sayyed Rahim Moosavinia
Shahid Chamran University of Ahvaz, Iran, *moosavinia@yahoo.com*

Somayyeh Atayi
Shahid Chamran University of Ahvaz, Iran, *somayeh.ataei93@gmail.com*

High stake tests are known to drive curricula and affect teachers and learners in numerous ways. In applied linguistics, this has come to be known as test washback. Thus far, almost all washback studies have focused on national or international language proficiency tests. There are scarce, if any, research studies addressing the washback of high stakes tests of English literature. The current study examined the washback effect of the English Literature Module of the Admission Test of English Literature (ATEL) on test takers' learning and attitudes as well as on test preparation materials. For this purpose, 100 graduate students of English literature from Iranian state universities completed a questionnaire designed to measure the washback of ATEL. The collected data were analysed using descriptive statistics and t-test. The results revealed that the test significantly influenced both test takers' attitudes and their learning of English literature. More specifically, it was found that the test, being beyond the zone of proximal challenge, causes most test takers to forsake the entire English literature module in favour of the general proficiency module and to hold negative attitudes towards the test.

Keywords: washback, literary competence, construct ambiguity, test difficulty, English literature

INTRODUCTION

The influence of tests on education has attracted the attention of a wide range of stakeholders in recent decades. The power that tests have is so evident that often policy makers use them as the de facto levers of change for effecting policies and practices

Citation: Razavipour, K., Moosavinia, S. R., & Atayi, S. (2018). Construct Ambiguity and Test Difficulty Generate Negative Washback: The Case of Admission Test of English Literature to Graduate Programs in Iran. *International Journal of Instruction*, 11(4), 717-732. https://doi.org/10.12973/iji.2018.11445a

(Andrews, 2004; Elana Shohamy, 2011)). In applied linguistics, the influence of tests on language learning and teaching has been established as a separate strand of research, often termed as test washback or backwash. Research into the impacts of tests on education suggests that the washback phenomenon is highly complex and multifaceted (Wall & Alderson, 1993; Watanabe, 2004). Hence, the extent to which a test influences education and the stakeholders hinges on factors such as the test format and content (Hamp-Lyons, 1998), the stakes associated with the test (Alderson & Hamp-Lyons, 1996), the prestige of the content matter tested (E. Shohamy, Donitsa-Schmidt, & Ferman, 1996) and other macro and micro-contextual factors (Watanabe, 2004b). The existing, extensive literature on test washback is almost solely concerned with national or international English language tests, which seek to measure test takers' proficiency in English.

In regard to tests of literary competence, although suggestions have been made concerning the dangers lying in attempts at quantifying appreciation of literature (Cooper, 1971; Gaston, 1991), studies addressing the washback of tests of English literature on participants, processes, or outcomes of education in English literature in English as a Foreign Language (EFL) contexts remain scarce. This study aims to narrow this lacuna by investigating the washback of ATEL on test takers' preparation practices, attitudes, and motivation. Given that in current validity theories, a test's validity is judged not only by its precision in measuring the intended construct but also in its educational usefulness (Messick, 1996), the present study contributes to examining the consequential validity of ATEL. The consequential validity of tests becomes more serious when the test in question carries important consequences for test takers or other stakeholders. Being a gate-keeping test, very important stakes are attached to ATEL as it is the only channel Iranian students can seek entry to graduate programs of English literature at state universities.

REVIEW OF LITERATURE

Whereas concern over the consequences of tests for education is not a new phenomenon, it was only after the publication of Messick's well-known unified matrix of validity that did systematic research into test washback begin. By integrating test validity and ethics, Messick placed consequences of tests at the heart of validity theory. This was consistent with the dominant utilitarian theory of moral philosophy, according to which the ethicality of an act judged in view of its consequences. In other words, an ethical act is one which brings the most benefit to the maximum number of people (Davies, 1997; Kunnan, 2010). Accordingly, the degree to which a test brings about beneficial consequences should constitute, at least, one source of validity evidence for the test.

Test washback is known to be a complicated, multi-faceted phenomenon (Alderson & Hamp-Lyons, 1996; Wall & Alderson, 1993; Watanabe, 2004a) since the effect that tests exert on education is mediated by numerous factors. Building on previous models of test washback (Bailey, 1996; Hughes, 1993)), Watanabe proposed a comprehensive model of washback, which explicates washback in terms of its aspects, dimensions, and the mediating factors (2004). The aspects of washback refer to those components or variables in the education system that are influenced by a test. Learners, teachers,

