Volume 10, Issue 2
Summer and Autumn, 2022
pp. 9-34

Journal of Applied Linguistics and Applied Literature: Dynamics and Advances

Towards the Development of an Assessment Literacy Questionnaire: The Case of Iranian EFL Teachers

Behnaz Rastegar[1], Abbas Ali Zarei[2*], and Rajab Esfandiari[3]

PhD Candidate, Imam Khomeini International University, Qazvin, Iran
Email: rastegar.beh@gmail.com

[2]Corresponding author: Associate Professor, Department of English, Faculty of Humanities, Imam Khomeini International University, Qazvin, Iran,
ORCID: 0000-0002-6062-6633
Email: a.zarei@hum.ikiu.ac.ir

[3]Associate Professor, Department of English, Faculty of Humanities, Imam Khomeini International University, Qazvin, Iran, ORCID: 0000-0002-2305-762X
Email: esfandiari@hum.ikiu.ac.ir

Abstract

Substantial research has been done on assessment literacy (AL), and several questionnaires have been developed to measure AL. However, little (if any) research has attempted to provide a comprehensive assessment literacy questionnaire. To fill this gap, the present study attempted to develop an assessment literacy questionnaire which encompasses not only the areas identified by previous research, but also those not identified by those studies. Moreover, attempt was made to identify the components that were better predictors of Iranian EFL teachers' assessment literacy. To this end, first previous AL questionnaires were explored and their main items were identified. Then, researchers-made items were added. Meanwhile, interviews were conducted with experts, who suggested some additional items. Then, 386 Iranian teachers of English were selected through convenience sampling on the basis of availability to fill in the first draft of the questionnaires that assessed different aspects of assessment literacy in order to validate it. Finally, a Principal Component Analysis (PCA) was conducted, and a questionnaire with 35 items which evaluated nine components of assessment literacy was developed. In addition, the validated, final version of the AL questionnaire was distributed among 146 EFL teachers to identify the better predictor components of AL among Iranian EFL teachers. Multiple regression analysis revealed that "administering, rating, and interpreting test" was the best predictor of teachers' AL in comparison to other components. The theoretical as well as practical implications of the findings are also discussed.

Keywords: assessment literacy, language assessment, questionnaire construction, validation

ARTICLE INFO

Research Article
Received: Sunday, January 2, 2022
Accepted: Wednesday, August 24. 2022
Published: Thursday, 1 December, 2022
Available Online: Wednesday, August. 4, 2022
DOI: http://dx.doi.org/10.22049/JALDA.2022.27579.1378

Online ISSN: 2821-0204; Print ISSN: 28208986

Introduction

The discussion of the role teachers can have in assessing their students has given popularity to the concept of "assessment literacy". Assessment literacy (AL), usually defined as teachers' knowledge of how to assess, what to assess, and how to interpret assessment results (Scarino, 2013), is among the most significant aspects of teachers' development. Research suggests that teachers of English, especially those in the EFL context of Iran have major problems with AL. Lan and Fan (2019) hold that most EFL teachers have only insufficient knowledge of assessment literacy. Pastore and Andrade (2019) believe that this is a result of incomprehensive models in developing assessment literacy. Moreover, a number of studies have discussed the shortcomings of AL in the context of Iran (e.g., Ashraf & Zolfaghari, 2018). Farhady and Tavassoli (2018), for example, noted that Iranian EFL teachers' AL has many aspects that are not all taken into account. Ashraf and Zolfaghari (2018) noted that to measure AL, both theoretical and practical knowledge should be assessed. To do so, a comprehensive instrument is needed. Without such a comprehensive instrument, not all aspects of AL could be targeted and improved. These shortcomings of AL suggest that there is a need for an instrument that could measure teachers' AL in all aspects.

