Assessing Assessment Literacy: Insights From a High-Stakes Test

Kioumars Razavipour

Shahid Chamran University of Ahvaz, razavipur57@gmail.com

Received: 01/11/2013 *Accepted: 04/01/2014*

Abstract

This study constitutes an attempt to see what Language assessment literacy (LAL) is for three groups of stakeholders, namely LAL test developers, LAL instructors, and LAL test-takers**.** The perceptions of the former group were derived from the content analysis of the latest version of the LAL test, and those of the latter 2 groups were assessed through a survey designed by the researcher. Participants were 54 M.A. TEFL students sampled conveniently. Descriptive statistical analysis of the data revealed that for test designers LAL is mainly a matter of knowledge and theory with little importance accorded to skills and even less so to principles. For instructors and test-takers, LAL was perceived to be mainly a matter of skills. Moreover, test-takers perceived of LAL as the most challenging module of the test because of its dealing with statistics, its theoretical nature, and test-takers' lack of hands-on experience with language tests.

Keywords**:** Language Assessment Literacy; Language Testing; High-Stakes Tests

1. Introduction

Assessment literacy is claimed to be the key to effective teaching (Popham, 2006, 2009). By the same token, LAL is pivotal in how effective language teaching can be implemented and evaluated, especially so given the new perspectives on testing such as assessment for learning (Stobart, 2008), assessment as learning (James, 2008), dynamic assessment (Lantolf & Poehner, 2008). As the concept is rather new in the field, a consensual definition is yet to emerge as to what precisely LAL is and how it is to be defined for groups with different needs and levels of involvement in language assessment practices. Therefore, the LAL required of a professional testing scholar is different from the LAL expected of classroom teachers; the demands are much higher for the former. Inbar-Lourie (2013a, 2008) identifies two dimensions to the LAL construct: generic and specific. The former dimension entails the knowledge base in assessment which is common to other areas of educational testing, but the specific dimension constitutes the body of assessment knowledge that is required for implementation and use in language assessment. To date, because of the many possible complexities and dimensions that are involved in the construct of LAL, almost no test has been developed for its measurement, especially for high-stakes purposes. The LAL component in the university entrance examination which is used to screen candidates to M.A. TEFL programs is probably one of the few LAL tests that operationalizes the construct of LAL.

To the best of our knowledge, to date no study of LAL has been conducted based on an actual test purported to measure the knowledge base of candidates in language assessment. Moreover, another gap that is apparent in the literature is that no evidence exists as to the nature of the washback of a test of this type. Hoping that the gap has been rightly noticed, this paper is based on a preliminary study on the language testing module of the annual test that is administered to M.A. TEFL candidates nationwide annually. More specifically, this study examines what LAL is for LAL test designers, what it is for LAL course instructors, and what constitutes LAL for test-takers. It is worthy to note that given the existing literature on test washback, how LAL is defined for test designers is highly likely to mediate what it is for the other two groups of stakeholders: LAL course instructors, and test-takers.

2. Literature Review

The field of language testing has evolved in tandem with advances in language teaching, though at some stages in its evolution with a slower pace. Looking at language testing in terms of its three historical stages, as described by Weir (2005), we find the stages more or less parallel to those of language teaching. Over the last several decades, language testing scholars have attempted to keep the field vibrant and dynamic by investigating it through the lenses of neighboring and parent disciplines like psychometrics, general education, and applied linguistics (McNamara & Roever, 2006). In so doing, language testing researchers have turned a blind eye to the dissemination of the findings of their studies to those expected to put them into actual practice. Efforts to make the body of knowledge produced in the field accessible to the wider audience have been scant, small, and limited (see Inbar-Lourie, 2008). Only recently has the community realized that pushing the boundaries of the field ahead without corresponding rises in its stakeholders' knowledge base in language testing renders the findings and innovations of limited practical use. The last decade, however, has seen an increasing awareness of the importance of enhancing the LAL of various groups who are to varying degrees involved in evaluating, through measurement, the language behavior of test-takers.

As the whole business of language assessment comes to developing valid measures of the constructs for the intended uses, one might wonder if assessment literacy has anything to do with test validation. Following Messick (1996), O'laughlin (2013) captures the role of LAL in test validity by maintaining that "the validity of a test hinges critically on the interpretation of test scores and the uses to which they are directed" (p.365). Looking at validity as interpretation of test scores, it is inferred that validity of tests depends on who is going to make the interpretations and draw inferences and who decides to what uses test scores be directed. Thus, the legitimacy and appropriateness of the interpretations becomes a matter of how qualified the interpreter is. The more assessment literate the one who makes the interpretations the more likely they are to be valid. It follows that a language test designed by an institution or an individual with poor LAL is of little chance of surviving the scrutiny of validity criteria.

Taking on board the crucial role of LAL for effective teaching, Fulcher (2012) maintains that lack of LAL disempowers teachers and makes them vulnerable to and defenseless against the mandated policies.

