Iranian Journal of Applied Linguistics (*IJAL*)

Vol. 19, No. 2, September 2016, 115-154

A Mixed-methods, Cross-sectional Study of Assessment Literacy of Iranian University Instructors: Implications for Teachers' Professional Development

Rajab Esfandiari*, *Imam Khomeini International University, Qazvin, Iran*
Razieh Nouri, *Imam Khomeini International University, Qazvin, Iran*

Abstract

Professionalism requires that language teachers be assessment literate so as to assess students' performance more effectively. However, assessment literacy (AL) has remained a relatively unexplored area. Given the centrality of AL in educational settings, in the present study, we identified the factors constituting AL among university instructors and examined the ways English Language Instructors (ELIs) and Content Instructors (CIs) differed on AL. A researcher-made, 50-item questionnaire was constructed and administered to both groups: ELIs (N = 155) and CIs (N = 155). A follow-up interview was conducted to validate the findings. IBM SPSS (version 21) was used to analyse the data quantitatively. Results of exploratory factor analysis showed that AL included three factors: theoretical dimension of testing, test construction and analysis, and statistical knowledge. Further, results revealed statistically significant differences between ELIs and CIs in AL. Qualitative results showed that the differences were primarily related to the amount of training in assessment, methods of evaluation, purpose of assessment, and familiarity with psychometric properties of tests. Building on these findings, we discuss implications for teachers' professional development.

Key words: Assessment literacy; Content instructors; English language instructors, Professionalism

Article Information:

Received: 23 March 2016 **Revised: 20 July 2016** **Accepted: 25 July 2016**

Corresponding author: Department of foreign languages, Imam Khomeini International University, Qazvin, Iran Email address: esfandiari@hum.ikiu.ac.ir

[DOI: 10.29252/ijal.19.2.115]

1. Introduction

Stiggins (1991) coined assessment literacy (AL), using it as the understanding of the principles of sound assessment. Although Brindley (2001) did not use Language Assessment Literacy (LAL), he was the first language tester credited with outlining the principles of LAL in applied linguistics. Some researchers have hypothesised possible components of AL. For example, Brindley (2001) identified three core modules, namely, the 'why', the 'what', and the 'how'. Davies (2008) proposed a three-dimensional model, including 'skills', 'knowledge', and 'principles'. DeLuca and Klinger's (2010) theoretical approach to AL included 'practice', 'theory', and 'philosophy'. Finally, Fulcher (2012) suggested 'contexts', 'principles', and 'practices' as triple components of AL.

However, these proposed elements of AL need to be backed up by empirical evidence (Harding & Kremmel, 2016). As Medland (2015) clearly stated, "Within the HE context, the concept of assessment literacy is in its infancy and accompanied by very little literature" (p. 23). The special issue of London Review of Education (2015) on AL was, therefore, timely, and as William (2015) asserted, it could "make valuable contributions to a much-needed debate about what assessment literacy might mean in practice" (p. 3). In the following sections, we elaborate on AL in education and on LAL in language assessment, examining origins, developments, controversies, competing theories, and other AL/LAL-related issues. We, then, review the literature on AL and summarise the findings from studies conducted in educational settings.

2. Review of the Related Literature

2.1. AL in Education

Messick (1989) posited that those who use assessment need to understand what the results of assessments mean and what assessment does to all people involved in the assessment process. In other words, "assessment literacy has both *evidential* and *consequential* aspects" (William, 2015, p.3). These two aspects were incorporated into Stiggins' (1991) two key questions about assessing students' achievements

[DOI: 10.29252/ijal.19.2.115]

about two decades ago: "(1) What does this assessment tell students about the achievement outcomes we value? [and] (2) What is likely to be the effect of this assessment on students?" (p. 535).

Posing these two key questions, Stiggins coined assessment literacy, defining it as the understanding of the principles of sound assessment and as a way of defining certain kinds of assessment skills teachers need. Expanding on Stiggins' brief definition, Webb (2002) proposed a more comprehensive definition with three main elements, including "the knowledge of means for assessing what students know and can do, how to interpret the results from these assessments, and how to apply these results to improve students learning and program effectiveness" (p. 1).

Recognizing the importance of AL, Ainsworth and Viegut (2006) pointed out that teachers were ill-prepared and were not given the tools they needed to help students succeed. As a result, Ainsworth and Viegut introduced a newly defined conceptual framework of AL that examined the broad picture of teachers' AL. Ainsworth and Viegut's model "has the potential to provide teachers with assessment data that they could use to intervene instructionally whether for remediation or acceleration" (Braney, 2010, p. 10). Using this model, Ainsworth and Viegut defined AL as "the ability to understand the different purposes and types of assessment in order to select the most appropriate type of assessment to meet a specific purpose" (p. 53).

Newfields (2006) addressed another aspect of AL, placing a stronger emphasis on the people involved. Rather than conceptualizing AL as a single concept with some sort of unitary meaning as a set of given skills shared among all people, Newfields believed that AL represents a wide matrix of skills varying significantly from population to population. In Newfields' words, "for students [AL] largely means knowing how to perform well on exams. For teachers, it is associated with the ability to grade students ethically and accurately. And for professional test developers, every facet of their work hinges assessment literacy" (p. 50).

As the above paragraphs in this section show, some experts just use AL for a knowledge base regarding test design and measurement endeavors, while others have taken a broader view and emphasize the social aspects of assessment and the

[DOI: 10.29252/ijal.19.2.115]

influence of context within which it is carried out. In the following section, we discuss AL as it is conceptualized in language assessment.

2.2. Assessment Literacy in Language Assessment

Although Brindley (2001) did not specifically address LAL, he was the first language tester who offered an outline of LAL for development in language assessment, which includes language knowledge components required for conducting assessment in educational contexts. Brindley proposed a framework consisting of three core modules: (a) the theoretical basis for language tests and the description of traits (the what), (b) methods of language test development and evaluation process (the how), and (c) the reasoning and rationale for language assessment (the why) (Inbar-Lourie, 2008; Scarino, 2013; Shohamy, 2008).

Reviewing Brindley's framework, Inbar-Lourie (2008) elaborated on LAL and what it meant for language teachers. First, Inbar-Lourie defined LAL as "having the capacity to ask and answer critical questions about the purpose of assessment, about the fitness of tools being used about testing conditions and about what is going to happen on the basis of the results" (p. 389). Inbar-Lourie, then, added that LAL comprises "layers of assessment literacy skills combined with language specific competencies form a distinct entry that can be referred to as language assessment literacy" (p. 389). Similarly, Pill and Harding (2013) argued that LAL might be understood as possessing "a repertoire of competencies that enable an individual to understand, evaluate and, in some cases, create language tests and analyze test data" (p. 382). However, Pill and Harding pointed that LAL includes all individuals engaging in language assessment practices, not just language teachers.

A few leading figures in language assessment have made serious attempts to develop a working model of LAL and specify its underpinning elements. Following Boyles (2005) and Inbar-Lourie's (2008) suggestion for establishing a framework of core competencies of AL, Davies (2008) suggested a three-dimensional model which provides a detailed description of what LAL entails. From Davies' point of view, LAL can be described as consisting of necessary training regarding

[DOI: 10.29252/ijal.19.2.115]

appropriate methodologies such as item writing, statistics, test analysis, and software programs for test delivery (skills), relevant background about different models of language learning, language teaching, and language testing (knowledge), and issues such as fairness, impact, ethicality, and professionalism in the field (principles).