parents, textbooks, learning and teaching are some of the main aspects of test washback. The dimensions of washback include specifity, intensity, length, intentionality, and value. Specifity refers to whether the influence of a test is general or specific. For instance, if students or test takers are pushed to study more in anticipation of an incoming test, the washback would in this case be placed on the general end of the continuum. The intensity dimension relates to the magnitude of the influence that the test has; whether the effects of the test are strong or weak. The length dimension refers to the duration of test effects. Thus, washback may last long or it may be short-lived as in the case where the effects of a test disappear immediately after test takers take the test. The intentionality dimension is about whether what happens subsequent to test administration is intended or unintended. Washback is believed to be intentional when tests are used to engineer desired changes in the education system (Andrews, 2004). Finally, the value dimension of test washback concerns with whether the consequences of a test are positive or negative. It is often assumed that intended washback effects are positive but consequences that are not intended might be positive or negative. Yet, some scholars have taken issues with this, arguing that the intentions of policy makers are not necessarily benign (Fulcher, 2009; Shohamy, 2001). Thus, scholars do not rule out the possibility that tests might be utilized to advance the agendas and interests of those in power, which would mean detrimental effects to the wider society. Therefore, scholars have called for critical language testing and democratic language assessment, which are believed to keep in check the power of tests (E. Shohamy, 2001). As washback is not an essentially linear process from test to education, mediating factors constitute the third component in Watanabe's model. Macro and micro contextual factors, test factors, personal factors, and prestige factors are at play in determining both the aspects and the dimensions of test washback.

Awareness as to the potential, political misuse of tests has alerted researchers to take test makers and users accountable for test consequences including washback (Hamp-Lyons, 1996). Others have argued that the ethical responsibility of testers and test users for the consequences of tests cannot be so open-ended and far reaching. To Davies, it is a heresy to place all the blame for test consequences on testers and test users (Davies, 2003). Instead, he takes a within-reason stance regarding the extent to which test consequences should be seen as the responsibility of testers.

Likewise, to Messick (1996), only influences that can evidentially be linked to the test should be considered as washback, which would in turn count as evidence of consequential validity or lack thereof. Messick maintains that a good test may bring about negative washback because of other factors in the educational system. On the contrary, a bad test is also likely to induce positive washback because of non-test factors. Alderson and Wall (1993) seem to agree with Messick on limiting the scope of washback to only those influences that can be evidentially attributed to the test. Accordingly, if test washback is in fact due to the mediating factors noted above, not exclusively because of test factors, such influences should be excluded from language testers' concerns and responsibilities. Yet, the direction that scholarship in test washback has taken over the past few decades seems not to be in keeping with this within-reason stance towards the scope of testers' responsibility for test consequences. In particular,

personal factors including both teacher and test taker factors have featured frequently in studies on test washback.

Test takers' understandings, beliefs, and perceptions of test content and uses have been shown to mediate various dimensions of test washback (e.g., (Qin, 2011, 2015; Samai & Mohammadi, 2017; Xie & Andrews, 2012; Zhan & Andrews, 2014) .There is evidence suggesting that psychological factors like motivation and perceived self-efficacy affect test takers' language learning directed at test preparation (Xie & Andrews, 2013). Using structural equation modelling, Xie and Andrews postulated a model based on expectancy-value motivation theory, according to which test preparation was the ultimate endogenous variable and test takers' perceptions of test design, of test value, and of expectation of success were exogenous variables. Acceptable indexes of fit were found between the postulated model and the data. Besides, it was found that students' perceptions of test design and value mediate test preparation. Additionally, expectation of success and test preparation were found to be correlated. More recently, washback to test takers' perceptions of test content and uses were studied by Razavipour, Gooniband Shoushtari, and Mansoori (2018). Framing their study within expectancy-value theory of motivation, they used a partial least squares structural equation modelling approach to determine whether differential washback occurs to test takers in different tertiary education institutions and if values accorded to the test and test takers' self-efficacy moderate test preparation. Findings from this study suggested that test takers' perceptions of test content explained the largest portion of variance in test takers' preparation practices. Overall, though there is a considerable body of work on the washback of language proficiency tests, far less has been done in the area of testing literary competence and achievement, which might be partly due to the perceived incompatibility of aesthetics and quantification (Gaston, 1991).

Reluctance to quantification and measurement characterize the field of literature. Gaston (1991) maintains that "measurement, it would appear, would be unkind to beauty. Quantification and appreciation rarely coexist easily" (p. 11). Similarly, Hanauer (1996) states that "from a historical point of view the field of literature does not have a tradition of systematic test construction or evaluation" (p. 143). Despite this perceived incompatibility of literature and measurement, there has been awareness of the impacts that tests might have on literature education. Half a century ago, Cooper (1971) warned against objective literary tests because of the detrimental effects they induce in the curriculum**. "**The danger, obviously, is that with tests limited to easily objectified and verifiable matters of literal-level comprehension and facts from author-biography and historical backgrounds, literary study in the classroom will be just so limited" (1971, p. 18). Similarly, there is evidence that standardized high stakes testing of literature reduces literature instruction to reading comprehension instruction (Beach, 2014).

Hanauer (1996) justifies the noted incompatibility by explaining the tension between validity and authenticity on one hand and reliability on the other. He states that a valid test of academic literary competence should allow for the measurement of multiple literary interpretations. It should also assess both the product and the process of literary interpretation. For these reasons, he calls for performance based tests of literature. For

Hanauer (1996), such measures would promote approaches to the teaching and learning of literature that are educationally defensible. Hanauer (1996) empirically showed that a literary test based on a substantive theory of literary competence allows for the diagnosis of problem areas in learning literature. Using a limited sample, Hanuer showed that a rating scale developed according to a theory of literary interpretation can yield ratings that are as reliable as traditional holistic scoring of literary interpretation.