Most of the commonly used assessment literacy questionnaires have referred to only some aspects of AL, but neglected some significant variables. Determining the components of language AL is a real and complex challenge (Inbar-Lourie, 2008; Rea-Dickins, 2008). In this study, the researchers tried to develop a comprehensive questionnaire that included many of these variables. Therefore, the purpose of this study was to determine the components of language teachers' assessment literacy that have already been measured using different questionnaires. Another important goal was to develop a new comprehensive questionnaire and to validate it. A final objective was to see which of the components of the newly developed AL questionnaire can better predict Iranian EFL teachers' AL. More specifically, this research addressed the following research questions:

1. What are the main items of AL in the currently used questionnaires?
2. What are the potential items to be used in a newly developed AL questionnaire?
3. What are the components of the newly developed AL questionnaire?
4. Which of the components of the newly developed questionnaire is a better predictor of Iranian EFL teachers' assessment literacy?

Review of the Related Literature

The concept of AL has been around for decades. However, only in recent decades have scholars come to the conclusion that it should be standardized. Indeed, improvements in the field of assessment and the expansion of assessment types such as formative, summative, alternative, formal, informal, and dynamic, have made it necessary for language teachers to know exactly the purpose of each type of assessment, the context where it fits, and the approach to implementing it (Malone, 2013).

Assessment literacy has traditionally been assessed through questionnaires, surveys and interviews (mostly semi-structured). A number of instruments have been used to measure assessment literacy; e.g., teacher assessment literacy questionnaire (Plake et al., 1993), classroom assessment literacy inventory (Mertler, 2003), language assessment knowledge needs questionnaire (Fulcher, 2012), successful Iranian EFL Teacher Questionnaire (SIETQ) (Moafian & Pishghadam, 2009), and assessment literacy Likert scale questionnaire (Esfandiari & Nouri, 2016).

Different aspects of AL have been investigated. Regarding teachers' views about the purpose and nature of assessment and the effect of these views on practices and outcomes significant studies have been conducted (Fulmer et al., 2015). It has been reported that not only possessing AL skills is requisite for lecturers, but also language assessment knowledge has spread itself as a necessity for alternative stakeholders inside the educational testing culture like policy makers, examination boards, and parents (Taylor, 2009). Recent studies have shown inadequate proof that researchers provide thought to the views of main stakeholders in the context of language AL.

In one study, Farhady and Tavassoli (2018) studied assessment related issues from the perspective of language teachers. To operationalize the concept of AL, 14 concepts and components of AL were taken from the literature including Roever and McNamara (2006), Popham (2009), and Shohamy (2001), and interview questions were developed. The major recognized concepts of AL were reliability, validity, assessment bias, construction of selection and construction test, scoring constructed-response test item, developing alternative assessment, formative assessment, interpretation of students' performance on standardized tests, assessing students with disabilities, high-stakes test preparation, the effect of test on teaching quality, social consequences of test results, giving feedback after assessment, and Students' involvement in test construction.

Regarding the effect of training and teacher education programs on teachers' language AL, some previous studies have reported insufficiency of training (Jeong, 2013; Lam, 2015). However, other studies have shown that training in language LA can improve the language AL of lecturers (Volante & Fazio, 2007).

As for the effect of AL on testing and testing resources, since one of the most integral parts of every educational system is testing, teachers' AL background is crucial for conducting assessment. In this regard, O'Loughlin (2013) sought to investigate the AL needs of a group of university members of staff with regard to the IELTS test in the context of Australian higher education and the possible approaches that could potentially be adopted to satisfy such needs. These needs were related to issues such as test purpose and content, test scores meaning, cut-off levels appropriateness, test validity and reliability, and predictive power. The results suggested that the participants of the study had to be provided with information about IELTS to advise prospective students about the requirements for gaining admission into English language courses and to make decisions about admissions.

A study by Razavipour et al. (2011) aimed to highlight the effects of teachers' AL on language assessment. Results of teachers' self-assessments of their own

general readiness to assess the performance of their students showed that Iranian EFL teachers acknowledged that their assessment background was rather weak. Similarly, Farhady and Tavassoli (2018) examined EFL teachers' AL and found that most of the participants believed they needed more knowledge of assessment. Khanjani et al. (2017) go further and report that even teacher training programs cannot successfully enhance EFL teachers' AL.

As to the effect of AL on teaching practices, Ashraf and Zolfaghari (2018) found a direct correlation between EFL teachers' AL and their reflective teaching. In another study, Esfandiari and Nouri (2016) explored assessment literacy and its implications for teachers' professional development. The findings showed that AL is a multifaceted construct consisting of three interrelated factors, rather than a unitary concept. Mellati and Khademi (2018) found that that teachers' AL provides them with essential information about the efficiency of their pedagogy and has a significant impact on learners' achievement.