Fulcher (2012) conducted a study to elicit the assessment training needs of instructors. Criticizing DeLuca and Klinger (2010) for the oversimplification of the definition of LAL, Fulcher offered an expanded definition of LAL which embraces the missing sociopolitical perspective. Based on his findings, Fulcher offered the following comprehensive and empirically-driven definition of LAL for language teachers as

> The knowledge, skills and abilities required to design, develop, maintain or evaluate large-scale standardized and/or classroom based tests, familiarity with test processes, and awareness of principles and concepts that guide and underpin practice, including ethics and codes of practice. The ability to place knowledge, skills, processes, principles and concepts within wider historical, social, political and philosophical frameworks in order understand why practices have arisen as they have, and to evaluate the role and impact of testing on society, institutions, and individuals. (p. 125)

Assessment specialists and researchers have provided various definitions and an integration of different competencies as explained in the foregoing sections. In light of the burgeoning research on AL/LAL and further advancements in search of its components from different perspectives, LAL still remains "a riddle, wrapped in mystery, inside an enigma" (Kumaravadivelu, 2006, p. 20). As a matter of fact, this concept has evolved over time, and there has never been unanimity regarding its definition, competencies, and possible constituencies. As Harding and Kremmel (2016) rightly asserted,

> Establishing an agreed-upon base of component areas of LAL, charting a realistic trajectory of development, and ensuring that LAL is tailored to the

[DOI: 10.29252/ijal.19.2.115]

> needs of different stakeholders (including the different needs of those within stakeholder groups) might thus present the greatest challenges to be faced by those involved in promoting LAL. (425)

2.3. A Selective Review of Empirical Studies on AL

As aptly put by Stiggins (1991), "we are a nation of assessment illiterate" (p. 535). As such, Popham (2004) regarded lack of appropriate training as "professional suicide" (p. 82). Researchers, therefore, have empirically investigated AL from different perspectives. As a result, the last two decades have witnessed a burgeoning number of studies addressing AL in education and language assessment. These studies are summarized chronologically below.

Plake, Impara, and Fager (1993) conducted one of the first studies to examine AL. Five hundred and fifty-five teachers and 286 administrators completed a two-part assessment questionnaire. According to the results, teachers performed well on administering, scoring, and interpreting results and poorly on communicating test results. The teachers who received training scored significantly higher on the standards than those who did not.

Volante and Fazio (2007) explored the AL of 69 primary teacher candidates enrolled in a four-year program in a Canadian college. Using convenience sampling, candidates were asked to complete a survey consisting of several open-ended and close-ended items. The findings revealed that participants rated themselves low in AL regardless of their participation in various levels of the program, demonstrated lack of knowledge regarding assessment methods especially formative assessment, asked for additional training in authentic assessment approaches, and endorsed the development of specific courses to improve classroom assessment and evaluation.

DeLuca and Klinger (2010) also examined an assessment education program in Canada. They administered a questionnaire to 288 teacher candidates in all subject areas enrolled in a teacher education program. The results also revealed that participants choosing to enrol in an educational assessment course had considerably higher levels of confidence than those who did not have any

[DOI: 10.29252/ijal.19.2.115]

instruction in assessment. The authors attributed the increased confidence to participants' enrolment in assessment courses. The findings supported the need for the instruction of specific topics including reporting achievement, modifying assessment, and developing items in courses for the development of AL among teachers.

Employing a mixed-methods approach, Ogan-Bekiroughlu and Suzuk (2014) conducted a study to address 28 pre-service Turkish physics teachers. The findings from both phases of the study showed teachers better understand theoretical dimensions of AL; however, teachers had considerable difficulty in bridging the gap between the theoretical and practical knowledge of AL.

AL in second language education has paramount importance as "it is a commodity needed by teachers for their own long-term well-being, and for the educational well-being of their students" (Popham, 2009, p. 11). Therefore, the following two studies have examined the AL of language instructors.

Fulcher (2012) was one of the well-known assessment researchers who developed, piloted, and delivered a survey on the Internet to uncover the assessment training needs of 278 language teachers that can be used for the creation of educational materials and programs in language assessment. From the detailed analysis of the survey regarding the topics which instructors recognize as necessary to be included in a language testing course, four factors labelled as test design and development, large-scale standardized testing, classroom testing and washback, and validity and reliability were identified.

To measure 878 foreign language teachers' AL in different areas of language testing and to gain an understanding of their perceived training needs in this area, Vogt and Tsagari (2014) conducted a mixed-methods study. Vogt and Tsagari reported that the majority of instructors received either very little or no training in AL. In other words, teachers' AL seemed to be undeveloped and they learned the important elements of classroom practices such as giving grades based on experience. Given the insufficient knowledge, the instructors believed that their training did not prepare them sufficiently for their work.

As the above review reveals, none of the studies have investigated AL of English language instructors and content instructors. In Iran, similar to other

[DOI: 10.29252/ijal.19.2.115]

countries which have test-driven educational contexts, teachers determine whether students should pass a course, be promoted to the next higher level, or repeat it. The prime objective of this study was, therefore, to determine whether there is an AL for English language instructors (ELIs) and content instructors (CIs). The evaluation of their AL will contribute substantially to touching upon the central question of how we understand and define AL and how it develops and matures over time. The present study was also aimed at identifying in what ways ELIs differ from CIs in the underlying components of AL. Finally, we were interested in knowing if the differences between ELIs and CIs in terms of Al were significant.
Following the foregoing goals, the research questions addressed in this study are as follows:

1. What is assessment literacy for university instructors?
2. In what way(s) does the assessment literacy differ between English Language Instructors (ELIs) and Content Instructors (CIs)?

3. Method

3. 1. Participants

A convenience sample of three 340 Iranian male and female university instructors (ELIs and CIs) teaching BA, MA, and PhD students participated in this study. ELIs included those university instructors who were teaching English Language Teaching, English Literature, Linguistics, and Translation courses, but CIs included those university instructors teaching content areas including Philosophy, Psychology, Economics, Political science, History, Theology, Management, Law, Engineering, Physics, Chemistry, Mathematics, Architecture, Agriculture, Geology, Accounting, Statistics and Geography at Iranian universities. The instructors were MA and PhD holders and their teaching experience generally ranged from 5 to 20 years.

Approximately, 10% of the participants were also chosen to participate in a follow-up semi-structured interview. Thirty instructors, 15 ELIs and 15 CIs, were interviewed.

[DOI: 10.29252/ijal.19.2.115]

3.2. Materials

3.2.1. Questionnaire for Assessment Literacy (QAL)

Both ELIs and CIs were administered QAL consisting of 50 items. The items required the participants to self-rate their current level of QAL. All items were rated on a 5-point Likert-type scale ranging from 1 to 5. The instructors were asked to decide where they fit in categories 'not at all' (1), 'small degree' (2), 'moderate degree' (3), 'high degree' (4), and 'very high degree' (5). The highest and the lowest possible self-ratings in this regard are 250 and 50, reflecting maximised and minimised self-perceived literacy. The completion of the questionnaire lasted approximately between 15~20 minutes (See appendix). More information about the development of the questionnaire is given in 'Procedure' section.

3.2.2. Semi-structured Interview

The second source of data collection was a semi-structured interview including 14 questions developed by the researchers. The questions were mostly concerned with common test methods, alternative assessment techniques, self-evaluation of the amount of instructors' knowledge regarding assessment, the training that the instructors had received regarding assessment, their perceived needs for further training, and suggested methods for promoting AL in universities and keeping themselves abreast of the latest developments in testing, modification of their teaching methodology according to feedback received on test results, ethical issues in testing, and the role of modern technology including computers in administering and scoring tests.

3.2.3. Procedure

An explanatory sequential design was implemented in this study for complementarity purposes (Cresswell & Plano Klark, 2011). The purpose of using a blend of methods was to elaborate, enhance, illustrate, clarify, and enrich the results from the integration of quantitative and qualitative data within the study.

[DOI: 10.29252/ijal.19.2.115]

The procedures used to collect data and statistical tests used to analyse data are described in the following sections.

3.2.3.1. Data Collection

To collect data quantitatively, the researchers used the QAL. The following stages were followed to develop the QAL. First, the literature on AL was fully reviewed and four major textbooks written by renowned experts in the field of assessment, including Bachman (1990), Hughes (2003), Fulcher and Davidson (2007), and Farhady, Jafarpur, and Birjandi (1994) were consulted for writing the items. We also consulted some existing questionnaires on AL. Some sixty items were written.

Next, the items were checked for clarity, comprehensibility, relevance, and wording. A few items were dropped on the ground of little relevance or considerable overlap with other items, wording of several others was modified, and some items were reordered to enhance the validity of the responses.