Yet, as Hanauer (1996) acknowledged, these assessments, despite their advantages, potentially add to the construct-irrelevant variance of test results. As a case in point, English language learners in the US consistently score lower on literature tests than their native, American counterparts (Bernhardt, 2005, cited in Beach 2014). To probe into the relevance of language proficiency in achievement tests of literature at tertiary level, Razavipour and Moosavinia (2017) conducted a survey study investigating the views of undergraduate students in English and Persian departments about the extent the constructs language proficiency should constitute part of the literary competence construct . It was found that the majority of students perceive of language proficiency as irrelevant to the construct of literary competence. Yet, disentangling the two, linguistic competence and literary competence, does seem to be an enormous challenge.

In one of the few empirical studies investigating the effect of standardized tests on teaching literature, Anagnostopoulos (2003) found that the extent to which state-mandated tests affect teachers' approaches to teaching literature was a function of two factors. One was the degree of convergence between teachers' teaching philosophy and the conceptions of literature implied in the tests. The other factor determining the washback of state mandated tests on teachers was teachers' curricular power: teachers with more curricular power were found to be less likely to change their teaching orientations because of the tests.

With the exception of a few studies, the literature reviewed above is almost exclusively focused on the washback of tests which seek to measure general language proficiency. Relatively, little research attention has been given to how high stake tests of literature might influence test takers' learning of literature and their attitudes and motivation. The few studies addressing washback of literature tests have had as their focus washback to teachers of literature and all have been carried out in the context of the U.S, where English is the official language of instruction and communication. Aimed at narrowing the noted gap, the current study sought to explore the washback of the Admission Test of English Literature (ATEL). Hence, the present study contributes to the existing slim literature on testing English literature in that first, it has as its focus the test takers, who are the most immediate stakeholders affected by tests, and secondly, it focuses on testing literature in an EFL context, to which extrapolation of research findings from inner circle countries should be done with extreme caution.

ADMISSION TEST OF ENGLISH LITERATURE (ATEL)

ATEL is annually administered nationwide during May by the Iranian National Organization of Educational Testing (NOET). The test is used to screen out candidates seeking admission to graduate programs in English literature at national universities. As the number of national universities offering graduate programs in English literature is

limited in the country, prospective students must engage in fierce competition to be granted a seat in such programs. As such, very high stakes are associated with ATEL and it presumably takes intensive preparation to succeed at it.

In terms of content, ATEL has two components: a general proficiency component (60 items) and a specialized literary component (60 items). All test items are of the multiple choice type and test takers must complete the test in no longer than 150 minutes. The focus of the present study is on the latter component, specialized literary component, which comprises of three major topic areas, namely, Literary Criticism, History of English Literature, and Literary Genres. Each of these topic areas is tested via 20 multiple choice test items. Each item on the specialized module carries a weight of three whereas each item on the general module carries a weight of two. Hence, given this differential weighting, one expects ATEL test takers to give priority to preparing for the specialized module.

METHODS

Participants

Selected through a convenience sampling approach, a total of 100 M.A students of English literature participated in this study. At the time of collecting data for the study, participants were all graduate students of English literature across eight Iranian state universities. Participants included both male (N = 43) and female (N = 57) students and their participation was on a voluntary basis. Given that direct access to all participants was not possible, participants were accessed either in person or electronically, via email and Telegram application. Table 1 summarizes information about the participants.

Table 1
Demographics of the participants

University	Number	2017	2016	Female	Male
Urmia University	8	3	5	6	2
Lorestan University	5	3	2	3	2
Shahid Chamran University of Ahvaz	16	8	8	10	6
Alzahra University	14	7	7	14	0
Kharazmi University	17	9	8	7	10
Allameh Tabatabai University	12	5	7	2	10
Shahid Beheshti University	14	8	6	6	8
Tehran University	14	9	5	9	5
Total	100	52	48	57	43

Instrumentation

A questionnaire, developed by the researchers for the purpose of this study, constituted the main data collection instrument. The questionnaire was developed and written in Farsi in order to neutralize the effect that differential English proficiency might have on participants' comprehension of questionnaire items. In order to develop the items of the questionnaire, 20 M.A students from Shahid Chamran University of Ahvaz were interviewed in groups of three or two. Eight interviewees had been admitted to their M.A programs in 2015, 6 in 2016 and 6 in 2017 academic year.

As the aim of the interviews was to generate the item pool for the questionnaire, the interviews were not structured. Rather, we set out with broad questions regarding how participants had gone about preparing for ATEL, what preparation behaviour they had utilized, what advice they would give to future test takers, what sources they had found useful, and what skills it entails to perform well on the test. We also inquired about participants' motivations and attitudes towards the test. The research questions of the study and our review of the related literature informed the questions asked during interviews.