> Teachers often seem unable to affect the policy, change the intended effect, or resist external imposition when they are aggressive. This may in part be because of a lack of conceptual assessment tools to evaluate and construct counter arguments, or the practical skills to investigate tests, test use, and deal with intended changes. (p.14)

One of the earliest studies on the LAL for teachers was carried out by Brindley (2001). To provide language teachers with the essentials of LAL, Bridnley suggested a program in a modular fashion that was composed of two core units and two additional units. In the first core unit, the "social context of assessment"(p.129), teachers develop an understanding of questions of accountability, values, ethics, and policies in assessment. The second core unit addresses the fundamental question of what it means and takes to know to use a language. Language proficiency models and issues of how to make judgments about human performance in communication are addressed in this module. Reliability and validity of language tests should also be covered in this core unit. The two additional units are to train teachers in language test construction and test evaluation. Discussions of classical item theory and item response theories are to be accommodated in such a unit. The final additional unit, according to Brindley, should be allocated to classroom assessment and criterion-referenced testing of achievement, progress, and diagnosis.

Focusing on epistemological orientations, Inbar-Lourie (2008) calls for fundamental shifts in understanding the philosophy, purposes, and procedures of language assessment. From this perspective, acquiring the capacity to carry out proper assessment procedures and policies in education is not a matter of learning a set of know-how. Rather, teachers and practitioners are urged to abandon the positivistic epistemology that epitomizes the testing culture in favor of an interpretive epistemology that seeks to establish an assessment culture, in which reality is socially coconstructed (p.387). Accordingly, to enhance the LAL of teachers, the way forward, is to provide them with opportunities to reflect on their own tacit assumptions so that they can gradually come to see learning not as a matter of individual cognition but a reality that is coconstructed by teachers and learners. Inbar-Lourie (2008) maintains that "onducting assessment within this constructivist contextually-situated framework necessitates internalization of assumptions and beliefs about assessment as a social practice and a social product" (p.387). The extent to which such an expectation is realistic or feasible remains to be investigated; however, given the resources available to classroom teachers it seems that the bar has been set too high.

With regard to what constitutes LAL, several scholars have attempted to capture the essentials of the construct. Davies (2008) reviewed the language testing course books written during the last four decades. He discerned two major trends in the language testing textbooks written during the time. One was the increasing professionalism, which is while a good cause for celebration, may insulate the field "from other potentially rewarding disciplines" (p. 327).The other trend discerned was the expanded view of the skills needed by the profession members. Unlike the earlier textbooks that were mainly concerned with teaching the skills of language testing, later textbooks gave more space to teaching the knowledge of both language and measurement. The third and the last to appear in textbooks, according to Davies (2008), is a concern for principles of language assessment, which includes issues of fairness, ethics, and social consequences of language assessment. The concern with principles has been taken up by other scholars. Shohamy (2001) pioneered the critical language testing field by proposing democratic assessment, which considers the current principles and practices current in the field as ideologically loaded. McNamara and Roever (2006) also lay emphasis on the central role of principles in training for language assessment: “In terms of academic training, we stress the importance of a well-rounded training for language testers that goes beyond applied psychometrics ... a training that includes a critical view of testing and social consequences” (p. 255).

As to the constituents of LAL, Inbar-Lourie (2008, 2013a) maintains that it all boils down to understanding the why, the what, and the how of language assessment (see Inbar-Lourie, 2008, pp. 390-394). The body of assessment knowledge that is recommended by Inbar-Lourie includes, but not limited to, what Brindley (2001) saw fit a decade ago. Inbar (2008) argues that for efficient integration of assessment and teaching, teachers learn to be well versed in new insights and findings in language teaching and learning in the what module. In particular, she suggests that teachers be familiar with issues of teaching English as an international language and integrated content and language programs.

Like other areas of education, where there is always a gap between what policy makers announce and what practitioners implement in their daily teaching practices (see Alderson, 2009), the LAL is perceived differently by various groups depending on their education background. Heejeong (2013) conducted an online survey to examine the effects of educational background on the content and structure of language testing courses. He found that opinions were divided as to what constitutes LAL among instructors who majored in the area of language testing and those whose background was in other language-related majors. The former were more likely to focus on the theories and the technical dimensions whereas the latter opted for classroom assessment, alternative assessment, and test accommodation. He also emphasized that the latter group of language testing instructors "are less confident in teaching technical assessment skills compared" (p. 345) to the former. A similar conclusion was arrived at in Malone (2013) study.

Taylor (2013) gives a couple of guidelines for effective dissemination of language assessment knowledge base. First, she discourages top-down approaches to building syllabuses of language assessment for various stakeholders, given that different groups need different levels of LAL. She maintains that:

> A profile for classroom language teachers, however, may end up focusing strongly on the practical know-how needed for creating tests but a much higher focus on measurement theory or ethical principles; the latter may need to be touched upon only briefly at a surface level. (p. 409)

Malone (2013) also underscores the need to be mindful of the needs of the end users of language assessment. "In promoting assessment literacy, material developers must be mindful that what interests language testing experts may not be of interest to language instructors" (p. 343).

The second guideline in Taylor (2013) relates to the ways ideas and concepts of language testing are packaged for communication. She maintains that the way we elucidate and talk about issues in language assessment should be as accessible as possible to lay audiences. To this end, terminology and phraseology should be kept to a minimum. Taylor believes that compared with 50 years ago, our discourse has become more inaccessible than otherwise. This challenge of simplifying the technical lingo of language assessment was also observed by Malone (2013) in developing an online tutorial for foreign language teachers. "The project revealed that practical issues involved in transforming complicated language testing issues into understandable, jargon-free language to communicate a clear message to the target audience" (p. 330).