As QAL was to be administered to both ELIs and CIs, the third stage was to translate the QAL into Persian. To attain this goal, the researchers adopted a functional equivalence rather than a literal translation approach. In order to check the quality of the translation, a number of procedures were followed. First, the researchers drew upon the fifth edition of one of the best-selling Persian books (Saif, 2009) covering almost all of the terminologies used in the questionnaire. Second, some assessment experts supplied feedback on the comprehensibility of translated items in terms of both linguistic and content-related issues. The feedback from the experts was used for further revision and refinement of the translation to avoid any confusion on the part of CIs.

Prior to the administration of QAL, the bilingual version of questionnaire was pilot tested on a group of volunteers (N = 57) in order to estimate its internal consistency reliability and re-word or re-scale any questions that are not answered as expected. Cronbach's alpha turned out to be .97 for QAL, indicating a high level of reliability.

The QAL was administered to both ELIs and CIs to seek their opinions about AL. The instructors were informed about the purpose of the study and they were

[DOI: 10.29252/ijal.19.2.115]

assured that their responses would remain confidential. Those who voluntarily participated in the study were provided a cover letter and the questionnaire along with brief instructions on the process of completing the questionnaires at the beginning of each section.

An online version of the QAL using Google Docs was created. During the period between June 22 and 22 September, 2015, ELIs and CIs in Iran were contacted via e-mails. The message in these e-mails included information about the research, request to complete the questionnaire, and the link of the QAL. Two weeks after the initial posting of the web-based version, instructors were sent a reminder about completing the instrument. As ELIs completed the questionnaire and submitted it, their responses were automatically loaded into a database on the web server, from which they were downloaded onto Microsoft Excel. The instructors completed and submitted 340 questionnaires, but the researchers decided to remove a few incomplete ones. Finally, 310 questionnaires were used for data analysis.

To collect data qualitatively, a semi-structured interview was developed. In the first step of designing the interview, broad questions were formulated. Then, in the second step, two assessment specialists omitted, or modified, a few questions, framed new ones, and put them in a logical sequence. Content validity of the interview was ensured through review of the items by two experts who assessed the questions in terms of comprehensibility, clarity, and relevance. Each interview session attended individually by the interviewees lasted approximately 15 minutes. The researchers found that instructors had very busy schedules; therefore, based on their own schedule, the instructors were interviewed at their own convenience. Interviews were held on an individual basis, because in group interviews, respondents may remain silent to preserve their professional face or their attitudes may have an impact on what others express (Krueger & Casey, 2000; Mackey & Gass, 2005).

3.2.3.2. Data Analysis

At the very first stage after data collection, quantitative data were processed using the 21th version of IBM SPSS software. For Likert-scale items of the questionnaire,

[DOI: 10.29252/ijal.19.2.115]

three exploratory factor analyses were run in order to identify the factors of AL. For this phase of data analysis, we got some additional information from another statistical program developed by Watkins (2000). Monte Carlo PCA (Principal Component Analysis) supported our decision to retain identified factors by SPSS. Next, an independent-samples t-test and fifty Mann-Whitney U tests were conducted respectively to determine if there were statistically significant differences between ELIs and CIs regarding AL.

In the second phase of data analysis, transcripts of interviews were content analysed and, during the analytic process, a set of codes was assigned to responses. Through this procedure, data were reviewed in a reiterative manner and coding system applied to one interview was used repeatedly throughout the remaining ones (Saldana, 2012) to discover thematic trends and patterns emerging from the coded data. Then, the patterns were clustered and grouped together not just because they were precisely and discretely bounded, but because they consisted of differences. This procedure is advocated by Charmaz (2014) who succinctly stated that "coding generates the bones of the analysis … [and] integration will assemble those bones into a working skeleton" (p. 45). Finally, the emergent patterns were compared to identify salient themes.

4. Results

4.1. Quantitative Results

4.1.1. Investigation of the First Research Question

To answer the first research question, 50 items of ALQ were subjected to factor analysis using a principal components analysis method to identify the underlying structure of AL. A variety of factor analysis criteria, including (a) sample size (+ 150 respondents or at least 300 cases); (b) Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy; (c) Bartlett's test of sphericity; (d) inter-item correlations equal to, or greater than .30; (e) greater-than-.3 communalities; and (f) cut-off factor loadings .30 were used. The procedure is described in greater detail step by step below.

Prior to performing rotation, the suitability of data for factor analysis was assessed. With the overall sample size including 310 instructors, after the removal

[DOI: 10.29252/ijal.19.2.115]

of 30 cases with missing data, this condition was satisfied in the present study. The inspection of the correlation matrix revealed the presence of many coefficients of .3 and above, with communalities ranging from .35 to .76. The Kaiser-Meyer-Olkin value was .957, exceeding the recommended value of .6 and very close to 1, which is 'superb', using the adjective Kaiser (1974) used. Also, Bartlett's test of sphericity reached statistical significance (p = .001) (requirement of p <.05), supporting the factorability of the correlation matrix.

Eigenvalues greater than 1 and a scree plot were used to determine the number of factors. Seven factors were identified, accounting for 67.708% of total variance, with 45.78% for Factor 1, 6.67% for Factor 2, 4.57% for Factor 3, 3.30% for Factor 4, 2.89% for Factor 5, 2.26% for Factor 6, and 2.21% for Factor 7.

To decide on the number of factors to retain, parallel analysis was run. The program asked for three pieces of information: the number of variables (in our case, 50 items), the number of participants (in our case, 310), and the number of replications (the program default requires 100). Then, we systematically compared eigenvalues obtained from SPSS for seven factors with the corresponding values from the random results generated by parallel analysis. The values larger than the criterion values from parallel analysis were retained. Results are summarized in Table 1. As can be seen in Table 1, only three of seven factors were to be retained.

Table 1

Comparison of eigenvalues from the first PCA and criterion values from parallel analysis

Component number	Eigenvalues from PCA	Criterion values from parallel analysis	Decision
1	22.890	1.871	Accepted
2	3.336	1.779	Accepted
3	2.288	1.715	Accepted
4	1.652	1.663	Rejected
5	1.448	1.610	Rejected
6	1.133	1.564	Rejected
7	1.107	1.524	Rejected

[DOI: 10.29252/ijal.19.2.115]

PCA was again conducted and three-factor solutions were examined. The results revealed the presence of three factors with eigenvalues exceeding 1. The three-factor solution explained a total of 57.02% of the variance, with Factor 1 contributing 45.78%, Factor 2 contributing 6.67%, and Factor 3 contributing 4.57% to the total variance.

The correlation between the components was not low; therefore, Oblique rotation with Direct Oblimin technique was conducted in order to aid in the interpretation of these three components. Using the highest loadings on factors, we labelled the factors as follows: theoretical dimension of testing, test construction and analysis, and statistical knowledge.

The first factor was labelled 'theoretical dimension of testing' on which Item 1 (accountability); Items 2 (validity), 3 (reliability), 4 (authenticity), and 6 (interactiveness) as aspects of test usefulness; Item 8 (proper use of tests); Item 9 (washback and impact); Item 10 (consequences of tests); Item 11 (test bias); Item 12 (theories of testing); Item 13 (testing models and frameworks); Item 15 (functions of tests); Item 16 (basic concepts in testing), Item 17 (history of testing); Item 18 (uses of tests in educational programs); and Item 20 (testing in relation to curriculum) highly loaded.

The second factor was labelled 'test construction and analysis', since Items 36 (determining test function and form), 37 (planning tests), and 38 (preparing items) as parts of test construction process and Item 24 (conducting item and test analyses) highly loaded on it.

Finally, the third factor was labelled 'statistical knowledge' which included Items 46 (measurement of central tendency) and 47 (measurement of variability) and Item 48 (using and interpreting inferential statistics).

4.1.2. Investigation of the Second Research Question

The second research question sought to examine in what ways ELIs differ from CIs in AL. To answer this question, two other factor analyses were run separately for ELIs and CIs, respectively. The results are presented below step by step based the procedures adopted for the first research question.