The third author held a total of seven interview sessions with the twenty participants. With the exception of one interview session which was held with two participants, in each of the other six interview sessions, three participants took part. As is the case with focused group interviews, there is always the possibility that one participant may dominate the interview or the flow of talk might go in a direction not of relevance to the aim of the interview. To prevent such threats to the validity of the collected data, the interviewer would intervene in a non-obtrusive manner whenever she felt that the interview was not proceeding in the desired direction or if one participant failed to contribute to the interview or if a participant dominated the talk for long stretches of time.

Data saturation was reached with the first seventeen participants as the final three participants did not provide any new information; hence, they were dropped from our final analysis. The interviews were audio recoded and simultaneous notes were also taken during interviews. The interviews were transcribed, coded, categorized and the most frequently occurring themes were chosen as possible questionnaire items.

In order to examine the psychometric properties of the questionnaire and to remove any remaining ambiguity regarding the clarity and characteristics of the items, the questionnaire was administered to another 29 participants from Chamran University, 8 males and 21 females. The questionnaire consisted of three sections, each measuring a separate construct, namely, test takers' learning of English literature, test takers' attitudes, and test preparation materials. The questionnaire items were particularly developed with the purpose of eliciting information on:

(a) Test takers' learning of English literature directed at ATEL (items 1-28)

(b) Test takers' attitudes towards ATEL (items 29-35)

The wording of the responses varied according to the semantics of each item. As such, in cases where the item was about the frequency of using a certain strategy, the options were not about the strength of agreement with the item but about the frequency of use. It should also be noted that for items 36-39, which asked participants about the preparation materials they had used, participants could tick more than one choice. Although the preparation sources that served as options in the noted items were derived from prior interviews, space was provided for participants to list sources they had used for test preparation, which were not included in the options,. It took the participants around 20 minutes to fill out the questionnaire.

A Cronbach's Alpha of .90 indicated that the questionnaire enjoyed a high level of internal consistency. As to the validity of the instrument, in addition to piloting, removing ambiguities arising from the wording of items, an expert's advice was also sought, leading to revisions in wording, elimination of some items, and addition of others.

Data Analysis

All items were of the Likert type format with six choices ranging from 'strongly agree' (1) to 'strongly disagree' (6). As such, higher scores on each item denote stronger disagreement with the proposition. Given the nature of the study, data were analysed using descriptive and inferential statistics via SPSS, version 18. In particular, mean and standard deviations of individual items were computed to explore test takers' perceptions of and test preparation practices for ATEL. In regard to inferential statistics, one-sample t-test was used to examine the degree to which participants' responses deviated from the neutral value of 3.5 (the mathematical average of the six Likert options). Since t-test makes distributional assumptions about the data, skewness and kurtosis values were examined to make sure the data do not violate such assumptions.

FINDINGS

This section summarizes the results in two rather detailed tables. Table 2 contains the descriptive statistics, mean and standard deviation, and Table 3 gives the results of one-sample t-tests run for each item on the questionnaire.

Table 2
Descriptive statistics for test taker learning and attitude items in the questionnaire

Items	Mean	SD
1. I studied for the Literature MA exam.	4.42	1.29
2. Studying Abrams' literary terms was instrumental in getting a high score in *figures of speech.*	3.92	1.43
3. Note taking from Abrams was effective in helping me gain a high score in *figures of speech*.	4.16	1.41
4. Making flash cards from Abrams was effective in gaining a high score in *figures of speech*.	4.28	1.60
5. Studying literary types was effective in gaining a high score in *literary criticism*.	3.52	1.54
6. Application of literary types to literary pieces was effective in gaining a high score in *literary criticism*.	3.92	1.57
7. Note taking from the relevant literary books was effective in gaining a high score in *literary criticism.*	3.56	1.42
8. Summarizing Norton's book was effective in gaining a high score in *literature history*.	3.40	1.52
9. Studying Norton's translation was effective in gaining a high score in *literature history*.	4.40	1.69
10. Studying translated pieces was effective in gaining a high score in *literature history*.	4.27	1.70
11. Watching films was effective in gaining a high score in *literature history.*	4.06	1.58
12. Although studying literary types was very time consuming for me, I studied it for success in the exam.	3.90	1.52
13. Studying Perrin's S*tructure, Sound and Sense* book was effective in gaining a high score in *literary types*.	4.15	1.62
14. The exam questions were beyond my English literature knowledge and thus decreased my self-confidence.	3.11	1.68
15 I owe my good performance to my quality undergraduate program.	3.23	1.54
16. I was shocked with the questions I was not familiar with in the exam.	3.53	1.45
17. The difficulty of the exam drained my motivation to respond to the specialized questions.	3.10	1.60
18. The unpredictability of exam sources decreased my hope for success in the exam.	3.34	1.63
19. I was admitted to the graduate program mainly because of my good general English proficiency.	2.45	1.33
20. Most M.A students of English literature have poor literature knowledge but high general English proficiency.	3.07	1.43
21. Taking part in M.A test preparation courses played a big role in my success.	4.59	1.54
22. The fact that exam sources are not specified deters many students from pursuing an M.A in English literature.	3.06	1.48
23. For their M.A, undergraduate literature students shift to TEFL, Translation Studies and Linguistics	3.06	1.52

