English and Persian Undergraduate Students' Perceptions of the Construct-(ir)Relevance of Language Proficiency in the Assessment of Literary Competence

Kioumars Razavipour
Assistant Professor
Shahid Chamran University of Ahvaz
razavipur57@gmail.com

Sayyed Rahim Moosavinia
Associate Professor
Shahid Chamran University of Ahvaz
moosavinia@scu.ac.ir

Abstract

Of the many dilemmas facing the assessment of literary competence, one is the extent to which language should constitute part of the target construct intended to be measured. Some argue for the construct-irrelevance of language and hence recommend that it be eliminated or minimized in favor of an exclusive focus on literary competence. In practice, this does not seem to be the case, as language proficiency considerations seem to creep into assessment, clouding assessment outcomes. The current study sought to examine students' perceptions of the degree to which knowledge of language constitutes part of the construct of literary competence in two departments of English and Persian literature. To this end, a total of seventy students in two poetry courses, one in the English department and the other in the Persian, responded to a questionnaire designed to gauge their perceptions of the extent to which language competence constitutes a component of the literary competence. Data were analyzed through one sample and independent samples t-tests. It was found that language competence is somehow construct-irrelevant in testing literary competence. Interestingly, measurement-invariance was observed regarding Persian and EFL students' stance on the construct-irrelevance of language in tests of literary achievement and competence.

Keywords: Language Proficiency, Assessment, Literary Competence

Received: January 2016; Accepted: February 2017

1. Introduction

The field of literary studies is conspicuously insulated from the innovations and developments in assessment policies and methods that have transpired in neighboring fields over the last few decades (Paran, 2010). For instance, the field of language testing and assessment has by now been established as an independent field of inquiry with its own research agenda, journals and conferences. Borrowing from psychometrics, linguistics, and educational measurement, language testing has made significant contributions to the field of applied linguistics in general and to assessing verbal constructs in particular. Assessing literature, however, seems to have escaped the attention of scholars and hence it has remained insulated from the scholarship produced in language testing in particular and educational measurement in general. Strange as it may first appear, assessing literature comes the closest to assessing English for Specific Purposes (ESP).

Not unlike testing English for Specific Purposes (Douglas, 2001), where drawing a borderline between content knowledge and language knowledge is a perennial issue, it seems that part of the insulation of testing literature from mainstream language testing has to do with the nature of literature, where defining the construct of literary competence and entangling it from other neighboring constructs is no easy undertaking. In particular, given the lack of a comprehensive theory of literacy competence wherein the relationship between communicative language ability and literary competence can be clearly articulated, it is not yet known whether and the extent to which literary competence is contingent on communicative competence and if so, how it is possible to draw borders between where one ends and where the other begins. This complexity makes testing literature a choice between Scylla and Charybdis (Paran, 2010). Bachman (1990) has, nevertheless, provided some brief hints

regarding where literary competence must fall in a general theory of communicative competence. In his oft-cited model of communicative language ability, Bachman (1990) seems to suggest that understanding literature is to be subsumed under the imaginative functions which enable us:

> To create or extend our own environment for humorous or aesthetic purposes, where the value derives from the way in which the language is used. Examples are telling jokes, constructing and communicating fantasies, creating metaphors or other figurative uses of language, as well as attending plays or films and reading works such as novels, short stories or poetry for enjoyment. (Bachman, 1990, p. 94)

A similar observation has been made in the Common European Framework of Reference (CEFR), where aesthetic aspects of language are accorded value not only for their educational benefits but also for their inherent cultural values (Paran, 2010). However, neither Bachman nor CEFR guidelines elaborate on how the aesthetics of language can be subjected to the valid and reliable quantification. This essentially boils down to questions about how it is possible to delineate the imaginative functions of language from the organizational competence of the model, what weight should be given to each, and whether it is ever possible to do so. Yet, the inevitability of testing in all educational programs, including literature programs, especially in the accountability era, makes it imperative that measurement in literature be subjected to the type of systematicity that is current in educational measurement in general and language testing in particular.

This paper is a modest attempt at addressing the relationship between literary competence, a component of Bachman's imaginative functions of communicative competence, and the general organizational competence. This would hopefully contribute to building a case for the validity of tests in

literature, which are quite common but are rarely, if ever, subjected to empirical scrutiny. In so far as, tests of literature are counted as grounds for making decisions about test takers (Fulcher & Davidson, 2008; Shohamy, 2001), debates surrounding the complexity of testing aesthetics taste, should not prevent us from subjecting those tests to critical rebuttals. Another reason that adds to the significance of the issue is the washback effect (Alderson & Wall, 1993; Messick, 1996) that tests of literature exert on programs of literature education.