Pill and Harding (2013) suggest that the dichotomized way of conceptualizing LAL is less helpful than one that takes it as a continuum. Drawing on literacy development in other areas such as math, Pill and Harding propose a five-level scale of LAL: illiteracy, nominal literacy, functional literacy, procedural and conceptual literacy, and multidimensional literacy. In their study, they found that stakeholders, parliament debate participants, were either illiterate or had minimal LAL. Working with members of Australian Parliament, Phil and Harding found that that parliament members, as policy makers who pass laws for the language requirements of prospective immigrants, were unaware of the existence of a field of expertise called language testing. Thus, they call on the language testing community to go more visible and become more policy literate so that their opinions are sought when needed. They conclude that whereas practitioners are in need of more dozes of LAL, the LAL experts need heavy dozes of policy literacy to make the field known to outsiders. Pill and Harding make the important point that sometimes addressing nominal literacy is more of a challenge than illiteracy "because it requires the dismantling of a misconception before an accurate conception can be achieved" (p.398). This central role of tacit, unconscious assumptions in literacy has been recognized by scholars in other areas of literacy as well. In discussing physical literacy, Whitehead (2010) maintains that:

> It is interesting to note that the overall conscious awareness which human beings experience has been described as 'the tip of iceberg' in the working of consciousness. It is proposed that the subconscious, working below our awareness, manages the majority of our everyday functioning. (p. 19)

Razavipur, Riazi, and Rashidi (2011) studied the mechanism through which the washback of external standardized tests is mediated by teachers' competence in language testing. They administered a test of assessment literacy along with a scale of test washback. It was found that teachers with higher scores on the assessment literacy measure were more likely to find a healthy balance between meeting the demands of powerful, external tests and those of classroom assessment. Such findings are consonant with Scarino's (2013) argument that teachers armed with a sound understanding of the two contrastive paradigms of language assessment, that is, standardized and local assessments, are more likely to practice self-awareness. Although similar studies on the complex interaction of LAL and language teachers' teaching practices are still very welcome and timely, this study is to address the construct of LAL itself and the perceptions of various groups of stakeholders of language testing in connection with a high-stakes test of LAL. More specifically, this study addresses the following questions:

- What are the perceived needed areas of LAL by language testing scholars?
- What is LAL for language testing instructors?
- What are test-takers' perceptions of the LAL test in terms of its difficulty?
- What are the problems in acquiring LAL in the view of test-takers?

3. Methods

3.1 Participants

Fifty-four M.A. TEFL students from Shahid Chamran University of Ahvaz and Ahvaz Azad University participated in the study. All the students had a course with the researcher in the academic year of 2013, fall semester. The participants were not asked to provide any demographic information to avoid a halo effect which might contaminate the data. However, based on the class list that the researcher had, 18 were men and 36 were women. They were typical M.A. students who were in their twenties, with a few in their early thirties. One requirement for participants was that they should not have passed the language testing course in their M.A. program. The rationale behind this limitation was that students who had already taken the language testing course in their M.A. program could not provide us with the reliable data because the M.A. language testing course might have colored their memories of their attitudes toward and preparation for the LAL. The other justification for the requirement was that such students were still fresh on their experience of taking the LAL test and the way they had prepared for it.

3.2 Instruments

A questionnaire instrument and the recent version of the LAL test were the instruments used in this study. The questionnaire, which was designed by the researcher, was written in participants' native language, Persian, in order to make sure that differential English proficiency does not interfere with the data elicitation process. It comprised seven items, three of which are not reported in this study (see Appendix). The omitted items were about participants' self-evaluation of their performance on the LAL test and the extent to which the sources they had studied for preparation proved helpful. With hindsight, however, we realized that the answer to these questions are better to be based on more reliable data like the participants' test scores on the LAL test rather than on such self-assessment measures. On the other hand, because the Center of Educational Measurement (CEM) reports only an aggregate score, not separate scores for each test module, accessing the participants' LAL test scores proved impossible.

Of the four questions reported in this study, one was about the language testing sources the participants had studied in their B.A programs, and another was about the sources they had studied in their preparing for the language testing module of the M.A. exam. The participants were instructed to supply either the name of the textbook's author or its title or both. In many cases, they remembered none but could recollect the textbook's cover. In such cases, they gave the description of the color or design of the cover page and, the researcher provided them with the name and the title of the textbook they were referring to. The next item asked participants to rank on a scale of difficulty the difficulty level of language testing test module compared to the other two modules, namely, linguistics and methodology. The last item was about the reasons why the LAL module of the tests was perceived to be more challenging.

The second instrument used in the study was the actual LAL test administered by the Center of Educational Measurement (CEM) in 2012-2013 educational year. This was the LAL test that the majority of participants of the present study had taken. Table 1 details the content domains of the test, the number of items in each content area, and other relevant information.

To ensure the reliability and validity of research instruments, the common practice in most applied linguistics articles is to hastily report an internal consistency index (i.e., test-retest, point-biserial correlations, Cronbackalph, KR-20, etc.) without considering the logic and rational of each of these procedures and whether they are appropriate for the kind of instrument at hand. The two pivotal underlying assumptions in almost all of the abovementioned procedures are unidemensionality and adequate variance in the collected data. Both assumptions are rooted in psychological testing and language tests (see Borsboon, 2005; Fulcher, 2014). These assumptions are not met in many questionnaires especially those with open-ended items or, instruments that for some reason fail to generate sufficient variance. Nor does it make sense to expect all data to meet the assumption of unidimensioanlity or divergence. Arguing on a similar ground, Hawkey (2006) invites researchers to make informed choices when it comes to assessing the reliability and validity of research instruments (see Hawkey, 2006). He further states that in washback-related questionnaires, which often contain a variety of item formats, it is "the veracity of participant responses" (p. 55) that matters not its linguistic accuracy, appropriacy or consistency.