because the exam sources are not specified for an M.A in literature.		
24. The wide range of sources for the exam can be a motivating factor for studying different sources.	3.82	1.43
25. The wide range of sources for the exam confuses test takers.	2.76	1.42
26. Studying Mahan preparation books for the MA exam was effective in gaining a high score in the exam.	5.01	1.32
27. Studying Modarresan Sharif preparation books for the MA exam was effective in gaining a high score in the exam.	4.61	1.48
28. Studying questions from recent MA exams was effective in gaining a high score in literature questions.	3.83	1.70
29. Admission to the M.A program was important for me because I want to get a PhD.	2.63	1.59
30. Admission to the M.A program was important for me because it increase my chances of employment.	2.81	1.48
31. Admission to the M.A program was important for me because I wanted to get a Master's degree.	2.93	1.51
32. Admission to the M.A program was important for me because I want to get a scholarship.	3.39	1.71
33. I was confident that I would have a good performance in the specialized literature section of the exam.	3.30	1.63
34. If I had studied better, I could have had a better performance.	3.07	1.43
35. If I had worked harder, I could have had a higher score in specialized literature section.	3.31	1.58

The reasoning behind conducting t-tests was that descriptive statistics does not tell us the degree of confidence we can have in the results; hence recourse to inferential statistics. To determine the strength of the results provided in Table 2, one-sample t-tests were run. Table 3 gives the outcome of one-sample t-test with a neutral value of 3.5, computed as the mathematical average of the six points on the Likert scale instrument used in data collection. Given the manner numerical values were assigned to points on the Likert scale, negative t-values in Table 3 denote agreement and positive t-values point to disagreement with the statement.

Table 3
One sample t-test results

	Test Value = 3.5					
	T	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
1.	7.097	99	.000	.92	.6628	1.1772
2.	2.930	99	.004	.42	.1356	.7044
3.	4.650	99	.000	.66	.3784	.9416
4.	4.868	99	.000	.78	.4621	1.0979
5.	.130	99	.897	.02	-.2857	.3257
6.	2.668	99	.009	.42	.1076	.7324
7.	.421	99	.674	.06	-.2225	.3425
8.	-.656	99	.513	-.1	-.4024	.2024
9.	5.295	99	.000	.9	.5627	1.2373
10.	4.517	99	.000	.77	.4317	1.1083
11.	3.527	99	.001	.56	.2449	.8751
12.	2.619	99	.010	.4	.0969	.7031
13.	3.990	99	.000	.65	.3267	.9733
14.	-2.320	99	.022	-.39	-.7236	-.0564
15.	-1.742	99	.085	-.27	-.5775	.0375
16.	.206	99	.838	.03	-.2596	.3196
17.	-2.492	99	.014	-.4	-.7185	-.0815
18.	-.979	99	.330	-.16	-.4843	.1643
19.	-7.858	99	.000	-1.05	-1.3151	-.7849
20.	-2.992	99	.004	-.43	-.7152	-.1448
21.	7.057	99	.000	1.09	.7835	1.3965
22.	-2.968	99	.004	-.44	-.7342	-.1458
23.	-2.876	99	.005	-.44	-.7435	-.1365
24.	2.225	99	.028	.32	.0346	.6054

25.	-5.203	99	.000	-.74	-1.0222	-.4578
26.	11.358	99	.000	1.51	1.2462	1.7738
27.	7.484	99	.000	1.11	.8157	1.4043
28.	1.941	99	.055	.33	-.0073	.6673
29.	-5.439	99	.000	-.87	-1.1874	-.5526
30.	-4.656	99	.000	-.69	-.9840	-.3960
31.	-3.768	99	.000	-.57	-.8701	-.2699
32.	-.643	99	.522	-.11	-.4495	.2295
33.	-1.222	99	.224	-.2	-.5246	.1246
34.	-2.992	99	.004	-.43	-.7152	-.1448
35.	-1.197	99	.234	-.19	-.5049	.1249

As noted earlier, higher values in Table 2 denote less agreement with the propositions expressed in each item. This being so, the general message from Table 1 is that despite the high stakes attached to it, ATEL does not induce much preparation as the high disagreement in item one indicates (M=4.42, SD=1.29). Test preparation does not seem to be effective in turning things around in helping candidates with the specialized module of the literature exam. This inefficiency holds true about both test preparation materials and materials not being specifically test directed. The corresponding t value in Table 3 indicates that this lack of preparation for the specialized section of ATEL is far beyond a chance incident; t (99) = 7.097, p= .0001.

Items related to test preparation materials and preparation courses (21, 26, & 27) clearly indicate that when it comes to preparing for the specialized module of ATEL, preparation materials and courses do not seem to be of much help (M = 4.59, SD = 1.50; M = 5.01, SD = 1.32; M = 4.61, SD = 1.48; respectively). The related t-values in Table 3 show that the observed disagreement is significant. All the three observed values are significant at .0001, pointing to overwhelming agreement among participants that preparation courses and materials are not of much help in preparing for ATEL.