[DOI: 10.29252/ijal.19.2.115]

4.1.2.1. Factor Analysis for ELIs

Factor analysis (using PCA) was conducted and the suitability of data was assessed: the value of the KMO measure of sampling adequacy yielded .934, which is 'marvelous', using the adjective Flynn and Kunkel (1987) used, Bartlett's test was statistically significant (p = .001), and the inspection of the correlation matrix revealed the presence of many coefficients of greater than .3 with communalities ranging from .35 to .78.

Extracting factors with eigenvalues greater than 1 and using scree plot as a guide left us with eight factors accounting for 70.46% of total variance including the following amounts of variance: Factor 1 (46.23%), Factor 2 (6.21%), Factor 3 (4.17%), Factor 4 (3.72%), Factor 5 (2.79%), Factor 6 (2.63%), Factor 7 (2.49%), and Factor 8 (2.17%).

To identify the correct number of components to retain, the eigenvalues generated from SPSS were compared with those obtained from Parallel analysis (this time 155 respondents). The results presented only three components with eigenvalues exceeding the criterion values (Table 2).

Table 2
Comparison of eigenvalues from the second PCA and criterion values from parallel analysis

Component number	Eigenvalues from PCA	Criterion values from parallel analysis	Decision
1	23.118	2.312	Accepted
2	3.107	2.169	Accepted
3	2.089	2.063	Accepted
4	1.864	1.974	Rejected
5	1.399	1.885	Rejected
6	1.317	1.813	Rejected
7	1.248	1.746	Rejected
8	1.089	1.684	Rejected

[DOI: 10.29252/ijal.19.2.115]

Once the number of factors was determined, PCA was again run for the three factors explaining 46.23%, 6.21%, and 4.17% of the total variance, respectively.

Having reached a suitable component correlation, the factors were obliquely rotated. Using the highest loadings on factors, we labelled them for ELIs as follows: statistical knowledge, test construction and analysis, and theoretical dimension of testing.

The first factor, statistical knowledge, loaded on Items 46 (measurement of central tendency), 47 (measurement of variability), 48 (using and interpreting inferential statistics), 49 (using and interpreting advanced statistics), and 50 (using and interpreting more modern statistical tests).

The second factor, test construction and analysis, involved Items 29 (developing and using selected-response assessment), 30 (developing and using constructed-response assessments), and 31 (developing and using personal response assessments) constituting various test techniques; Items 36 (determining test function and form), 37 (planning tests), 38 (preparing items), and 39 (reviewing items) related to test construction process; and Item 24 (conducting item and test analyses).

The third factor, theoretical dimension of testing, was composed of Item 1 (accountability); Items 2 (validity), 3 (reliability), 4 (authenticity), and 6 (interactiveness) making up test usefulness; Item 7 (fairness and ethics in assessment); Item 9 (washback and impact); Item 10 (consequences of tests); Item 11 (test bias); Item 12 (theories of testing); Item 13 (testing models and frameworks); and Item 17 (history of testing).

4.1.2.2. Factor Analysis for CIs

The KMO value was .81, which is ‘great’, using the adjective Hutcheson and Sofroniou (1999) used. Bartlett’s test of sphericity was statistically significant (p = .001). Other factor analysis criteria were also met: The majority of items showed correlation coefficients of greater than .3, with commonalties ranging from 0.52 to 0.89. These pieces of information showed suitability of factor analysis for CIs.

[DOI: 10.29252/ijal.19.2.115]

Extracting factors with eigenvalues greater than 1 and using scree plot as a guide left us with 11 factors accounting for 69.06% of total variance including the following amounts of variance: Factor 1 (23.98%), Factor 2 (10.04%), Factor 3 (7.73%), Factor 4 (5.07%), Factor 5 (4.63%), Factor 6 (3.72%), and Factor 7 (3.16%), and Factor 8 (3.11%), Factor 9 (2.76%), Factor 10 (2.53%), and Factor 11(2.28%).

Parallel analysis was run and eigenvalues obtained from SPSS for eleven factors were systematically compared with the corresponding values generated by parallel analysis. As can be seen in Table 3, only six of 11 factors were to be retained.

Table 3

Comparison of eigenvalues from the third PCA and criterion values from parallel analysis

Component number	Eigenvalues from SPSS PCA	Criterion values from parallel analysis	Decision
1	11.99	2.30	Accepted
2	5.02	2.16	Accepted
3	3.86	2.06	Accepted
4	2.53	1.96	Accepted
5	2.31	1.88	Accepted
6	1.86	1.81	Accepted
7	1.58	1.74	Rejected
8	1.55	1.68	Rejected
9	1.38	1.62	Rejected
10	1.26	1.56	Rejected
11	1.14	1.50	Rejected

Rerunning PCA, the final six-factor solution accounted for 55.19% of the total variance: Factor 1 (23.98%), Factor 2 (10.04%), Factor 3 (7.73%), Factor 4 (5.07%), Factor 5 (4.63%), and Factor 6 (3.72%).

[DOI: 10.29252/ijal.19.2.115]

Imposing Direct Oblimin rotation, each factor was represented by a number of strongly loaded items. Considering the main loadings, we labelled them for CIs as: theoretical dimension of testing, knowledge of test construction, employment of test techniques, knowledge of descriptive and inferential statistics, testing in relation to education, and interpretation of test results.

Factor 1 was labelled 'theoretical dimension of testing', because based on loading patterns, the following items loaded strongly on this factor: Items 2 (validity), 3 (reliability), and 4 (authenticity); Item 8 (proper use of tests); Item 9 (washback and impact); Item 10 (consequences of tests); Item 11 (test bias); Item 12 (theories of testing); and Item 13 (testing models and frameworks).

Items 36 (determining test function and form), 37 (planning tests), and 38 (preparing items) as steps of test construction process loaded strongly on Factor 2.

Items 29 (developing and using selected-response assessment), 30 (developing and using constructed-response assessments), and 31 (developing and using personal response assessments) loaded strongly on Factor 3.

Items 46 (measurement of central tendency), 47 (measurement of variability), and 48 (using and interpreting inferential statistics) loaded strongly on Factor 4.

Item 18 (use of tests in educational programs), Item 19 (examination of different models of learning in testing), and Item 20 (testing in relation to curriculum) strongly loaded on Factor 5.

Finally, Item 24 (conducting item and test analyses), Item 27 (using different types of interpretation), and Item 28 (realizing limitations of test result interpretation) strongly loaded on Factor 6.

Scores for each item of the questionnaire and total scores for ELIs and CIs were compared by conducting an independent-samples t-test using Welch's procedure and 50 Mann-Whitney U tests.

Examination of the results revealed statistically significant differences between ELIs and CIs for the total score, $t = 14.52$, $p = .001$, $df = 269.998$ using Welch's procedure, $d = .40$, as well as scores on all Likert-scale items (shown in Table 4). In all cases, ELIs scored significantly higher than did CIs.

[DOI: 10.29252/ijal.19.2.115]

Table 4
Mann-Whitney U tests for all items of ALQ

Item	N		Mean Rank		Mann-Whitney U	Z	Sig.	d
	ELIs	CIs	ELIs	CIs				
1	155	155	179.27	131.73	8327.500	-4.925	0.001	0.07
2	155	155	204.48	106.52	4421.000	-9.893	0.001	0.31
3	155	155	210.68	100.32	3459.500	-11.094	0.001	0.39
4	155	155	206.06	104.94	4175.500	-10.178	0.001	0.33
5	155	155	182.07	128.93	7893.500	-5.498	0.001	0.09
6	155	155	190.80	120.20	6541.000	-7.221	0.001	0.16
7	155	155	172.31	138.69	9407.000	-3.505	0.001	0.03
8	155	155	188.79	122.21	6852.500	-6.973	0.001	0.15
9	155	155	192.36	118.64	6299.500	-7.634	0.001	0.18
10	155	155	199.34	111.66	5218.000	-8.889	0.001	0.25
11	155	155	189.27	121.73	6778.000	-6.832	0.001	0.15
12	155	155	211.50	99.50	3332.000	-11.244	0.001	0.40
13	155	155	206.16	104.84	4160.500	-10.212	0.001	0.33
14	155	155	199.72	111.28	5158.500	-9.041	0.001	0.26
15	155	155	207.79	103.21	3907.500	-10.611	0.001	0.39
16	155	155	201.74	109.26	4845.000	-9.406	0.001	0.28
17	155	155	207.94	103.06	3884.500	-10.575	0.001	0.36
18	155	155	194.35	116.65	5991.500	-7.958	0.001	0.20
19	155	155	199.50	111.50	5192.500	-8.924	0.001	0.25
20	155	155	190.58	120.42	6574.500	-7.208	0.001	0.16
21	155	155	211.17	99.83	3383.500	-11.228	0.001	0.40
22	155	155	188.04	122.96	6969.000	-6.654	0.001	0.14
23	155	155	198.64	112.36	5326.000	-8.713	0.001	0.24
24	155	155	187.84	123.16	7000.000	-6.585	0.001	0.14
25	155	155	178.22	132.78	8490.500	-4.607	0.001	0.06
26	155	155	205.56	105.44	4252.500	-10.090	0.001	0.32
27	155	155	193.68	117.32	6094.000	-7.729	0.001	0.19