Hawkey (2006) suggests a full list of alternative recommendations for validating the type of questionnaires we used in this study. Among such recommendations are retrospective interviews, multi-trait-multi-method qualitative procedure, and interevaluator consistency. We used retrospective interviews to ensure the veracity of responses. After each participant handed in the questionnaire, the researcher briefly reviewed the responses of each participant and in cases where there were ambiguities, students were asked for further clarifications, which were then jotted down in the margins of their completed questionnaires. As an informal mono-trait-multi-method approach, the participants were also invited to talk about their experiences of taking the LAL test around the general themes of the survey. Generally speaking, there was nothing in their talks which was in conflict with what they had reported on the survey sheets. This was taken as evidence of convergence validity of the responses.

The participants' responses to the open-ended item, which required the participants to write about the challenges they faced in learning language testing were subjected to interevaluator consistency procedure. The responses they provided lent themselves to coding and grouping into mutually exclusive categories because the nature of the question was such that participants did not produce extended narratives of personal, subjective experience. Despite this objective nature of the responses, a language testing scholar was invited to code the responses. Expectedly, the Kappa coefficient was moderately high (k=.6). It should be pointed out that the seemingly low coefficient is because of the procedure used in the estimation of Kappa. According to Brown (2005), "since Kappa represents the percentage of classification agreement beyond chance, it is usually lower than the agreement coefficient" (p. 203). Similarly, to enhance the reliability of the LAL content analysis, another language testing scholar was asked to do the categorization. The kappa coefficient of agreement was not very high but acceptable (k = .57).

3.3 Procedure

The questionnaires were administered by the researcher during two separate class sessions. The participants were under no time pressure to complete them. However, given the small number of items, it took them no more than twenty minutes. Participants seemed to be very interested in the subject of the questionnaire as it generated much debates among them.

Knowing that M.A. students at Azad University are a selected on the basis of a different entrance exam, the researcher made it clear to the participants that the questionnaire was about their preparation for the state universities entrance exam. Therefore, those who had not taken the LAL test administered by CEM or had taken it but had not prepared for it were asked not to participate. A few students announced that they had either not taken the test or had taken it but with no preparation. Such students were excluded from the study.

Data analysis was carried out using Excel 2010 and SPSS 18, with the former being utilized in describing sample characteristics and the latter for doing inferential statistics. In keeping with the idea that "a picture is worth a thousand words" (Hatch & Lazartan, 1991, p. 147), the descriptive statistics is mainly presented using histograms or column charts in Excel terms. Because of the nature of data, nonparametric statistics is mainly used in the inferential statistics section. In particular, nonparametric tests of group differences are made use of in comparing groups' attitudes about the difficulty level of the LAL test.

The rationale behind the content analysis of the LAL along the three components of assessment knowledge was to juxtapose the domain of the test with that of the language testing courses offered in undergraduate programs, and how such a gap, if it appears to exist, is realized in test-takers' preparation practices. To this aim, the language testing textbooks which test-takers reported to have studied either in their undergraduate or preparation programs was also examined. For assigning textbooks to the categories of the framework, we relied on Davies (2008) and A. Davies (personal communication, November 28, 2013). In Davies's (2008) framework, the knowledge base needed in the language assessment profession consists of practical know-how or the skills of language testing, knowledge of language and measurement, and principles that serve to guard against unethical practices and uses of assessments. It is worth noting that this framework of LAL was not imposed on the data *a priori*, rather; an inductive examination of the test's content led to the selection of the above framework among many others that could potentially serve as a framework of reference. Moreover, compared with other frameworks, it enjoyed a certain degree of generality and comprehensiveness that could be used to model the content of the LAL test as well as those of the language testing textbooks.

4. Results

4.1 Content Analysis of the LAL Test

Table 1 summarizes the content of the LAL test under study. The first point that captures attention is the disproportionate allocation of items to the three broad areas of LAL. We see that the test is mainly made up of items that tap the knowledge aspect of the LAL and less heed is paid to either the skill or the principle components:

Table 1. *Classification of the LAL Test Items*

Category	Subcategory	Item Number	Number of items	Total
Skills	Item analysis	96	1	2
	Statistics	99	1	
Principles	Critical Language testing	102	1	1
	Performance assessment	105, 92	2	
	Alternative assessment	103, 104	2	
Knowledge	Reliability and validity	100, 91	2	12
	Criterion-referenced testing	97, 98	2	
	Integrative testing	94,	2	

		95	
	Testing language skills	101	1
	Rasch Measurement	93	1

Of the couple of items related to the skills of language testing, not to be taken as testing of language skills, one question is on item analysis and the other on statistics. More specifically, the first question is about the relationship between item difficulty and item discrimination. The other item measures candidates' ability in choosing appropriate types of correlations for different research situations.