Similarly, items related to routine sources (items 4, 9, 10, 11, 13,) not specially directed at the exam, were found to be of little use in helping candidates with the test items (M = 4.28, SD = 1.60; M = 4.40, SD = 1.69; M = 4.27, SD = 1.70; M = 4.06, SD = 1.58; M = 4.15, SD = 1.62; respectively). The probability that the observed values might be due to chance occurrence is indeed very low as the t-value for item 11 is p = .001 and it was even smaller for the rest four items (p = .0001). This overwhelming disagreement with regard to the effect of default literary textbooks in helping candidates prepare for ATEL combined with a similar disagreement with preparation materials, noted earlier in this section, show that it is a rather curious situation; a test for which neither conventional textbooks nor test preparation materials can play a role in helping candidates master the test content.

The other general trend discernable from Table 2 pertains to the fact that the specialized module of the exam is beyond the candidates' level of ability, hurting both their self-confidence (Item 14: M = 3.11, SD = 1.68) and motivation (item 17: M = 3.10, SD = 1.60). As the mean scores show, candidates highly agreed with these two propositions. T-test values also indicate that their agreement was significantly high; for item 14, t (99) = -2.320, p = .02, and for item 17, t(99) = -2.492, p = .01.

Quite related to the above-noted theme, the other take away from Table 2 is that admission to M.A programs is seemingly a competition not based on literary competence but based on aspects of general English proficiency such as reading comprehension, vocabulary, and grammar. Learners agreed that their admission to the graduate program was mainly due to their good general English proficiency (M = 2.45, SD = 1.33) (Item 19). In fact, this was the item participants most highly agreed with. T-test value also confirmed this high agreement; t(99) = -7.858, p = .0001. This result points to the construct-underrepresentation of ATEL, where admission decisions are made not on the basis of an adequate coverage of content but on a partial sampling of the content universe. Similarly, Item 20 asked candidates to indicate their degree of agreement with the following statement: Most M.A students of English literature have poor literature knowledge but high general English proficiency (M = 3.07, SD = 1.43). One-sample t-test value further verified participants' agreement with the statement that ATEL is not functioning efficiently; t(99) = -2.992, p = .004.

Respondents also expressed profound dissatisfaction with the unpredictability and wide diversity of exam sources, which apparently discourages many students from pursuing an M.A in English literature (M = 3.06, SD = 1.48) (Item 22). T-test value for this item in Table 3 indicates that participants' agreement with the item is significantly high; t(99) = -2.968, p = .004.

Participants also remarkably agreed with the proposition that 'undergraduate literature students shift to TEFL, Translation Studies and Linguistics because the exam sources for ATEL are not specified' (M = 3.06, SD = 1.52) (Item 23). Agreement with this item was also significantly high; t(99) = -2.876, p = .005. They also strongly agreed with the idea in item 25 that the unpredictability of exam sources induces confusion among participants (M = 2.76, SD = 1.42). The strength of agreement with this item significantly departed from the neutral value too; t (99) = -5.203, p = . 0001.

In regard to candidates' motivation in taking the exam, participants appeared to have both so-called instrumental and intrinsic motivations. Getting a PhD (M = 2.63, SD = 1.59) or an M.A degree (M = 2.93, SD = 1.51) ranked specifically high among the reasons for taking the tests. Candidates also agreed that seeking employment was a major drive for their taking the exam (Item 29): M = 2.81, SD = 1.48. The observed t-values for all the three items addressing test takers' motivation were significant, indicating that the majority of participants strongly agreed with the propositions carried by the items.

DISCUSSION

The current study sought to probe into the washback that ATEL might have on test takers. Particularly, it sought to examine test takers' preparation practices as well as their attitudes towards and motivations for taking ATEL. The former question essentially translates into whether and how the learning of English literature is promoted or weakened in the process of preparing for ATEL.

Generally speaking, it was found that ATEL exerts a detrimental influence on the learning of English literature. In the first place, it was found that the test is beyond the

zone of proximal challenge; hence, lack of incentive and drive on the part of students to prepare for ATEL. It is known that for a test to induce positive washback, it should not be either too easy or too challenging (Alderson & Hamp-Lyons, 1996; Hamp-Lyons, 1998) because either way it would kill test takers' motivation to engage in proper test preparation. Secondly, it was found that most test takers harbor negative attitudes towards the test and give up on it, leading them to either focus on boosting their general English proficiency or opt for other language related fields such as Linguistics, Translation Studies, or TEFL. The former observation, the test being beyond the zone of proximal challenge, can in fact account for test takers' negative attitudes. Moreover, test takers strongly believed that neither the conventional literary sources nor commercial ATEL-directed materials were much helpful in preparing for ATEL.