Despite the above-mentioned benefits of AL for language teachers, Rezvani Kalajahi and Abdullah (2016) conclude that there is a theory-practice gap within the assessment context. Similarly, Janatifar and Marandi (2018) believe that language assessment literacy (LAL) is an understudied construct in the EFL context of Iran. In other words, despite the studies on the different aspects of AL, there still appear to be many unanswered questions surrounding this issue. For one thing, the data collection instruments that have been used in the previous studies do not seem comprehensive enough to capture the various aspects of the multi-faceted concept of AL. Another issue is that different studies on AL have reported different underlying components for the construct of AL. It is, therefore, not quite clear, to date, what constitutes AL. Another understudied area is which of the components of AL is more strongly associated with and, therefore, a better predictor of teachers' LAL. This study was an attempt to bridge part of the mentioned gap in the relevant literature.

Methodology

This study employed an exploratory sequential mixed methods design. First, the qualitative phase of data collection and content analysis of the data (already existing questionnaires on AL) was carried out. This was followed by a quantitative phase, which involved the validation of the assessment literacy questionnaire as well as extracting the main components of assessment literacy and checking the predictive power of each component of this questionnaire over EFL teachers AL.

Participants

To answer the third question of the study, 386 Iranian EFL teachers (221 males and 165 females) were selected through convenience sampling based on availability. Since the purpose of the study was to assess the AL of Iranian EFL teacher, the only criterion that was used was experience in teaching and testing English. The AL questionnaire was sent online to as many teachers of English as could be found. Whoever returned the filled-out questionnaire was taken as a participant. The participants varied in age from 21 to 65 years old with the mean age

of 32.5. Their teaching experience ranged from three to 35. They lived in various cities of Iran. 73 of them had B.A. in TEFL; 217 had M.A.; and 96 were Ph.D. holders or Ph.D. students.

To address the fourth research question, 146 Iranian EFL teachers were selected through convenience sampling. Their age ranged from 21 to 65 years with the mean age of 31.5. Their teaching experience ranged from four to 40 years. 22 of them had B.A. in TEFL, 75 had M.A., and 49 were Ph.D. students or Ph.D. holders.

Instruments

The instruments for the qualitative section of the study were the following AL questionnaires: teacher AL questionnaire (Plake et al., 1993), language assessment knowledge needs questionnaire (Farhady & Tavassoli, 2018), classroom AL inventory (Mertler, 2003), LAL (Fulcher, 2012), questionnaire for AL (Esfandiari & Nouri, 2016), classroom AL inventory (Mertler, 2003), language assessment knowledge scale (Ölmezer-Öztürk & Aydin, 2018), and LAL survey (Janatifar & Marandi, 2018). These questionnaires assessed various aspects of teachers' AL. All these are established questionnaires, and each of them has been used in one or more studies with acceptable indices of reliability.

The instrument for the first part of the quantitative section of the study was the first draft of the newly developed of AL questionnaire. The instrument for the second part of the quantitative section of the study was the finalized and validated version of AL questionnaire. The validity of the new questionnaire was established through expert opinion and the subsequent factor analysis. Its reliability was estimated using Cronbach alpha to be .83.

Procedure

First, the questionnaires listed in the previous section were collected. A thorough search for the related questionnaires was done. To make sure the researchers did not miss any published questionnaires, valid databases such as Wiley, Science Direct, Sage Publications, Springer Nature, SCOPUS, Web of Science, Sage publications and Oxford Publications were screened carefully. Other than these databases, the publishers and journals that publish extensively in the Iranian EFL context were checked. This was done to ensure that the researchers collected all questionnaires on AL. Then, their content was explored and their similarities and differences were determined. Next, the researchers added some missing parts that they believed should be included in such questionnaires. The result was an 80-item questionnaire (54 items from the existing questionnaire and 26 researcher-made items).