[DOI: 10.29252/ijal.19.2.115]

28	155	155	197.35	113.65	5526.500	-8.528	0.001	0.23
29	155	155	193.92	117.08	6057.000	-7.851	0.001	0.19
30	155	155	193.43	117.57	6133.000	-7.799	0.001	0.19
31	155	155	191.41	119.59	6447.000	-7.296	0.001	0.19
32	155	155	172.38	138.62	9396.000	-3.539	0.001	0.04
33	155	155	195.10	115.90	5875.000	-8.026	0.001	0.20
34	155	155	190.65	120.35	6564.500	-7.157	0.001	0.16
35	155	155	166.59	144.41	10293.500	-2.285	0.022	0.01
36	155	155	186.48	124.52	7210.500	-6.412	0.001	0.13
37	155	155	173.59	137.41	9208.000	-3.762	0.001	0.04
38	155	155	176.63	134.37	8737.500	-4.473	0.001	0.06
39	155	155	190.43	120.57	6598.000	-7.198	0.001	0.16
40	155	155	191.51	119.49	6431.000	-7.237	0.001	0.16
41	155	155	194.15	116.85	6022.500	-7.798	0.001	0.19
42	155	155	185.77	125.23	7320.500	-6.135	0.001	0.12
43	155	155	202.39	108.61	4745.000	-9.453	0.001	0.28
44	155	155	168.85	142.15	9943.000	-2.788	0.005	0.02
45	155	155	178.17	132.83	8498.500	-4.593	0.001	0.06
46	155	155	180.46	130.54	8144.000	-5.035	0.001	0.08
47	155	155	191.81	119.19	6384.000	-7.296	0.001	0.17
48	155	155	204.71	106.29	4385.000	-9.895	0.001	0.31
49	155	155	201.54	109.46	4877.000	-9.519	0.001	0.29
50	155	155	191.13	119.87	6489.500	-7.561	0.001	0.18

4.2. Qualitative Results

In this part, instructors' responses are examined. Interview questions are presented one by one, and the responses of all 30 instructors are examined on each question. Themes are reported, and wherever necessary, similarities and differences are noted. Abbreviations are used for anonymity purposes.

[DOI: 10.29252/ijal.19.2.115]

Question 1: What common test techniques do you use to test your students' knowledge and ability? Please, explain your choice.

Analysis of responses to this question showed that although ELIs prefer to strike a balance between different types of test techniques, they normally use one, or sometimes two of them, based on the nature of the courses they teach. Although CIs were aware of different test techniques, they mostly preferred to use essay-type tests to measure their students' amount of learning. For CIs, the choice of techniques is also mainly determined by the nature of the course. For example, ELI-2 commented that

> *I would favor a mixture of all types of techniques because it makes evaluation more accurate by allowing students who do not do well with a testing format to excel in other areas. But, obviously, for specific courses it is not possible due to the manageability matters.*

CI-10 responded that

> *There are many different types of test techniques, each with its own strengths and weaknesses. But almost all the time, open-ended questions are my choice because psychology students are required to explain the reasoning behind their answers. This can only be waived in very exceptional circumstances.*

Question 2: Do you ever try to use more modern test techniques to make a test and test your students? Please, name them (if any).

Analysis of responses indicated the popularity of a variety of alternative assessment methods among ELIs. ELIs use more modern test techniques, including portfolios, journal reports, and performance assessment, to help students to be autonomous, to master learning objectives, and to evaluate themselves. Although some CIs commented that they would use a range of modern assessment

[DOI: 10.29252/ijal.19.2.115]

techniques, others noted they would apply traditional methods. Interestingly, the analysis of the responses revealed that most ELIs' responses were aligned with more recent formative (i.e., assessment for learning) assessment, but their CIs counterparts used assessment primarily for traditional summative purposes (i.e., assessment of learning). The following two quotes illustrate these points.

> *I use not only exams but also alternative assessment methods. I require my students especially MA students to do research, deliver lectures, or hand over projects, thus they become responsible for their own learning. (ELI-3)*

> *No, I am ill-informed about modern test techniques. Besides, I believe that modern techniques are not always the best and the most effective ways to evaluate students' performance. Therefore, I assess my students as I was assessed as a student. (CI-5)*

Question 3: Do you ever try to attend testing workshops to update your knowledge on testing issues?

While, on the basis of the analysis of the responses, it can be inferred that a number of ELIs tended to attend workshops, the majority of them as well as CIs did not get the chance due to the unavailability of workshops and lack of information about the workshops, as the following quotes illustrate: "Actually, not yet. Few numbers of workshops, almost all of which were far away from here, have been held" (ELI-11). "I had not the chance to do so, because no workshop was held, or I was not informed about it" (CI-15).

Question 4: How often do you keep yourself abreast of the latest developments in testing?

Some ELIs believed that they regularly updated themselves because there was a need to do so, but some ELIs never deemed it necessary to keep themselves up–to-date. Similarly, CIs do not usually tend to keep up with the latest developments in

[DOI: 10.29252/ijal.19.2.115]

testing because they see no necessity for it. One of the language instructors, for example, remarked that *"Most of the time we do, because there are some new discoveries which are specific to our courses and can only be followed by those teachers who are into testing" (ELI-10).* A content instructor commented that *"Never do I update myself. Up to now, our background knowledge in testing obviates the need for further learning" (CI-8).*

Question 5: Have you been trained to construct, administer, and score a test?

Almost all of the ELIs received some training, although the training came through university courses at BA, MA, and PhD programs. On the other hand, either CIs received no training, or they referred to the assessment courses they needed to pass when they were students, or when they were being prepared as would-be teachers at Teacher Training Centers.

The following comments show what ELIs and CIs thought of training. *"Indeed, as BA, MA, and Ph.D. students, we completed several prescribed credits of testing" (ELI-14). "Yes. Actually I suppose that only Teacher Training Center provides this opportunity for pre-service teachers. Nevertheless, the subjects covered in these courses are only a set of theories and methods but not advanced ones" (CI-3).*

Question 6: Do you ever consider psychometric properties of your test?

The majority of ELIs argued that it is generally unlikely that all psychometrically desirable properties could be optimized simultaneously for all tests. They proposed considering the importance of the tests, decisions to be made and type of the tests. By contrast, CIs do not seem to be in a position to critically evaluate their tests; that is, establish reliability and validity, or do statistical analysis to gauge the quality of the questions, even though they seem to be vaguely familiar with item facility and item discrimination. For example, ELI-10 stated that "*These issues including item facility, item discrimination, choice distribution, reliability, validity,*

[DOI: 10.29252/ijal.19.2.115]

and even practicality should be dealt with, but every single test should not possess all the characteristics." CI-10 noted that

> *I have a passing acquaintance with these properties. For example, three types of questions constitute the students' final exam: about 30% of the questions are easy, about 30% are medium, and the remaining are somehow difficult. These types of tests discriminate between upper-level, mid-level, and lower-level students.*

Question 7: To what extent do you think statistics may come in handy in testing?