It was also found that the majority of candidates had given up on studying English literature for the exam and had limited their test preparation to improving their general English proficiency so that they can compensate for their low scores on the literature module by obtaining high scores on the general test. This finding is consistent with expectancy value theory of motivation (Xi & Andrews, 2013). When a task is too challenging, the learners' expectation of success decreases to the point where they may completely give up on the task. The other factor known in the literature to mediate and inform test takers' test preparation is their understanding of the weight of the different components of a test (Qin, 2015; Zhan & Andrews, 2014). Zhan and Andrews (2014) using a diary study found that test takers' understanding of test weight mediated their study plans and strategies in preparing for the test. Similar findings were reported in the study by Qin (2015). Both of the above mentioned studies indicated that test takers allocate more time to test components that carry more weight in their final total scores. Nevertheless, in the present study, results were contrary to expectations as well as to previous research, in that participants invested more heavily in the General English section, which carries less weight than the Specialized English literature component. One possible reason for this outcome has to do with the complex interplay between test takers' perceptions of test construct, of test difficulty, and of test component weighting. It seems that perceptions of construct and of component difficulty override component weighting. In other words, even when a test component carries more weight, if it is perceived to be more difficult, it is sacrificed in favour of components that carry less weight but whose mastery seems more doable. This observation is consistent with Powers (1987), which found that test takers invest more time and resources in components that are more coachable. The General proficiency test is obviously more amenable to coaching than the specialized English component. In regards to the relationship between linguistic competence and literary competence, ATEL washback can be seen as rather negative because though language proficiency and literary competence are difficult, if not impossible to disentangle, there is evidence that the two constructs are not the same. In Razavipour and Moosavinia's study (2017), students across Persian and English departments expressed discomfort over factoring in linguistic competence in rating their literary achievement and competence. Thus whereas students endorse keeping the two constructs separate in achievement tests, when it comes to preparing for graduate programs of English literature, they choose to focus exclusively

on boosting their linguistic competence; hence, threatening both the construct and the washback validity of ATEL.

Another finding of the present study was that the diversity and unpredictability of the sources based upon which test items are derived is one of the major reasons driving students away from seeking admission to graduate programs in English literature. Yet, such complaints might not be considered legitimate if we think of ATEL as a test of literary competence that is not necessarily designed based on any specific set of textbooks or other learning materials. Nevertheless, we believe that given the participants' overwhelming agreement with the noted findings, most t-values were significant at .0001, test takers' complaints might warrant further theorizing and explanation. In the remaining of this section, we strive to put forward a consistent theoretical justification addressing why ATEL fails to fulfil its purported function efficiently. Needless to say, its purported function as a test of selection is to discriminate among test takers with regard to their literary competence. In statistical terms, items in selection tests should be designed so that test takers are placed along a normal distribution. In its current state, we infer from the findings that the test yields a positively skewed distribution given that it seems to be too difficult, as participants reportedly stated.

The main reason for the failure of ATEL might have to do with whether and how the construct to be tested is defined. In other words, it seems that in the design of ATEL, crucial stages in modern test development are skipped. Principled language test development begins with a theoretical model of the ability to be measured, followed by formulating a framework which establishes links between the purpose and context of testing on one hand and the theoretical model on the other. Scholars in the field of literary studies have also called for more systematic approaches to the development of tests of literary competence. Hanauer (1996) maintained that for tests of literary competence to be valid, they need to be "generated by a theory of literary competence, and analysis of the required task or tasks within the field of literary studies" (p. 143). Only after such grand theoretical issues are resolved, will we be in a position to start thinking about developing other components of test specification ((Davidson & Fulcher, 2007; Fulcher, 2013). This does not seem to be the case with ATEL. In fact, when it comes to assessing literary competence for high stake testing, ATEL is a measure of general knowledge about literary terms and figures not a test of literary skills. This being so, those who engage in test preparation are perplexed as to what they should study because the range of sources from which public knowledge items may be sampled is so vast that one can never exhaust the range of all possible sources. Had the test been founded on clearly defined literary abilities, the test taker would engage in learning the target skills regardless of the sources from which test items might be extracted. This echoes the concerns that Cooper (1971) had almost half a century ago; that testing literature if not founded upon a sound theory of literary competence would reduce it to testing reading comprehension and memorized bits of information about authors' biographies.

This way of test design seems to be (mis)informed by a traditional understanding of test validity which is rooted in authority not in empirical evidence. As far as the excerpts are taken from authoritative sources, tests are considered to be valid. The following quote from Fox (1938, cited in Cooper, 1971, p. 5-23) lucidly captures this mindset regarding the validity of tests of literature:

We must select our material from acknowledged classical sources of literature. It is not the psychologist's business to evaluate the worth of such literature. He must accept it until such time as it is overthrown by something which is equally widely acknowledged and has stood the test of time.

When tests are not designed according to a detailed, principled test specification, they are highly likely to suffer from construct irrelevant variance or construct underrepresentation or both.

IMPLICATIONS FOR TESTING ENGLISH LITERATURE

In its current state, the specialized literature section of ATEL is obviously not serving the function it is designed for. A number of initiatives can be taken to bolster the validity of ATEL. First, more psychometric analysis should be done on individual items on the Specialized literature component. One way forward is to use item response theory so that items whose difficulty indexes do not match test takers' ability levels be identified and deleted or improved. One reason for the current practice on the part of test takers in ignoring the more important module of ATEL is the complementary approach adopted in interpreting scores on the exam. In a complementary approach, a composite total score is estimated for the entire test so that a low score on one component can be compensated for by a high score on another. One solution is to adopt a non-complementary approach, which requires that test takers reach a threshold score on all test components (Qin, 2015). Such an approach would discourage test takers from ignoring the literature component, which happens to be the main target construct of the test.