The second component of LAL, principles, is notably underrepresented in the test. As shown in Table 1, only one item targets the candidates' awareness of the principles of language assessment and that single item is about the idea that language tests are always ideologically-driven, or what McNamara (2000, p. 76) considers a radical reading of critical language testing.

Four-fifth of all the items tap the candidates' knowledge component of LAL. Performance and alternative assessment receive two items each. Likewise, an equal number of items target issues related to reliability and validity of language assessment. Two items are on criterion-referenced and integrative testing. One item is about enabling skills in testing reading comprehension and one is about the theoretical bases of Rasch measurement.

Combined together, skills and principles form only one-fifth of the LAL measure used for screening candidates. Although no recipe exists as to how much space should go into each of the triple areas of LAL, as adumbrated by Davies (2008), and much depends on the purpose, test developer, a myriad of other factors, the above distribution is yet, at most, eccentric. Davies calls for balance among the three, thus:

> For teaching, as for learning, there is a need for careful balancing of the practical (the skills) with the descriptive (the knowledge) and the theoretical (the principles). All are necessary but one without the other(s) is likely to be misunderstood and/or trivialized. (p. 336)

Given that little information is available about the rationale or the theoretical framework that test designers had based their test on, it is likely that the LAL test designers have had a test specs different from the criteria suggested by Davies above.

4.2 Language Testing Textbooks

Figure 1 illustrates the language textbooks that language testing instructors use in the B.A. programs. The first major point that Figure 1 demonstrates is the lack of diversity observed in the textbooks used, with the textbook authored by Farhady, Jafarpur, and Birjandi (1994) being used with a frequency larger than that of all the other textbooks combined. The other frequently used textbooks are Heaton (1975) and Harris (1969), with frequencies of 15 and 11, respectively:

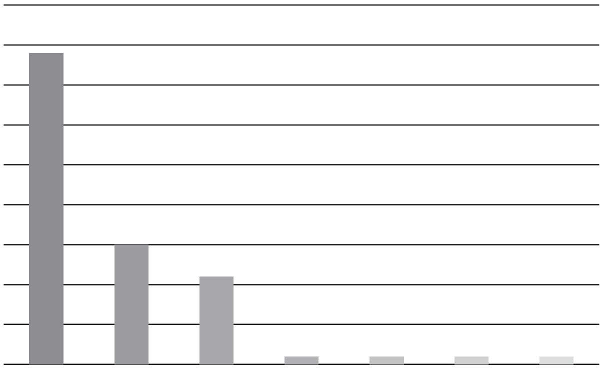

Figure 1. Language Testing Textbooks Introduced in B.A. Programs

Four other textbooks are used each with a frequency of one. Taken together, it can be concluded that the three textbooks with the highest frequencies dominate almost all language testing courses in B.A. programs. The common feature of all the three textbooks in current use is that all three have been written decades ago. The most recent among the three is Farhady et al. (1994), which was written two decades ago. The other two textbooks are each three and four decades out of date.

This reliance of language testing instructors on a few language testing textbooks, which all happen to be somehow outdated, can be taken as evidence of what they think constitutes the knowledge base in language assessment. Both the Heaton and the Harris are textbooks that deal almost exclusively with the how or the skills of language testing. As to the Farhady et al. (1994), although not among the textbooks surveyed in Davies (2008), it is very much like the Heaton in its coverage of topics. The bottom line, then, is that all the three most frequently used textbooks are heavily toward the skill trend of language testing with less, if any, coverage of the what of language assessment and no space given to the principles of language assessment.

Two scenarios are then discernible; one is that language testing instructors are not aware of or familiar with other more comprehensive language testing materials. The second possibility is that they are aware of such resources but choose the above textbooks because they believe that these textbooks cover what they think is of primary importance in language testing. If the former is true, it is surmised that language testing instructors suffer from poor LAL. If we assume the latter scenario to be the case, it is inferred that course instructors have failed to keep thoroughly abreast of the state of the art of the field, leading them to believe that the repertoire of competencies needed for one to speak the language of language assessment (Inbar-Lourie, 2008) boils down to the practical know-how of language testing.

Figure 2 depicts the sources test candidates choose to study to prepare for the language testing module of the test. The first point about Figure 2 is that except for the test preparation materials, test-takers in their test preparation practices choose to stay faithful to the same materials they studied in their B.A. programs. It is observed that the textbook most candidates choose to study is that authored by Farhady et al. (1994), followed by test preparation materials written by different private or public organizations, mainly for commercial reasons. The next two textbooks in order of frequency of use are the Heaton and Harris, as was the case in textbooks studied in B.A. programs. The only change compared to the previous diagram, though not considerable, is Bachman (1990) with a frequency of four. Overall, the pattern of textbooks in use in B.A. programs and those chosen for test preparation seems not vary considerably, as a comparison of Figures 1 and 2 reveals.