The results show statistics serve a useful purpose and help ELIs to compare groups of students, evaluate the amount of learning, and measure students' performance. Similarly, CIs think that statistics is beneficial, although they mostly know basic information about statistics such as descriptive statistics. Two examples are provided to confirm the findings. *"Of course, it would be very useful. Statistical methods allow us to collect information about students and evaluate them in a better way, that is, in a sense without bias" (ELI-9). "I believe that it might be of use to some extent, but in fact my familiarity with it is pretty limited … maybe just central tendency" (CI-11).*

Question 8: Do you ever try changing, or modifying your teaching methodology according the feedback you receive on test results?

The results indicated a perceived alignment of testing and teaching. Both ELIs and CIs' attitudes highlighted the fact that assessment is tightly interconnected with instruction and best suited to individual teachers' needs to take steps in adjusting their teaching methodology. ELI-14 argued that

> *Naturally it happens. Assessment provides a great deal of information for teachers. It tells what areas the students have or have not learned and*

[DOI: 10.29252/ijal.19.2.115]

inform teachers what needs to be reinforced and perhaps what strategies did or did not work.

CI-14 opined that *"Yes. Notwithstanding teachers' desperate attempts, sometimes all students get poor grades. This may be the direct result of teaching. Therefore, the content will be presented refreshingly different from previous time".*

Question 9: How do you prepare your students for tests? Do you teach them strategies, do you ask them to cover content or objectives, or do you use other methods?

Analysis of the responses revealed that although ELIs and CIs did not seem to prepare students for tests, they provide them with two important and necessary pieces of information regarding what is covered in the tests (test content) and how it is administered (test format). ELI-8, for example, stated that *"I inform the students the way I will evaluate them on the content of their textbooks during the course. For example, I explain whether the test is going to be subjective, objective, or a combination of both."* CI-7 also commented that *"Students are personally interested to know about what is covered in the test. Therefore, we provide them with this range of information as they expected us".*

Question 10: Do you find modern technology, including computers, useful in administering and scoring test?

Incorporating the technology types, especially computers, into testing provides versatile assessment options for ELIs at different stages of testing process from the construction of the test to test scoring and analysis of results. Although CIs acknowledged that technology types may be helpful, they just used them to type their questions. The following quotations from instructors illustrate these points.

[DOI: 10.29252/ijal.19.2.115]

> *Yes, they are potentially useful. Benefiting from equipment such as computers, for example, tests can be taken independently of time and place, test questions can be displayed in different orders in the electronic versions of the tests, and large item banks can be created. (ELI-7)*

> *The usefulness of computers is proved in teaching due to the audiovisual materials provided by them. It may be useful in testing too. But, up to now, we only have benefited from the computers for typing the questions. (CI-9).*

Question 11: Do you ever have your colleagues review and comment on the test you make?

Almost all ELIs and CIs were unanimous in not seeking their colleagues' opinions on the tests they develop because of time constraints, colleagues' tight work schedules, and their colleagues' unwillingness. For example, ELI-4 stated that *"Actually seldom. Because we do not have enough time or they do not show willingness, but I feel like doing that."* In like manner, CI-14 noted that *"Yes, but not often, because they are tied up almost all days, so they cannot do anything else."*

Question 12: Do you ever consider ethical issues in testing?

What can be inferred from the analysis of responses is that, for overwhelming majority of both ELIs and CIs, ethical issues are equal to fairness; as a result, tests should be fair not to discriminate against certain test takers. ELI-14 commented that *"To me it is very important. As an integral part of testing, everyone follows some codes of ethicality. Students with the same ability levels do not obtain remarkably different scores"*. Similarly, one content instructor remarked that *"Yes, it is important to judge each person on his/her merits. For me the only criterion of students' evaluation is their performance not anything else"* (CI-11).

[DOI: 10.29252/ijal.19.2.115]

Question 13: What problems do you think teachers may have in testing?

The analysis of responses revealed that, from ELIs' point of view, the fundamental problem lies in the test construction process as teachers do not assume responsibility for it and some of them deliberately ignore some properties under certain circumstances. Agreeing with ELIs, CIs went a step further and added that instructors have trouble in assigning students fair ratings. ELI-6 contended that *"I suppose construction of the test itself is the major source of problem. Some of the teachers do not take seriously some issues regarding designing affective tests, and as a consequence they will be affected."* CI-3 remarked that "*Providing a balanced and, more importantly, unbiased scoring system may be one of the problems that teachers should deal with".*

Question 14: What suggestions do you make to have teachers be updated?

Almost all ELIs and CIs suggested that instructors improve their testing by attending conferences and workshops or reading related materials especially published testing papers via surfing the Internet. Both groups also recommended considering pressure on teachers to improve and enhance their knowledge and skills in assessment.

The following two quotes illustrate these suggestions: "To improve their knowledge, not only can teachers start reading relevant books, but they can also participate in workshops" (CI-7). "It is better that teachers be bound to update their testing knowledge by authorities and organizations that are in charge of education" (ELI-13).

6. Conclusion

The results of exploratory factor analysis showed a three-factor solution for AL. This first finding echoes the hypothesis Brindley (2001), Davies (2008), DeLuca and Klinger (2010), and Fulcher (2012) suggested, theorising that AL can be thought of a three-dimensional construct.

[DOI: 10.29252/ijal.19.2.115]

The results of this study partly lend support to Davies's (2008) theoretical model of AL. Davies proposed three components constituting AL: knowledge, principles, and skills. In Davies's model, the first component, 'knowledge', includes relevant background information about different theories and models of assessment and the second component, 'principles', contains underlying concepts of testing, that is, validity, reliability, ethics, fairness, and impact. In common with Davies's first two components, 'theoretical dimension of testing' as the first factor of our model is comprised of different theories and models formulated in testing as well as test characteristics, proper use of tests for decision making, and their fairness and impact. In other words, the first factor of our model represents two separate components of Davies's model.

On the other hand, the subcomponents that constituted 'skills' are broken down into two discrete factors in our model, namely 'test construction and analysis' and 'statistical knowledge'. Davies identified testing expertise including item writing, statistics, test analysis, and software programs for test delivery, analysis, and reporting as the constituent elements of his third component—skills; in our model, 'test construction and analysis' and 'statistical knowledge' are made up of similar subcomponents.

The results of exploratory factor analysis showed a three-factor solution for ELIs; however, they yielded a six-solution factor for CIs. Two factors may have contributed to the differences of AL between ELIs and CIs in this part. The amount of training received by ELIs and CIs in assessment can account for these differences. While lack, or complete absence of any kind of training in assessment, prevailed among CIs, all ELIs reported having received various degrees of assessment training—either in their BA, MA, or PhD programs while they were students or in the workshops they attended while they were teaching. Apart from degree of training in assessment, discipline can be a second possible reason why such differences exist between ELIs and CIs. All the ELIs majored in English Language. By contrast, CIs majored in a wide range of subject areas.

Results of an independent samples t-test and 50 Mann-Whitney U tests showed statistically significant differences between ELIs and CIs. We did not find any study to compare AL between ELIs and CIs. Therefore, we used the results of

[DOI: 10.29252/ijal.19.2.115]

qualitative phase, as presented and explained in the Results section, to explain the reasons for the differences.

The first reason relates to the way ELIs and CIs assessed their students. CIs cleave to the methods they were traditionally assessed by their instructors in their college courses. Therefore, they did not deem it necessary to keep up with the latest developments in this area. Although ELIs also evaluated their students' performance using some of those methods they were assessed by their instructors, they used more modern testing techniques and were interested in catching up with the most recent testing techniques.

The second reason deals with the purpose of testing. Although all ELIs' AL was primarily on assessment for learning in order to help students to be autonomous, to master learning objectives, and to evaluate themselves, most CIs used assessment primarily to be able to grade students, that is, assessment of learning. This highlights the claim that instructors' assessment practices are influenced by their beliefs on assessment (Quilter & Gallini, 2000; Tierney, 2006).

The third reason concerns training in assessment. Almost all of the ELIs received some training, although the training came through university courses at BA, MA, and PhD programs, but necessarily not in pre or in-service programs. However, the majority of the CIs did not receive any formal training. For example, when asked to provide details of their experience with participation in assessment workshops, the majority of the CIs stated that they seldom attended these workshops or have no routines they could fall back on in order to update their knowledge on testing issues. However, a large number of ELIs tended to attend testing workshops or regularly updated themselves because improvement, they commented, in this area was needed.