One of the recommendations for inducing positive washback of exams is to use direct testing (Bailey, 1996) because indirect tests drive the kind of test preparation that might attenuate the validity of scores (Qin, 2011) and encourage the kind of learning that is not consistent with best practice in fostering literary competence (Hanauer, 1996). Therefore, instead of testing knowledge about literature, the literary competence of students and the associated skills must preferably be assessed. This would certainly necessitate theorizing as to the precise nature of literary competence. It would also call for more open-ended question types which would allow for deeper insights into individualized response to literature.

REFERENCES

Alderson, C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing, 13*. doi:10.1177/026553229601300304

Anagnostopoulos, D. (2003). Testing and student engagement with literature in urban classrooms: A multi-layered perspective. *Research in the Teaching of English*, 177-212.

Andrews, S. (2004). Washback and curriculum innovation *Washback in language testing* (pp. 59-72): Routledge.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*. doi:10.1177/026553229601300303

Beach, R. (2014). Assessing responses to literature *The companion to language assessment*.

Cooper, C. R. (1971). Measuring appreciation of literature: a review of attempts. *Research in the Teaching of English, 5*(1), 5-23.

Davidson, F., & Fulcher, G. (2007). Language Testing and Assessment: an advanced resource book. New York, Routledge. 148 p.

Davies, A. (1997). Introduction: the limits of ethics in language testing. *Language Testing, 14*(3), 235-241. doi:10.1177/026553229701400301

Davies, A. (2003). Three heresies of language testing research. *Language Testing, 20*(4), 355-368.

Fulcher, G. (2009). Test use and political philosophy. *Annual Review of Applied Linguistics, 29*, 3-20.

Fulcher, G. (2013). *Practical language testing*: Routledge.

Gaston, P. L. (1991). "Measuring the Marigolds": Literary Studies and the Opportunity of Outcomes Assessment. *The Journal of the Midwest Modern Language Association, 24*(2), 11-20.

Hamp- Lyons, L. (1998). Ethical test preparation practice: The case of the TOEFL. *TESOL Quarterly, 32*(2), 329-337.

Hanauer, D. (1996). ACADEMIC LlTERARY COMPETENCE TESTING. *Journal of Literary Semantics, 25*(2), 142-153.

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

Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing, 27*(2), 183-189. doi:10.1177/0265532209349468

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*. doi:10.1177/026553229601300302

Powers, D. E. (1987). Who benefits most from preparing for a "coachable" admissions test? *Journal of Educational Measurement, 24*(3), 247-262.

Qin, X. (2011). Is Test Taker Perception of Assessment Related to Construct Validity. *International Journal of Testing, 11*(4), 324-348. doi:10.1080/15305058.2011.589018

Qin, X. (2015). Do component weighting and testing method affect time management and approaches to test preparation? A study on the washback mechanism. *System, 50*, 56-68. doi:10.1016/j.system.2015.03.002

Razavipour, K., Gooniband Shoushtari, Z., & Mansoori, M. (2018). The Structural Invariance of a Model of Washback to Test Takers' Perceptions and Preparation: The Moderating Role of Institutions. *Journal of Teaching Language Skills*, -. doi:10.22099/jtls.2018.28165.2440

Razavipour, K., & Moosavinia, S. R. (2017). English and Persian Undergraduate Students' Perceptions of the Construct-(ir)Relevance of Language Proficiency in the Assessment of Literary Competence. *Iranian Journal of Applied Language Studies, 9*(2), 103-142. doi:10.22111/ijals.2017.3543

Samai, M., & Mohammadi, S. (2017). On the Development and Validation of a Scale of Test Impact on Test Takers. *International Journal of Language Testing, 7*(2), 155-177.

Shohamy. (2001). Democratic assessment as an alternative. *Language Testing*.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing, 18*(4), 373-391. doi:10.1177/026553220101800404

Shohamy, E. (2011). Assessing multilingual competencies: Adopting construct valid assessment policies. *The Modern Language Journal, 95*(3), 418-429.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: washback effect over time. *Language Testing, 13*. doi:10.1177/026553229601300305

Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10*. doi:10.1177/026553229301000103

Watanabe, Y. (2004a). Methodology in washback studies *Washback in language testing: Research contexts and methods* (pp. 19-36).

Watanabe, Y. (2004b). Teacher factors mediating washback. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Xie, Q., & Andrews, S. (2012). Do test design and uses influence test preparation? Testing a model of washback with structural equation modeling. *Language Testing, 30*. doi:10.1177/0265532212442634

Zhan, Y., & Andrews, S. (2014). Washback effects from a high-stakes examination on out-of-class English learning: insights from possible self theories. *Assessment in Education: Principles, Policy & Practice, 21*(1), 71-89. doi:10.1080/0969594X.2012.757546