Before proceeding any further, we deem it necessary to emphasize that the arguments laid out in this paper are of relevance only in situations where the aim of a literature program is to foster literary competence. Therefore, cases where literature serves as content or input for teaching language are not the concern of this study. We should also pinpoint that our own experience with teaching and testing literature is with undergraduate and graduate university English literature programs. Mindful of the centrality of context in shaping pedagogical decisions (Freeman & Johnson, 1998), including testing ones, we caution that findings should not be easily extended to other programs of literary study.

2. Review of Literature

The literature on testing literary competence is rather scarce, perhaps due to the common understanding that quantification distorts beauty and taste, which are often the very essence of literature. As early as 1967, Purves observed the incompatibility of "the humanistic encounter with literature and the mechanical appraisal of education" (p. 310). In a similar vein, Gaston (1991) maintains that "Measurement, it would appear, would be unkind to beauty. Quantification and appreciation rarely coexist easily" (p. 11). The apparent

hostility noted above between measurement and aesthetics is perhaps due to the incompatibility of measurement tools with target constructs of measurements. Nearly half a century ago, Forehand (1966, cited in Cooper, 1971, p. 7) captured this incompatibility quite cogently: "What we want to measure is complex but subjective; the methods we have to work with are objective but simple. The problem, then, is to make our goals more objective and our measures more complex".

This rather scant literature is reviewed in the following order: We would first discuss, in light of the limited, existing literature, why the time has come for programs of liberal arts including literary studies to come out of their comfort zones to report their outcomes in meaningful, quantitative methods. We then review the challenges that face such programs in their efforts to measure their outcomes.

The expansion of accountability movement at all educational levels, particularly in higher education, however, has made it difficult for any educational program to survive the scrutiny of monitoring bodies, without being able to transparently document its gains for the stakeholders. Gatson lists four main reasons for this acceleration in accountability. First, unlike before, given the mushrooming of numerous public and private higher education institutions, institutions must compete to ensure adequate enrollments for their programs to stay alive. Secondly, most educational organizations rely on public funding to run their programs. As such, there is an increased pressure both from the political hierarchy and the general public to make educational institutions more accountable. In Iran, the recent pressure from the Ministry of Science, Research and Technology to tie the promotion of faculty members to evidence of measurable scholarly track records is evidence of this pressure from the hierarchy to take individual and institutions accountable for the space they

occupy in higher education institutions. Thirdly, for higher education administrators, academic accomplishments are increasingly being displaced by administrative skills. This is because universities are no longer pure academic, intellectual institutions insulated from the corporate world. They should assume responsibility for their fringing budgets. In so doing, they need administrators who are familiar with the rules of the game in the corporate market. It logically follows that such administrators would push for more outcomes assessment. Lastly, there are mounting pressures on institutions of higher education to offer academic programs that are economically sound. This would drive colleges to review their programs, identify the least efficient and vulnerable ones and replace them with programs that make more financial sense. This would in turn call for more outcomes based assessments. Language and literature programs are in a weak position in the face of the accountability tide because they are commonly unable to express in a market friendly language, language of numbers and statistics, the competencies such programs develop in their students.

The looming shadow of outcome assessment is not all evil for literature education programs, however. According to Hutchings (1990), outcome assessment "can shift attention from credits earned to competencies developed, and inspire a sense of individual accountability among students and faculty members alike" (cited in Gaston, 1991, p. 15). Currently, for students in most language and literature departments to graduate, the only requirement is to show proof of the number of credits they have passed. Departments seem to never bother to ask whether the sum of the credits earned by their students has any substantial meaning in terms of students' intellectual growth and competencies. Outcome assessment will likely put an end to this complacency.

Given the current situation that programs of arts and literature face, there are two choices before them in relation to the accountability movement: to resist or to comply (Gatson, 1991). They may choose to argue that the existing instruments of measuring literacy competence are not sophisticated enough to capture the competences that instructional programs of literature seek to develop and thus they must be exempted from outcome assessment. The other option is for the programs to "set forth more easily measurable objectives, make sure those objectives are addressed, and document their accomplishment" (p. 15). None of the above strategies is in the best of interest of programs of literature education. The former argument does not prevail and will most likely fall on deaf ears of policy makers, whose demands for accountability spares nobody. Opting for the easily measurable outcomes would lead to the deterioration of the content of such programs. In other words, the former strategy would lead to damage from outside and the latter would to damage from inside. The only viable option for those of us in the language and literature programs is to efficiently measure ourselves before they simplistically measure us (Paron, 2010; Gatson, 1991).