Next, five research experts were asked to scrutinize the questionnaire and provide us with their comments. They deleted 12 researcher-made items and the questionnaire was reduced to a 68-item questionnaire. Then, 30 items were combined and a 38-item Likert scale questionnaire was produced. This questionnaire was given to 386 participants to fill in. Then, after a factor analysis, 35 items were loaded on 9 factors, and three items were omitted. In the next step, the final version

of the AL questionnaire was administered to 146 participants to identify the best predictor of AL from among the extracted components.

Data Analysis

For the qualitative phase, a thematic analysis was used. The researchers used Hsieh's and Shannon's (2005) model in qualitative content analysis (QCA) to address the first research question. A Principal Components Analysis (PCA) was used to identify the underlying components of the newly developed AL questionnaire. Furthermore, multiple regression analysis was used to examine the relative contribution of each component of the newly developed AL questionnaire to predict teachers' AL.

Results

The First Research Question

The first research question sought to find the main items of AL in existing questionnaires. To do this, the 11 AL questionnaires mentioned earlier were collected. Then, they were compared, and 54 items were extracted. The number of items that were extracted from each source was different due to the fact that some questionnaires were scenario-based and others were field specific. Also, the wording of the selected items was different based on the purpose and length of the questionnaires. Hence, the present researchers attempted to list those items in a simple and unified way (See Table 1).

Table 1

The Items Extracted from the Currently Used AL Questionnaires

1. Doing planning (determining / specifying the content of tests) / (deciding what to test)
2. Compiling table of test specifications (writing test specifications / blueprints)
3. Preparing items
4. Reviewing items (modification and improvement of the quality)
5. Doing pre-test (item facility, item discrimination, choice distribution)
6. Developing and using recognition type assessments (true-false, matching, multiple choice)
7. Developing and using suppletion type assessments (fill in the blank, short answer and performance assessments, short essay)
8. Developing and using personal response assessments (checklists, journals, videotapes, audiotapes, self-assessment, peer assessment)
9. Teacher observation, portfolios, conferences, diaries
10. Validity (predictive, concurrent, content, construct, face, response)

11. Rating performance tests (speaking / writing)
12. Rating receptive tests (listening / reading)
13. Classroom assessment
14. Large-scale testing
15. Standard setting
16. Preparing learners for tests through utilizing test taking strategies
17. Washback and impact (the effect of tests on teaching / learning, society, and educational systems, on the classroom)
18. Test administration
19. Fairness and ethical considerations in testing or assessment
20. Consequences of tests (social, educational, political) / (the uses of tests in society)
21. Proper use of tests (correct interpretation of test results)
22. Alternative assessment
23. Familiarity with authentic test (test content which is related to students with a specific cultural heritage)
24. Have deep cultural awareness which informs test creation, dissemination and evaluation
25. Using and interpreting descriptive statistics, including measurement of central tendency (mode, mean, median)
26. Using and interpreting descriptive statistics, including measurement of variability (range, variance, standard deviation)
27. Using and interpreting inferential statistics (parametric versus nonparametric)
28. Using and interpreting advanced statistics
29. Using and interpreting more modern statistical tests (multilevel modelling)
30. Research methods in setting up experiments in testing (quantitative, qualitative, and mixed-methods approaches)
31. Using computer software programs in testing (test construction, test analysis, and test scoring)
32. Using different types of interpretation (norm-referenced and criterion-referenced interpretation)
33. Realizing limitations of test result interpretation (indirectness, incompleteness, imprecision, subjectivity, relativeness)

34. Recognizing test distinctions (formal versus informal tests, traditional versus alternative tests, low-stakes versus high-stakes tests, teacher-made versus standardized tests)
35. Developing a detailed scoring system for rater mediated assessments (holistic, primary trait scoring, multiple traits scoring)
36. Using scales of measurement (nominal, ordinal, interval, ratio scale)
37. Scoring and administration of paper and pencil, or oral tests
38. Administering and scoring computer-adapted testing and Internet-based testing (TOEFL IBT)
39. Functions of tests (achievement, proficiency, aptitude, selection, placement, diagnosis)
40. Providing test security
41. Test bias (due to reasons such as cultural background, ethicality, sex, native language, background knowledge)
42. Interactiveness (interaction between test takers' characteristics and test tasks)
43. Practicality (ease of administration, ease of scoring, ease of interpretation and application, availability of resources)
44. Authenticity (situationally authentic tests, interactionally authentic tests)
45. Accountability (obligation of instructors to accept responsibility for students' performance)
46. Inform and justify students on how their work was derived
47. Guide students on what steps to take to improve (constructive feedback)
48. Give on time feedback
49. Grade based on a student performance a range of assessments
50. Avoid bias in grading (grade based on one test)
51. Calculate final score according to the relative importance of assessment range (short quiz, mid-term, final)
52. CAT: computer adaptive tests or tailored testing (tests that are adapted to examinee's ability level)
53. Multimodal assessment (use more than one modality to create meaning: visual, aural)