The fourth reason relates to establishing psychometric properties of tests. These issues were not addressed by CIs as they had a passing acquaintance only with item difficulty. By contrast, ELIs were somewhat knowledgeable about these properties, although they reported that depending on the importance of the tests, decisions to be made, and type of the tests, psychometric properties were different.

The findings show that AL is not a unitary concept (Smith, Worsfold, Davies, Fisher, & McPhail, 2011), but it is a multifaceted construct consisting of three

[DOI: 10.29252/ijal.19.2.115]

interrelated factors. However, little is still known about what exactly constitutes AL and whether AL may be affected by factors including teaching experience, the amount of training instructors receives, and the setting in which instructors teach. These issues warrant further research to shed light on the complexity of AL. As Taylor (2013) asserted, further empirical studies on AL "are urgently needed not just to inform and underpin existing policy and practice but also to inspire and shape new and innovative initiatives for disseminating core knowledge and expertise in assessment" (p. 405).

The second conclusion drawn from the findings is that the AL is not equally shared by ELIs and CIs. Both ELIs and CIs can, to varying degrees, be assessment literate. Although "teachers spend as much as a quarter to a third of their professional time involved in assessment-related work" (Stiggins, 2014, p. 68), very few CIs receive the essential training needed to do it well, and other CIs may be quite bereft of any training in educational assessment. Koh (2011), Popham (2006, 2009, 2011), and Vogt and Tsagari (2014) advocate designing continuous educational assessment training courses which address teachers' needs of assessment knowledge, enabling them to acquire what they need to know for classroom practice.

The third conclusion relates to the distinction between assessment for learning and assessment of learning. If we want to empower teachers, they should be taught that while assessment is the process of gathering information to inform instructional decisions (assessment of learning), it serves as an instructional tool to help students to learn more (assessment for learning) (Carless, 2015; Sainsbury & Walker, 2008; Scarino, 2013; Stiggins, 2014; Taras, 2002). "All assessment can be oriented for learning" (Deeley & Bovill, 2015). This gap will be bridged by changing teachers' old beliefs by providing them with opportunities to engage in assessment for learning. In this context, when we seek to improve teachers' AL in order to enhance students' learning, educational assessment training may be one of the keys to success.

Building on the findings, we, therefore, argue that professionalism and autonomy among teachers require that teachers be assessment literate so that they will be able to use a wide variety of assessment methods to assess students'

[DOI: 10.29252/ijal.19.2.115]

performance more effectively. Teachers need to be familiar with, skilful at, and knowledgeable about the assessment methods. Our findings suggest that, to be autonomous, teachers need to have access to a wide range of assessment techniques, know how to construct the methods and analyse the assessment results, and possess statistical knowledge to correctly interpret the results.

One way toward such professionalism, we believe, is instructors' consciousness need to be raised regarding how important a role AL can play in evaluating their students' performance. This involves launching workshops, producing more introductory, user-friendly assessment textbooks, and establishing websites (Fulcher's Language Testing resources: http://languagetesting.info/ is an excellent example) so that they can be trained regarding assessment literacy issues because AL enriches instructors' understanding of their current state in assessment and sheds light on their strengths and weaknesses (Vogt & Tsagari, 2014). A second way to contribute to professionalism relates to policy makers in charge. Qualitative findings showed that CIs did not feel they needed to keep up with the latest developments in testing. As a result, Ministry of Education, or other responsible bodies for teacher education, may consider introducing continuous assessment programs. "A cascade from the theoretical to the practical" (Harding & Kremmel, 2016, p. 423) may be a third way to help teachers toward professionalism. This involves translating different components of LA into syllabuses (e.g. theoretical dimension of testing, test construction and analysis, statistical knowledge, as the results of this study showed), effectively teaching teachers these syllabuses, and critically evaluating the outcomes for possible improvement.

Although we empirically investigated AL of two groups of instructors and proposed a three-layered AL competence, we did not examine what components of AL should be taught and prioritised. In the future, researchers may consider developing more creative research designs and methodologies to investigate core components of AL for teachers and how the components should be best taught and prioritised to better empower them in their teaching endeavours. We did not examine mode of delivery for AL instruction in the present study. As Harding and Kremmel (2016) also remind us, "little research effort has gone into the evaluation and comparison of the effectiveness and accessibility of [modes of delivery]" (p.

[DOI: 10.29252/ijal.19.2.115]

425). In future studies, researchers may also consider examining the efficacy of various forms of promoting AL between language instructors and content instructors in L1 and L2 settings to arrive at more solid conclusions about which mode of delivery best works in diverse educational settings for diverse populations.

7. References

Ainsworth, L., & Viegut, D. (2006). *Common formative assessment: How to connect standards-based instruction and assessment*. UK: Sage Publication Ltd.

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

Boyles, P. (2005). Assessment literacy. In M. Rosenbusch (Ed.), *National assessment summit papers* (pp. 11-15). Ames, IA: Iowa State University.

Braney, B. T. (2010). *An examination of fourth grade teachers' assessment literacy and its relationship to students reading achievement* (Doctoral Dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3434570)

Brindley, G. (2001). Language assessment and professional development. In C. Elder, A. Brown, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honor of Alan Davies* (pp. 126–136). Cambridge: Cambridge University Press.

Carless, D. (2015). *Excellence in university assessment*. London: Routledge.

Charmaz, K. (2014). *Constructing grounded theory* (2nd ed.). London: Sage Publications.

Cresswell, J. D., & Plano Klark, V. L. (2011). *Designing and conducting mixed method research* (2nd ed.). California: Sage Publications.

[DOI: 10.29252/ijal.19.2.115]

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing*, *25*(3), 327–348.

Deeley, S. J., & Bovill, C. (2015). Staff student partnership in assessment: Enhancing assessment literacy through democratic practices. *Assessment & Evaluation in Higher Education*.

DeLuca, C., & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 419–438.

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

Flynn, C. P., & Kunkel, S. R. (1987). Deprivation, compensation, and conceptions of an afterlife. *Sociological Analysis*, *48*(1), 58-72.

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

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. New York: Routledge.

Harding, L., & Kremmel, B. (2016). Teacher assessment literacy and professional development. In Tsagari, D. and Banerjee, J. (eds) *Handbook of Second Language Assessment* (pp. 413-429). Berlin: Walter de Gruyter.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Hutcheson, G. D., & Sofroniou, N. (1999). *The multivariate social scientist*. London: Sage Publications Ltd.

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

Kaiser, F. H. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31-36.

Koh, K. H. (2011). Improving teachers' assessment literacy through professional development. *Teaching Education*, *22*(3), 255-276.

Krueger, R. A., & Casey, M. A. (2000). *Focus groups: A practical guide for applied research* (3rd ed.). Thousand Oaks: Sage Publications.

Kumaravadivelu, B. (2006). *Understanding language teaching: From method to post-method*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

[DOI: 10.29252/ijal.19.2.115]

Mackey, A., & Gass, S. (2005). *Second language research: Methodology and design*. Mahwah: Lawrence Erlbaum.

Medland, E. (2015). Examining the assessment literacy of external examiners. *London Review of Education, 13*(3), 21-33.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed.) (pp. 13-103). Washington, DC: American council on Education.

Newfields, T. (2006). Teacher development and assessment literacy. Retrieved from http://jalt.org/pansig/2006/PDF/Newfields.pdf

Ogan-Bekiroglu, F., & Suzuk, E. (2014). Pre-service teachers' assessment literacy and its implementation into practice. *The Curriculum Journal, 25*(3), 344-341.

Pill, J., & Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. *Language Testing, 30*(8), 381-402.

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice, 12*(4), 10-12.

Popham, W. J. (2004). Why assessment illiteracy is professional suicide. *Educational Leadership, 62*(1), 82–83.

Popham, W. J. (2006). Needed: A dose of assessment literacy. *Educational Leadership, 63*(3), 84-85.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice, 48*(1), 4-11.

Popham, W. J. (2011). Assessment literacy overlooked: A teacher education's confession. *The Teacher Educator, 46*(4), 265-273.