To start thinking what to measure in literature programs and how to measure it, we need to be clear about why to measure. Given that all testing is done for some purpose (Shohamy, 2001) and such purposes are often educational, it seems plausible to think that the purposes of assessing literature are closely tied to the purposes of literature education. According to Purves (1986, 1979), there seems to be three general aims for teaching literature.

1. Transfer of knowledge within literary/cultural texts of a group
2. Training qualified readers and critics of such texts
3. Promotion of personal empowerment by literary texts through the other two aims

According to Purves, the functions literature tests are to serve are contingent upon which of the above-mentioned purposes literature is taught. Borrowing an analogy from Universal Grammar, which draws a distinction between deep and surface structure of linguistic utterances, Purves (1979) classifies literary curricula in three structures of imitative, analytic, and generative. In the imitative structure, Purves argues, literature is taught for social cohesion; it serves a social and political purpose. It is used as a means for transferring ideal cultural heritage so that citizens grow up feeling proud of their national affiliation. In such a structure, where transfer of knowledge is intended, testing literature is expected to focus on measuring the recall of literary information. The imitative curriculum, poses the least number of challenges for assessment, for question of validity, defined as alignment between curriculum content and tests, is easier to address. When literature, like history, is taught to engineer national cohesion, assessment of literary competence is in fact assessing a bank of literary information that resides in the text and can be tested via multiple choice items. In the analytic structure, on the other hand, literature teaching seeks to develop a set of skills in learners to be critics of literary texts. Training critics in teaching literature raises the question of testing skills; and finally, in the generative structure, which seeks to promote personal empowerment, testing literature would be about testing attitudes and personal response to texts.

Beach (2014) traces three developments in the assessment of literary competence. In the first stage, during the heyday of Formalism/New Criticism approaches, the literary text was seen as an autonomous unit of meaning, which allowed for universally uniform understanding and interpretation. This is close to what Purves terms imitative structure in literature curriculum. Advances in cognitive theory and cognitive processes of reading, according to Beach,

stimulated interest in how readers interact with texts, including literary texts. Hence, readers' schemata and world knowledge were entered into the equation. It was no longer the literary text per se that mattered, but also the readers' cognitive attributes and processes. The third development in literature assessment has to do with embracing insights in the socio-cultural turn in social science and humanities. In the socio-cultural school, learning is a collective phenomenon taking place in a social milieu, not a product of individual cognitive processing in isolation. Therefore, response to literature is to be viewed as a collaborative endeavor that should be accomplished in groups. This view would demand for an assessment for learning approach to assessing literature. In particular, alternative assessments (Brown & Hudson, 1998) such as portfolio assessment, dynamic assessment (Poehner & Lantolf, 2013), as well as peer and self-assessment (Douglas, 2011) would be consistent with a socio-cultural view of response to literature. This evolvement from literary text as an autonomous object of learning to the role of reader's cognition in responding to literature, to the socio-cultural approach has increasingly made the assessment of literary competence more challenging because the construct has expanded to include not only text features but also test takers' characteristics and their social environment. To Beach (2014), assessing response to literature via objective, multiple choice items is a thing of the past.

What complicates matters further is that goals set for teaching literature are not always straightforward to tell. Indeed, very often various goals coexist during a literature education program or even in the span of a single course in literature. It should also be borne in mind that the goals teachers set themselves in teaching are not necessarily the same as the goals policy makers have in mind in designing curricula (Wall, 1996). The often implicit nature of

goals or their mixing together add to the complexity of making decisions concerning the function testing literature is to fulfill.

Lastly, it is high time for teachers and faculty members to get engaged in literary courses and should fill the vacuum projected by the nature of literary studies through "some sort of questionnaire or informal interview" (Purves, p. 323). As Gatson (1991) maintains, to practically save the value of literary studies in a measuring and measured world, it is time to carry out this important task. However difficult it is for professionals in the liberal arts and especially literary studies to perform such a task and resolve the dilemma, determination is always looming on the horizon to settle the problem of measuring outcomes in literary studies programs.