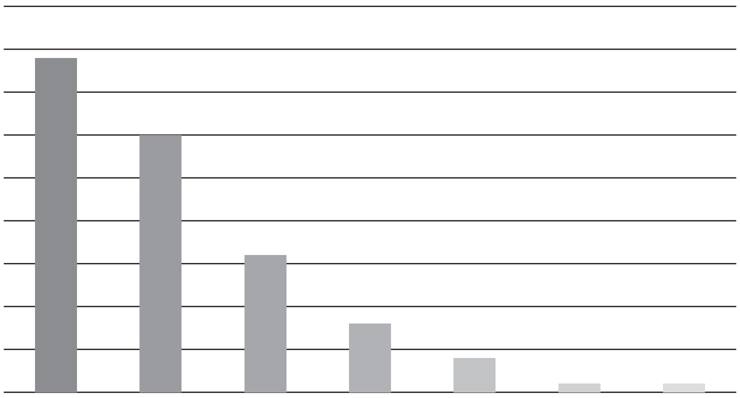

Figure 2. Materials Used for Test Preparation

As to the test preparation materials that are produced and marketed for language testing, the textbook with the highest frequency of use was examined for its content coverage (its bibliographic information is withheld to avoid working in or against the interests of the publisher or writer). The preparation textbook was found to be completely modeled on the Farhady et al. (1994) with the same chapters in the same order. The entire content consists of same-format presentation of materials, which is of dubious ethicality and educational defensibility (see Hamp-Lyons, 1998). Therefore, the content of language testing courses in B.A. programs is not at any significant variance with that of the materials used in preparing for the high-stakes test of LAL. This observation indicates that the LAL test, unlike other high-stakes tests, has failed to exert a considerable influence on the teaching and learning content that precedes it. The literature on backwash has it that the first aspect of learning and teaching that is influenced by a powerful test is the what or the content of teaching and learning (see Alderson, 2004). Why this test has not generated the conventional washback effect, though important, is beyond the scope of this paper.

4.3 The Perceived Difficulty of Language Testing Module

Figure 3 depicts the results of the question that asked participants to rate the language testing section of the M.A. entrance exam in terms of its difficulty compared to the other couple of test modules, namely, language teaching methodology and linguistics. Thirty five participants ranked it as the most difficult test module whereas those who ranked it as moderately difficult or the easiest were nineteen, combined. To test the difference between the first group, the participants who ranked LAL module as the most difficult, and the latter two groups combined, a chi-square test was run, $\chi^2(1, N =) = 4.74$, $p = .03$, which indicates that the observed difference between the number of participants who find it the most challenging test module and the number of those who do not is significant:

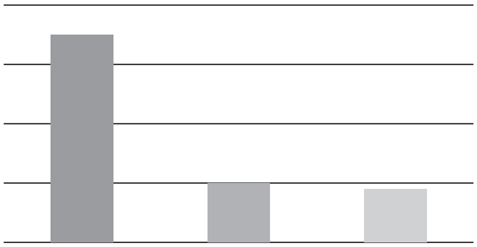

Figure 3. Perceived Relative Level of Difficulty

Two-thirds of the participants rated it as the most difficult test module; 20% of the participants rated it as the second most difficult and the rest rated language testing module as the easiest test section. Such findings are alarming for the language testing community, indicating that the field has failed to reach the wider audience it seeks to inform. It indicates that a decade after Brindley's (2001) recommendations to alleviate the alien-nature problem of language testing, not much progress has hitherto been made to alleviate fears of the stakeholders. There are multiple explanations for this observation. First, the importance accorded to language testing in B.A. programs is far below its actual importance in language teaching as well as in the entrance examination for M.A. programs. Many other authors have taken on board this disproportionate attention that is given to language testing in language programs (see Bridnley, 2001; Fulcher, 2012, among others). Juxtaposing the role of language testing in M.A. entrance exam, which we believe it deserves the status it has in the exam, with the importance given to it in language programs, we find that the two are widely discrepant. One possibility is that the lack of congruence between test content and the content of textbooks taught in B.A. programs or studied for preparation can be accounted for by what Heejeong (2013) found out about the way LAL is defined differently by nonspecialist and specialist language testing instructors. Such differences are here between test designers, who are presumably language testing scholars, and language testing instructors, for the majority of whom language testing is not the primary research area of interest.

4.4 Why Is Language Testing Perceived to Be Challenging?

The last research question posed was to see what the reasons are for the perceived difficulty level of language testing compared to other test modules. Figure 4 summarizes the reasons mentioned by participants along with their frequencies. Clearly, statistics stand out among the reasons with a frequency of 12. Next comes the lack of practical experience and instructors' reliance on lock-step, theoretical teaching. Nonexpert instructors and their failure to proceed with a consistent syllabus were also among the major complaints aired by test-takers:

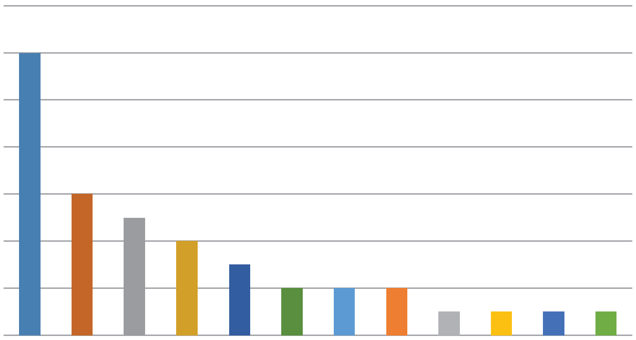

Figure 4. Reasons Given for the Difficulty of Language Testing

Other factors with less frequencies included lack of access to resources, inadequacy of teaching hours allocated to language teaching in B.A. syllabus, and confusing terminologies. The participants also cited the time constraints of the test, lack of alignment between test content and instructional materials content, inadequate attention to testing in the educational system, and abstract concepts of the field as factors which contribute to the difficulty of building a knowledge base in language assessment.