Quilter, S. M., & Gallini, J. K. (2000). Teachers' assessment literacy and attitudes. *The Teacher Educator, 36*(2), 115-131.

Saif, A. A. (2009). *Educational measurement, assessment, and evaluation* (5th ed.). Tehran: Dowran Publication.

Sainsbury, E. J., & Walker, R. A. (2008). Assessment as a vehicle for learning: Extending collaboration into testing. *Assessment & Evaluation in Higher Education, 33*(2), 103–117.

[DOI: 10.29252/ijal.19.2.115]

Saldana, T. (2012). *The coding manual for qualitative researchers* (2nd ed.). London: Sage Publications.

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. (2008). Introduction to volume 7: Language testing and assessment. In N. Hornberger (Ed.), *Encyclopedia of language and education* (2nd ed.) (pp. xiii-xxii). New York: Springer Science and Business Media, Inc.

Smith, C. D., Worsfold, K., Davies, L., Fisher, R., & McPhail, R. (2011). Assessment literacy and student learning: The case for explicitly developing students' assessment literacy. *Assessment & Evaluation in Higher Education, 38*(1), 44-60.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan, 72*(3), 534–539.

Stiggins, R. J. (2014). Improve assessment literacy outside of schools too. *Phi Delta Kappan, 96*(2), 67-72.

Taras, M. (2002). Using assessment for learning and learning from assessment. *Assessment & Evaluation in Higher Education, 27*(6), 501–510.

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

Tierney, R. D. (2006). Changing practices: Influences on classroom assessment. *Assessment in Education, 13*(3), 239-264.

Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly, 11*(4), 374-402.

Volante, L., & Fazio, X. (2007). Exploring teacher candidates' assessment literacy: Implications for teacher education reform and professional development. *Canadian Journal of Education, 30*(3), 749-770.

Webb, N. L. (2002). Assessment literacy in a standard-based urban education setting. Paper presented at the annual meeting of the American Education Research Association, Neworleans, LA.

[DOI: 10.29252/ijal.19.2.115]

William, D. (2015). Foreword: Assessment literacy. *London Review of Education, 13*(3), 3-4.

Notes on Contributors:

Rajab Esfandiari is an assistant professor in English Language Teaching at Imam Khomeini International University in Qazvin, Iran. His areas of interest include teaching and assessing L2 writing, multifaceted Rasch measurement, L2 classroom assessment, and EAP teaching and testing.

Razieh Nouri obtained her MA in TEFL from Imam Khomeini International University. She is currently teaching English conversation classes at language institutes in Qazvin. Her areas of interest include language assessment and scale construction.

[DOI: 10.29252/ijal.19.2.115]

Appendix. Questionnaire for Assessment Literacy

The following items measure different aspects of assessment literacy.
Please, read them very carefully and indicate your response as follows.

Items		1.Not at all	2.Small degree	3.Moderate degree	4.High degree	5.Very high degree
1.	Accountability (obligation of instructors to accept responsibility for students' performance)	O	O	O	O	O
2.	Validity (predictive, concurrent, content, construct, face, response)	O	O	O	O	O
3.	Reliability (test-retest, parallel forms, split-haves, Kuder-Richardson formulae, Cronbach's alpha, scorer reliability)	O	O	O	O	O
4.	Authenticity (situationally authentic tests, interactionally authentic tests)	O	O	O	O	O
5.	Practicality (ease of administration, ease of scoring, ease of interpretation and application, availability of resources)	O	O	O	O	O
6.	Interactiveness (interaction between test takers' characteristics and test tasks)	O	O	O	O	O
7.	Fairness and ethics in assessment	O	O	O	O	O
8.	Proper use of tests (correct interpretation of test results)	O	O	O	O	O
9.	Washback and impact (the effect of tests on teaching/learning, society, and educational systems)	O	O	O	O	O
10.	Consequences of tests (social, educational, political,	O	O	O	O	O

[DOI: 10.29252/ijal.19.2.115]

	etc.)					
11.	Test bias (due to reasons such as cultural background, ethicality, sex, native language, background knowledge)	O	O	O	O	O
12.	Theories of testing (traditional testing, discrete-point testing, integrative testing, communicative testing)	O	O	O	O	O
13.	Testing models and frameworks (skills and components model, communicative ability)	O	O	O	O	O
14.	Different test types(objective versus subjective, essay type versus multiple choice)	O	O	O	O	O
15.	Functions of tests (achievement, proficiency, aptitude, selection, placement, diagnosis)	O	O	O	O	O
16.	Basic concepts in testing (tests, measurement, evaluation, test use, test type, test format)	O	O	O	O	O
17.	History of testing (pre-scientific, psychometric-structuralist, sociolinguistic-pragmatic)	O	O	O	O	O
18.	Uses of tests in educational programs	O	O	O	O	O
19.	Examination of different models of learning/learning in testing	O	O	O	O	O
20.	Testing in relation to curriculum	O	O	O	O	O
21.	Alternative assessment	O	O	O	O	O
22.	Test critique (critical evaluation of tests)	O	O	O	O	O
23.	Research methods in setting up experiments in testing (quantitative, qualitative, and mixed-methods approaches)	O	O	O	O	O
24.	Conducting Item analysis and test analysis	O	O	O	O	O
25.	Using computer software programs in testing (test construction, test analysis, and test scoring)	O	O	O	O	O
26.	Compiling table of test specifications	O	O	O	O	O
27.	Using different types of interpretation (norm-referenced and criterion-referenced interpretation)	O	O	O	O	O

[DOI: 10.29252/ijal.19.2.115]

28. Realizing limitations of test result interpretation (indirectness, incompleteness, imprecision, subjectivity, relativeness)	O	O	O	O	O
29. Developing and using selected-response assessments (True-false, matching, multiple choice)	O	O	O	O	O
30. Developing and using constructed-response assessments (Fill in the blank, short answer and performance assessments)	O	O	O	O	O
31. Developing and using personal response assessments (checklists, journals, videotapes, audiotapes, self-assessment, peer assessment, teacher observation, portfolios, conferences, diaries)	O	O	O	O	O
32. Preparing students for tests	O	O	O	O	O
33. Utilizing test taking strategies	O	O	O	O	O
34. Recognizing test distinctions (formal versus informal tests, traditional versus alternative tests, low-stakes versus high-stakes tests, teacher-made versus standardized tests)	O	O	O	O	O
35. Providing test security	O	O	O	O	O
36. Determining test function and form	O	O	O	O	O
37. Doing planning (determining/specifying the content of tests)	O	O	O	O	O
38. Preparing items	O	O	O	O	O
39. Reviewing items (modification and improvement of the quality)	O	O	O	O	O
40. Doing pre-test (item facility, item discrimination, choice distribution)	O	O	O	O	O
41. Validating the test	O	O	O	O	O
42. Developing a detailed scoring system for rater-mediated assessments (holistic, primary trait scoring, multiple traits scoring)	O	O	O	O	O
43. Using scales of measurement (nominal, ordinal,	O	O	O	O	O

[DOI: 10.29252/ijal.19.2.115]

	interval, ratio scale)					
44.	Scoring and administration of paper and pencil, or oral tests	O	O	O	O	O
45.	Administering and scoring computer-adapted testing and Internet-based testing	O	O	O	O	O
46.	Using and interpreting descriptive statistics, including measurement of central tendency (mode, mean, median)	O	O	O	O	O
47.	Using and interpreting descriptive statistics, including measurement of variability (range, variance, standard deviation)	O	O	O	O	O
48.	Using and interpreting inferential statistics (parametric versus nonparametric) (t-test, ANOVA, MANOVA, Chi-square, Correlation, Regression, Factor analysis, Kruskall-Wallace)	O	O	O	O	O
49.	Using and interpreting advanced statistics (Classical True Score theory, Generalizability theory, Item Response theory, Structural Equation Modelling, Path analysis)	O	O	O	O	O
50.	Using and interpreting more modern statistical tests (Multilevel modelling, Autoregressive SEM models, Latent growth curve modelling, Time series approaches, Event history analysis)	O	O	O	O	O

[DOI: 10.29252/ijal.19.2.115]