The above account was almost all about measuring literature in the mother tongue. In testing literature in a second or foreign language, all the issues involved in testing literature in the native language linger, plus a set of additional specific problems. Paran (2010, p. 153) has identified six dilemmas for testing literature in EFL teaching: whether testing is an external activity with a set of gate-keeping goals or an internal activity with a cluster of internal goals, such as individual growth and character development; whether to test language or test literature; to test literary knowledge or literary competence skills; to test public literature knowledge (efferent reading) or personal appreciation of literature (aesthetic reading); to introduce genuine everyday oral tasks or formal non-specialist pedagogic tasks; teaching skills and whether to test metalanguage or not. Our concern in this research is the second in Paran's list, that is whether to teach, and by way of modification test language or literature.

Thus Purves (1986) has clarified this question:

> Research indicates that the ratings of various aspects of performance are related to each other, but that raters aware of the relationships can make distinctions between the content and the form of a written or dramatic performance. For an overall grade in language arts, of course, teachers might want to combine the two, but for the literature aspect of the grade, the content is important. (P. 323)

To build a typology of test takers' responses to literary criticism, Purves (1967) turned to literary theories to no avail. In lieu of a grand theory, he adopted a bottom-up approach to group whatever test takers write about a work of literature. Analyzing essays in literary criticism written by nearly 500 students in a handful of languages, Purves concluded that whatever students write boils down to the four categories of engagement, perception, interpretation, and evaluation. Of the four categories, only perception, according to Purves, lends itself to objective testing, with the other three too broad constructs to be captured reliably through objective tests.

There are two more additional challenges to assessing response to literature in an L2. These challenges have been mainly researched for English Language Learners (ELLs) in the United States. In the first place, there is substantial evidence that "students' response processes or strategies do not readily transfer from L1 to L2 reading in that students need more than simply L2 linguistic ability to interpret L2 literary texts" (Bernhardt, 2005, cited in Beach, 2014, p. 90). Further, when response to literature is tested through open-ended written exams, which are superior to discrete point test on validity grounds as alluded to earlier, an enormous challenge for students is to state their interpretations in writing due to their limited language proficiency. As such, the meaning of scores assigned to students in such exams is confounded.

If scores vary by virtue of language proficiency rather than literary competence, it would be an obvious case of construct-irrelevant variance.

Studies focusing exclusively on how literary competence in English is assessed at tertiary education levels in EFL contexts are rather scant. Kadhim (2015) surveyed a few colleges of Arts and those of Education across Iraqi universities to see how testing literature varies across colleges as well as across courses such as poetry, novel, and drama. She also investigated whether language accuracy is a criterion in testing literature. It was found that the two types of colleges differed in their aims of teaching literature and the different aims in turn affected test types, formats, and tasks. She also found that assessment formats and types varied across courses. Finally, it was found that teachers of English literature in both college types confounded language accuracy with literacy competence.

Kadim's study was on how tests vary as a function of college and course type. Yet, tests are not always pliable means in teachers' repertoire of educational assets. In many situations tests affect decisions literature teachers make concerning both course content and instructional methods. Zancanella (1992) conducted a case study to see how state-mandated tests influence teachers of literature. Two teacher characteristics proved to be moderating the tests' influences. One was the convergence between teachers' favorite approaches to teaching literature and the approach reflected in the test content. The other factor had to do with teachers' power within the curriculum. Teachers in lower status were more inclined to teach to the test.

This brief review reveals that there is no escaping from assessing response to literature given the accountability movement across the globe. Now that measuring response to literature is inevitable, identification and purification of the construct of literary competence should top our agenda because unless we

know what we are to measure, questions of how to measure it would not get off the ground. The present study is a modest attempt in that direction: demarcating the boundaries between language ability and literary competence. Thus, the question we seek to answer is whether language proficiency should feature in assessing literary competence. This is a worthwhile question for two reasons. In the first place, it is in keeping with the democratic assessment paradigm in language testing, championed by Shohamy (2001, 2014), stakeholders' ideas and perceptions do matter in the act of assessment. Moreover, there is evidence that test takers' perceptions bear on the validity of test scores (Xie, 2011; Xie & Andrews, 2013). Our second objective is to pin down the possible differences of students' perceptions that might exist between assessing literature in the L1 Persian and L2 English at tertiary levels.

3. Methods

A total of 74 undergraduate junior and senior students comprised the participants of this study. They were students in two poetry courses of English and Persian. Twenty nine were studying English literature and 45 were undergraduate Persian literature students. One third of the participants were male and the rest were female and they aged between 21 and 27. They were chosen based on a purposive sampling procedure. As we were seeking two comparable classes in terms of the requirement of literary assessment, we ended up with two courses in poetry, one in English and the other in Persian. In doing so, we had two criteria in mind. First, the two courses had to be identical across the two departments, that is, we were looking for participants in classes that were similar in everything but the language (Persian or English). We were also interested in courses whose outcomes were to be assessed via essay type

questions rather than discrete point tests, because in the latter type tests productive language skills are not implicated.