5. Discussion and Conclusion

This study demonstrated that there is a gulf between what LAL is for scholars, what it is for test makers, and what it is for language testing instructors and test-takers. For scholars, a rounded education in language assessment is a healthy balance of skills, knowledge, and principles (see Davies, 2008; Malone, 2008). For test designers, as revealed by content analysis of the LAL test, LAL is mainly a matter of knowledge and marginal attention is given to either the principles or the skills of language testing. Language testing instructors and test-takers diminish the construct further down the ladder, strip it of its knowledge and principles components and reduce it to skills. Other scholars have drawn attention to the gap between language testing experts and nonexperts (see Heejeong, 2013; Malone, 2013). However, the discrepancy between LAL test designers and what test-takers choose to invest in for preparation is not consistent with the literature on consequential validity of tests.

The gap observed between the content of LAL test and the content of materials studied for test preparation was not expected because of the immediate influence that high-stakes tests exert on teaching content. There are many possible explanations for this failure in backwash. The type and amount of washback a test has on education depends on a number of factors such as the degree to which a test is counter to common practice, the extent to which teachers and material developers are willing to innovate, and how far teachers and textbook writers contemplate appropriate methods of test preparation (Hamp-Lyons, 1998, Alderson & Hamp-Lyons, 1996). That washback failed to take place for the LAL might be attributed to lack of willingness on the part of instructors and textbook writers to innovate. Yet, another possibility for this lack of impact is the fact that education innovations often take much longer than we expect to take root (see Stoller, 1994; Wall, 1996). However, for the observed state of affairs not all the blame can be placed on the LAL test for its lack of washback as participants gave a variety of reasons for their poor LAL.

Of the many reasons cited for their language testing phobia, statistics ranked conspicuously high. This heightens the anxiety levels of practitioners because of the nature of statistics. As Brown (2013) states, "statistics anxiety is a widespread phenomenon that is not isolated in the language teaching population and that the issues involved are complex and many" (p. 354). As statistics is to blame for a considerable part of the phobia associated with language testing, a good strategy, as Brown suggests, is to turn to the literature in statistics teaching for ideas and insights. However, the fact that there was only one item requiring statistical processing on the test shows that the prevalent psychometric image that most practitioners have of language testing is not much realistic. This debilitating misconception, at least partially, stems from language testing instructors' overemphasis of skills at the expense of knowledge and principles.

To enhance the collective knowledge base of language assessment for teachers the major hurdle to overcome is to get language testing instructors and language teachers notice that the bank of experiences they have had with tests as learners is a poor and shaky foundation to build on for effective language assessment. If the language testing community manages to raise teachers' awareness about such assumptions (Scarino, 2013), which are seldom articulated, half the job is done. As Pill and Harding (2013) maintain, the problem of illiteracy in assessment is less difficult to tackle than the problem of minimal literacy because the latter necessitates first, dispelling ingrained suppositions.

Another issue that seems to have become the default practice for language textbook authors is to keep the first language of their target audience at bay at any cost. Local language testing textbooks give the impression that the whole business of the profession is something imported from the west, with no roots or history in other parts of the world. This is partly because "our field has been remarkably ahistorical" as Spolsky (1995, cited in Davies 2008, p. 330) puts it. But history does have it that language testing has been around in all parts of the world for subjectivity and identity detection purposes (see McNamara & Roever, 2006). This English-only approach may be inevitable for textbook writers who write for an international audience but there is nothing inherently wrong for local authors to, at least occasionally, give reference to some aspect of learners' first language in their texts, exercises, or examples. This is in line with the multicompetence theory that is currently in favor in the field (Cook, 2010, 1999).

Writing about the stages in the history of language testing and the type of training that was required of language instructors at each stage, Malone (2008) states that in the psychometric era "a gulf developed between instructors and tests" (p. 227). This was because certain organizations assumed responsibility for measuring learners' achievement through large scale tests and teachers were in charge of providing instruction to learners. According to Malone, this phase in the history of language testing was followed by the sociolinguistic period, in which standards-based assessment (i.e., NCLB and CEFR) gained in importance and this led to coming together of testers and teachers, which made available to teachers training in assessment. It seems reasonable to claim that language testing in Iran is yet at the psychometric era, as no far-reaching outcome-based assessment policies have been implemented yet. As a result, the gaps in LAL of language testers, language testing instructors, and language testing practitioners is not likely to entirely disappear, at least not in a foreseeable future, unless the current centralized system with its "unfortunate division of labor" (Kumaravadivelu 2008, p. 166) goes through some fundamental changes.

The importance of LAL cannot be overestimated as any education endeavor devoid of appropriate assessment, no matter how carefully designed and implemented, is doomed to failure or to go unnoticed. Put is simply, without proper measures of learning outcomes, there is no way to know the difference between effective and ineffective education. By analogy, without adequate levels of LAL, one cannot choose, design, or implement effective language assessment. Nor can one differentiate between good and bad practices in language testing. This would in turn lead to an inability to judge whether language education activities and programs are effective or not.