A questionnaire and semi-structured interviews were the major data collection procedures of the study. As the results from the latter are already published (the reference is withheld to maintain the anonymity of peer review), in this paper we are going to limit ourselves only to the quantitative aspect of the study. The final version of the questionnaire consisted of 20 Likert-scale items written in Persian. Since Likert (1932) invented the Likert type scale, there has been controversies surrounding the proper way of describing and analyzing data from such scales (Boone & Boone, 2012). In particular, whether data from Likert item types should be considered ordinal or interval data or whether the resultant data from such scales are to be treated with parametric or nonparametric statistics has been at times subject of some debate. Through making a distinction between Likert type and Likert scale data, Boone and Boone (2012) maintain that the way out of this dilemma is to be clear about the nature of the variable of interest that is to be measured with the Likert instrument. If it is not conceptually plausible to add up items to compute composite scores, that is, items do not contribute variance to a single latent variable like attitudes, beliefs, or perceptions, then we would have a case of Likert type data and they should be analyzed using non-parametric statistics (each item should be analyzed separately). If, however, it conceptually makes sense to compute an aggregate or composite score for the items, the data is Likert scale and they should be considered interval data and should be described and analyzed using parametric statistics. Since all items in the current instrument were written in the pursuit of measuring participants' perception of the construct relevance of language proficiency, the data for the present study were of the latter type, Likert scale, and hence amenable to parametric

statistics. In other words, the items were not each tapping a different trait, as is the case with Likert type items used in research on, say, language learning strategies (Tseng, Dörnyei, & Schmitt, 2006).

To guard against any differential language proficiency that might play into the process of responding. The items were pooled drawing on our review of the pertinent literature, informal interviews with students, as well as the researchers' experience of testing and teaching English language and literature. The items were all designed to tap on participants' opinions regarding the extent to which they believed that the construct of language proficiency is part of the construct of literary competence or whether the two are distinct abilities that should be kept apart in testing students' literary competence. For instance, participants were asked to indicate their (dis)agreement with the inclusion of grammar, diction and writing issues in assessing their responses to open-ended, written poetry and prose examinations. Best practice in questionnaire design has it that in writing items there has to be a combination of positive and negatively worded items to counter bias. In keeping with this, we included a handful of such items and they were then reverse coded. To see if the data lend themselves to parametric statistics, statistical assumptions of skewness and kurtosis were examined and all items appeared to be within the acceptable range of minus and plus two (Bachman, 2004).

To enhance its validity, the questionnaire was subjected to multiple expert reviews, leading to the changes in wording of some items and the removal of some others. To our dismay, initial reliability analysis pointed to an index below the acceptable threshold. "Corrected item-total" correlation and "alpha-if-item-deleted" indexes (Hatch & Lazartan, 1991) were checked but none was found promising to increase its reliability in significant ways. We finally speculated that it might have to do with comparable validity of the

questionnaire for the two groups of participants, namely Persian and English literature students. We found that the questionnaire was more reliable for English literature students (alpha=.68) than for Persian literature students (alpha=.5). Despite its popularity, however, Conbach's alpha should not be seen as a versatile, infallible index of reliability because often its underlying assumptions cannot be met (Hair, Hult, Ringer, & Sarstedt, 2014; Brown, 2014). Equidistance and equal difficulty level of items are two such assumptions that are often not satisfied in Likert scale questionnaires (ibid). Another possibility is that students in TEFL departments are more Likert scale-savvy merely because they are more frequently exposed to such scales, thanks to the nature of research in English departments.

To collect the data, with prior arrangements with instructors, one of the researchers attended the classes and administered the questionnaires in person. In each class before handing in the questionnaires, he briefed the participants about the purpose of the study and they were assured of the anonymity of their responses. Participation was voluntary and participants were invited to ask questions should they need any clarification on any item. It took them about 20 minutes to complete the questionnaires.

To analyze the data, all questionnaire data were entered into SPSS, version 16. They were then analyzed using both descriptive and inferential statistics. In addition to examining the reliability of the questionnaire, one sample t-test, and independent samples t-test were run to answer the research questions.

4. Results

In this section, after checking the normality assumption of the data, we first present the descriptive statistics regarding participants' perceptions of the

construct-relevance of language proficiency in the assessment of their achievement of literary competence. Our first research question was the extent to which students of literature believe that language proficiency considerations should constitute criteria in achievement tests of literature. In Table 1, descriptive statistics of participants, English and Persian students combined, are given. The mean is 2.95 and the standard deviation of scores is .3. To interpret this mean, we must remember that given our five-point Likert scale questionnaire, the maximum score was five and the minimum was one. As Table 1 illustrates, the mean of 2.95 stands somehow between the two ends of the scoring scale, indicating that participants seem to have adopted a middle stance regarding the question of language issues in the assessment of literary competence.

Table 1 gives the kurtosis and skeness values for individual items on the questionnaire as well as those of the total scores.

Table 1. *Skewness And Kurtosis Values For Items and Sum Scores*

	N	*Mean*	*Std. Deviation*	*Skewness*		*Kurtosis*	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
item1	74	3.7973	.95055	-.562	.279	-.052	.552
item2	74	4.2432	.73672	-.633	.279	-.140	.552
item3	74	3.8784	.90588	-.322	.279	-.756	.552
item4	74	2.7027	.96130	.446	.279	-.218	.552
item5	74	1.9054	.68584	.123	.279	-.828	.552
item6	74	3.1622	1.03404	-.335	.279	-.465	.552
item7	74	2.8649	.86522	.399	.279	-.582	.552
item8	74	3.0000	.84400	-.141	.279	-.548	.552
item9	74	2.4324	.89260	.447	.279	-.008	.552
item10	74	3.0946	.99545	.063	.279	-.470	.552
item11	74	3.3649	.91523	-.028	.279	-.858	.552
item12	74	3.3649	.90014	.013	.279	-.240	.552
item13	74	2.1216	.96447	.504	.279	-.251	.552
item14	74	2.3784	.82267	.857	.279	.792	.552
item15	74	2.9865	1.09160	.287	.279	-.442	.552
item16	74	2.9054	1.03592	.118	.279	-.721	.552
item17	74	2.6622	1.11376	-.024	.279	-.954	.552
item18	74	3.0811	1.01707	-.246	.279	-.528	.552
item19	74	3.1892	1.01598	-.312	.279	-.177	.552
item20	74	1.9730	.92118	.595	.279	-.539	.552
total	74	2.9554	.30223	.096	.279	.494	.552
Valid N (listwise)	74						

As can be seen in Table 1, all the kurtosis and skewness value are within the required. According to (Bachman, 2004), kurtosis and skewness values within a -2 and +2 range are indicative of the normality of the data. That said, since all

the analyses in the current study are based on composite scores, it is the skewness and kurtosis values of the sum scores that are of special interest, which are again .09 and .49, respectively.

Table 2. D*escriptive Statistics of all Participants*

One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
total	74	2.95	.30	.035

To see whether the observed mean significantly deviates from the neutral value of 3 (the average value for a five-point Likert scale), a one sample t-test was conducted, the outcome of which is given in Table 2 (t=1.26, p=.2, df=73).

Table 3. *Results from One-sample T-test*

One-Sample Test						
	Test Value = 3					
	T	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
total	-1.269	73	.20	-.044	-.114	.025

The large p value indicates that participants' responses do not significantly deviate from the neutral value. To relate this statistic back to the research question, it is evident that participants' agreement to the construct irrelevance of language proficiency to literary competence is not strong enough to reach statistical significance. In simple terms, to participants, language remains somehow construct relevant in the assessment of literary competence.

Table 4 illustrates the means and standard deviations of participants' total scores on the questionnaire.

Table 4. *Group Descriptive Statistics of Sum Scores on the Questionnaire*

	Language	N	Mean	Std. Deviation	Std. Error Mean
dimension1	English	29	2.90	.32	.06103
	Persian	45	2.98	.28	.04213

The mean score for the English literature participants is 2.9 and that for the Persian students is 2.98. The standard deviation is .32 and .28 for English and Persian students respectively, indicating more diversity among the former group. Eyeballing the two means we discern some apparent difference; yet, this has to be statistically substantiated. To do so, an independent samples t-test was conducted.

Table 5. *T-test Results Comparing English and Persian Literature Students' Perceptions*

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	T	df	Sig.	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
total	Equal variances assumed	.879	.35	1.19	72	.23	-.08	.071	-.228	.057
	Equal variances not assumed			1.15	53.33	.25	-.08	.074	-.234	.063

T-test results are given in Table 5. It clearly demonstrates that the observed difference in table 1 is not of statistical significance (t=1.19, p=.23, df=72), indicating similarity between Persian and English literature students in regard of their stances concerning the construct-(ir)relevance of language proficiency in assessing literary achievement.