Given the centrality of LAL to effective language teaching and learning, it is imperative for language education policy makers, teacher educators, and language testers to take proper, long-term initiatives to enhance the LAL of all groups of LAL stakeholders: teachers, students, university instructors, language testers, etc. A first crucial step is to give more space to language testing in undergraduate and graduate language programs. It does not come with the territory that of a total of nearly 140 units on the undergraduate syllabus there is only one two-hour course allocated to language testing. To notice the Cinderella status of language testing in language programs, one only needs to note that topics of remote relevance to language teaching such as Persian literature, Arabic, history, theology, family planning, and so on, all outweigh language testing in undergraduate programs. A more ambitious step is for universities to offer graduate programs exclusively in language testing and assessment. Cambridge and Lancaster are pioneers in this regard, so other universities can benefit from their plans, experiences, and syllabuses.

This study suffered from a number of shortcomings which limit the generalizability of the findings. First, it was not based on a large number of participants; nor were the participants randomly selected. Second, the participants did not represent the population of LAL test-takers of the country, as they were mainly from the southwest province of Khouzestan and graduated from universities in the same region. Moreover, it did not measure the LAL of the participants or instructors in any direct way. Therefore, future studies with random, large samples using valid tests of LAL would be very welcome. Finally, no matter how many precautions are taken in keeping subjectivity at bay, the researcher's bias is certain to creep in in one way or another in the research process. Thus, more studies with more robust designs are direly needed to substantiate or refute the claims made in this study.

References

Alderson, J. C. (2004). Forward. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp.19-36).Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Alderson, J. C. (2009). The micropolictics of research and publication. In J. C. Alderson (Ed.), *The politics of language education: Individuals and institutions* (pp. 222-236). Bristol: Multilingual Matters.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Borsboon, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Brindley, J. (2001).Language assessment and professional development. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumly, T. McNamara, & K. O'Loughlin, (Eds.), *Experimenting with uncertainty: Essays in honor of Alan Davies* (pp. 126-127). Cambridge: Cambridge University Press.

Brown, J. D. (2005). *Testing in language programs*. New York: McGraw-Hill.

Brown, J. D. (2013). Teaching statistics in language testing courses. *Language Assessment Quarterly, 10*(3), 351-369

Cook, V. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly, 33*(2), 185-210.

Cook, V. (2010). *Translation in language teaching*. Oxford: Oxford University press.

Davies, A. (1990). *Principles of language testing*. Oxford, UK: Blackwell.

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing, 25*(3), 327–347.

Farhady, H., Jafarpur, A., & Birjandi, P. (1994).*Testing language skills: From theory to practice*. Tehran: SAMT Publication.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113-132.

Fulcher, G. (2014). Philosophy and language testing. In A. Kunan (Ed.), *The Companion to Language Assessment* (pp. 1-17). New York: John Wiley & Sons, Inc.

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

Harris, D. P. (1969). *Testing English as a second language*. New York: McGraw-Hill.

Hatch, E. M., & Lazaraton, A. (1991). *The research manual: design and statistics for applied linguistics*. Boston, Mass: Heinle & Heinle

Heaton, J. B. (1975). *Writing English language tests*. London: Longman.

Heejeong, G. (2013). Defining assessment literacy: Is it different for language testers and nonlanguage testers? *Language Testing, 30*(3), 345- 362.

James, M. (2008).Assessment and learning. In S. Swaffield, (Ed.)*, Unlocking assessment: Understanding for reflection and application.* New York: Routledge.

Kumaravadivelu, B. (2006). *Understanding language teaching: From method to postmethod*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Lantolf, J. P., & Poehner, M. E.(2008). Dynamic assessment In E. Shohamy& N. Hornberger (Eds.), *Encyclopedia of language and education* (Vol. 7)*: Language testing and assessment* (pp. 273-284). New York: Springer

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

Inbar-Lourie, O. (2013a).The special issue on language assessment literacy. *Language Testing, 30*(3), 301-307.

Inbar-Lourie, O. (2013b). *Language assessment literacy: What are the ingredients?* Plenary speech at the 4th CBLA SIG Symposium Program ΄Language Assessment Literacy- LAL, Cyprus.

Malone, M. (2008).Training in language assessment. In E. Shohamy& N. Hornberger (Eds.), *Encyclopedia of language and education* (Vol. 7)*: Language testing and assessment* (pp. 225–239). New York: Springer

Malone, M. E. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing, 30*(3), 329-344.

McNamara, T. (2000).*Language testing*. Oxford: Oxford University Press.

McNamara, T., & Roever, K. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.

Messick, S. (1996).Validity and washback in language testing. *Language Testing, 13*(3), 243-256.

O'laughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing, 30*(3), 363-380.

Pill, J.,& Harding, L. (2013).Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. *Language Testing, 30*(3), 381-402.

Popham, W. J. (2006). Needed: A doze of assessment literacy. *Educational Leadership, 63*, 84-85.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice, 48*, 4-11.

Razavipur, K., Riazi, M., & Rashidi, N. (2011).On the interaction of test washback and teacher assessment literacy. *English Language Teaching, 4*(1), 156-161.

Scarino, A. (2013). Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. *Language Testing, 30*(3), 309-327.

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

Stobart, G. (2008). *Testing times: The uses and abuses of assessment*. New York: Routledge

Stoller, L. F. (1994). The diffusion of innovations in intensive ESL programs. *Applied Linguistics, 15*(3), 300-327.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30*(3), 403-412.

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

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave.

Whitehead, M. (2010). *Physical literacy: Throughout the life course*. New York: Routledge.