5. Discussion and Conclusions

Whereas in the past tests were the considered the province of experts with test takers being only the objects of measurement, recent paradigms in critical language testing (Shohamy, 2001) and alternative assessments (Brown & Hudson, 1998) call for stakeholders' involvement in the act of assessment. Likewise, the sociocultural views of response to literature maintain that test takers' background and perceptions are undeniably important variables in response to literature (Beach, 2014). The current study set out to put to empirical scrutiny the extent to which in achievement tests of literature, language competence constitutes part of the target construct. The short answer to the question was that test takers' somehow endorsed the construct-irrelevance of language competence in assessing literary competence. From a theoretical standpoint, test takers' perceptions were found to be inconsistent with Bachman's (1990) model of communicative language ability, wherein the organizational competence and the ability to perform imaginative functions with language both constitute aspects of the general communicative competence in a language.

The results were also out of sync with the current practice in assessing literary competence. "In many contexts we end up employing at least two criteria when we are grading or evaluating any sample that we have as the result of our literature test, a literary criterion and a language one." (Paran, 2010, p. 147). The current practice seems to echo the inseparability dilemma known in testing ESP (Douglas, 2001, 2011; Brunfaut, 2014). Yet, the issue of inseparability of content and language in assessing literary competence is even more problematic in that more often than not the point of a piece of literacy work lies in the very language used in it. As such, it is even more difficult than in ESP testing to draw a solid, sharp line between content and language in

assessing literature. However, the primary focus of study programs must inform our scoring decisions to strike a healthy balance of language competence and literary competence in assessments. Paran maintains that "It is therefore important to be very clear about which competence we are tapping, and which aspect of performance in the test we are going to mark." (p. 148). In EFL and ESL programs whose primary goal is to teach language through literature, then the primary construct for measurement is language, with literary competence being involved only if the assessment is of a task-based nature.

Besides curricular orientations, individual differences and preferences are also crucial in deciding on the construct-relevance of language in tests of literature. Using a confirmatory factor analysis, Miall and Kuiken (1995) identified seven components for the construct of response to literature. Some of these components were more text-based than others. Obviously, depending on what aspect of literary competence is the object of measurement has a bearing on the degree to which language competence constitutes an aspect of literary competence.

Our research question addressed the possible difference between assessing literature in the mother tongue and that in EFL in relation to the role of language proficiency. Interestingly, Persian literature students and students of English literature held similar views concerning the construct relevance of language to testing literary competence (see Table 4). This finding runs counter to some previous findings which maintain that assessing literature in the L1 is essentially different from that in the L2 (Beach, 2014; Purves, 2010). Yet, those studies were based on expert views, not on those of the test takers. In the context of this study, this finding is plausible considering the fact that both groups of participants were studying literature as their field of study not as a means of improving their language competence. Clearly, more studies need to

be conducted to arrive at substantive conclusions regarding the weight that must be given to language proficiency in assessing literature.

Based on students' views, the relevance of language proficiency to literary competence is stable across the two languages of English and Persian. In other words, independent samples t-test results indicated that English and Persian students held similar views concerning the construct (ir)relevance of language proficiency to literary competence.

Participants' perceptions regarding the relevance of language proficiency to the assessment of literature seem to indicate that they are somehow of the opinion that in situations where a test of literary competence involves some written or spoken production, the quality of their language should not confound their literary competence. Adhering to such a separation of the two competencies in assessments of literature would promote positive washback (Messcik, 1996) too as test preparation would be directed toward achieving the tested construct: literary competence.

In conclusion, we barely managed to scratch the surface of a complicated, multi-faceted issue. Numerous tests of literature of various types for various functions including achievement, selection and credentialing have been and will continue to be given across institutions and countries. In many of these instances of testing, the future of numerous students is at stake. Thus, we cannot simply leave it to impressionistic evaluations subject to individual assessors' idiosyncratic preferences. Now that there is no escape from testing literature (Paron, 2010; Gatson, 1991), it has to be tested in keeping with the best knowledge and practice in language testing and assessment. And in doing so, one of the first questions that must be addressed is the weight that must be given to language in assessing literary competence.

6. Limitations and Future Research

As no research is final or perfect, a number of shortcomings in here in the present study. In the first place, we hypothesize that the outcomes of this study might have partly to do with the participants who perhaps counted themselves beneficiaries of the survey results. In other words, a halo effect might have crept into the collected data because the study was conducted by internal researchers, who were in a position to alter students' test scores. Future research by external researchers or on students who have already graduated is likely to rectify this pitfall.

Another possibility is that the questionnaire we designed might have clouded the outcomes. Despite our efforts at gathering expert views and examining internal consistency, we still surmise that a better case could have been made for the validity of the instrument we used. Particularly, the questionnaire was less internally consistent with Persian literature students, which we might ascribe to the fact that students in Persian departments are less Likert-scale-savvy than those in the TEFL departments. Secondly, it may have to do with the authors' background; both of us come from an English background. More importantly, our questionnaire was not founded upon a substantive theory or model of literary competence, which to our knowledge does not exist. Building data collection instruments based on a comprehensive model of literary competence with clear specifications is in and of itself a worthwhile future inquiry. Needless to say, such instruments would add credibility to future research findings.

Systematic research into the assessment of literary competence of the kind common in language testing is scant in testing literary competence. Future research should delve more deeply into the components of literary competence. In particular, studies with a cognitive bent using verbal protocol analysis hold

the promise to further our understanding of what constitutes literary competence. Moreover, as literature stands at the intersection of language and arts, the literature in quality assessment in other fields of art can be of help in systematizing the assessment of literary competence.

References

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

Bachman, L. F. (2004). *Statistical analysis for language assessment*. Oxford: Oxford University Press.

Beach, R. (2014). Assessing responses to literature. In A. J. Kunnan (Ed). *The Companion to Language Assessment* (pp. 85-101). John Wiley & Sons, Inc. DOI: 10.1002/9781118411360.wbcla017

Boone, H. N., & Boone, D. A. (2012). Analyzing likert data. *Journal of Extension*, *50*(2), 1-5.

Brunfaut, T. (2014). Language for specific purposes: Current and future issues. *Language Assessment Quarterly*, *11*(2), 216-225.

Brown, J. D. (2014). Classical theory reliability. In A. Kunnan (Ed). *The companion to language testing* (pp. 1149-1166). New York: Routledge.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, *32*(4), 653-675.

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

Douglas, D. (2010). *Understanding language testing*. London: Hodder Education.

Douglas, D. (2001). Three problems in testing language for specific purposes: Authenticity, specificity and inseparability. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, . . . & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 45–52). Cambridge, England: Cambridge University Press.

Freeman, D., & Johnson, K. E. (1998). Reconceptualizing the knowledge-base of language teacher education. *TESOL Quarterly*, *32*(3), 397-417.

Fulcher, G., & Davidson, F. (2008). *Language testing and assessment: An advanced resource book*. New York, NY: 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.

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2017). *A primer on partial least squares structural equation modeling (PLS-SEM)*. 2nd Edition. Thousand Oaks: Sage.

Hatch, E. M., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York, NY: Newbury House Publishers.

Kadhim, H. M. (2015). *Testing English literature at the university level in Iraq*. Unpublished M.A thesis. University of Thi-Qar.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140), 1–55.

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

Miall, S. D., &Kuiken, D. (1995). Aspects of response to literature: A new questionnaire. *Research in the Teaching of English, 29*(1), 37-58.

Paran, A. (2010). Between Scylla and Charybdis: The dilemmas of testing language and literature. In A. Paran, & L. Sercu (Eds). *Testing the Untestable in Language Education* (pp.143-165). Bristol: Multilingual Matters.

Poehner, M. E., & Lantolf, J. P. (2004). Dynamic assessment in language classroom. *Language Teaching Research*, *9*(3), 233-265.

Purves, C. A. (1986). ERIC/CRC report: Testing in literature. *Language Arts*, *63*(3), 320-323.

Purves, C. A. (1979). Evaluation of learning in literature. *Evaluation in Education, 3*(2), 93-172.

Purves, A. (1967). Literary criticism, testing, and the English teacher. *College English*, *28*(4), 310-313.

Shohamy, E. (2014). *The power of tests: A critical perspective on the uses of language tests*. Routledge: New York.

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

Tseng, W.-T., Dörnyei, Z., & Schmitt, N. (2006). A new approach to assessing strategic learning: The case of self-regulation in vocabulary acquisition. *Applied Linguistics*, *27*(1), 78-102.

Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing*, *13*(3), 334-354.

Xie, Q. (2011). Is test takers' perception of assessment demand related to construct validity? *International Journal of Testing, 11*(4), 324-348.

Xie, Q. & Andrews, S. (2013). Do test design and uses influence test preparation? Testing a model of washback with structural equation modeling. *Language Testing*, *30*(1), 49-70.