The Second Research Question

The purpose of the second research question was to find potential items to be used in the newly developed AL questionnaire. For this question, those main items

that were extracted from the currently used questionnaires were kept. Then, by considering the cultural aspects of second language tests, the ways of giving feedback, and the more up-to-date computer-based assessment of second language, the researchers added 26 items relevant to the cultural aspects of second language tests and the ways of giving feedback and some aspects of computerized assessment to the selected items. However, after consulting with five experienced experts, 12 of the items that the researchers intended to add were removed due to the overlap with each other, and 14 items were kept (Table 2).

Table 2

Researcher-Made Items

1. Administer dynamic assessment (test, teach, retest)
2. Acquaintance with multicultural assessment approaches
3. Eliminate cultural bias in language testing (test which represent perception and experiences of a specific group of people)
4. Eliminate content bias in language testing (familiarity of specific group of students with some specific vocabulary or interaction patterns included in the test)
5. Eliminate linguistic bias in testing (discrepancy between examiner dialect and that of students)
6. Computer-based testing (equivalent to paper and pencil tests as gold standards)
7. CALT: computer-assisted language testing (employ computer application for eliciting and evaluating test taker's performance)
8. Technology-based assessments (use of podcast, free online chats, mobile phones)
9. WBT: web-based testing (user friendly technology in language testing and administration)
10. Give meaningful feed back
11. Give verbal feedback
12. Give individual feedback
13. Give written feedback
14. Give negative feedback (wrong answer)

The total number of items (previously used items and researcher-made items) reached 68 items. Then, 30 of the items were merged because some items were subsets of other items; therefore, instead of stating them separately, the researchers mentioned them as more general items. Previous studies have also done the same (Farhady & Tavassoli, 2018). Finally, after revising and merging items, there remained 38 items (Table 3). The four new themes identified in this study (administering dynamic assessment; multicultural assessment approaches and

subjects related to cultural, content, and linguistic bias in language testing; or computer-based testing, WBT, CALT and technology-based assessment; and different types of feedback in assessment) are of crucial importance. For instance, dynamic assessment can lead to more accurate ways of assessing students' potential for future development. Influenced by Vygotsky's arguments, Feuerstein et al. (1981, p. 218) state that "what is at stake is not theoretical elegance, but issues that affect the lives and destinies of real people". This shows the importance of dynamic assessment, especially, regarding its effect on people's future lives. As for the second theme, multicultural assessment approaches and subjects related to cultural, content, and linguistic bias in language testing, according to Savignon (2018), multicultural assessment enables students from different communities to learn how to interact effectively with students from other cultural backgrounds. The third theme (computer-based testing, WBT, CALT and technology-based assessment) has a notable role in assessing second language (Zygouris & Tsolaki, 2015). Even some theoreticians have argued that in the future, it may be the main tool dominating second language assessment (All et al., 2016). The fourth one, different types of feedback in assessment, has also been supported by the literature – especially, theoretical research (Bangert-Drowns et al., 1991; McDaniel et al., 2007). Different types of feedback are associated with students' second language learning (Bangert-Drowns et al., 1991).

Table 3

Final Items of the AL questionnaire

1. Dynamic assessment (test, teach, retest)
2. Familiarity with authentic test
3. Multimodal assessment
4. Computer-based (CALT, CAT), technology-based (podcast, free online chats, mobile phones) and web-based testing (user-friendly technology in language testing)
5. Teachers' responsibility to prepare learners for test through utilizing test taking strategies
6. Rating performance (speaking, writing) or receptive tests (listening, reading) through developing a detailed scoring system (holistic, primary trait scoring, multiple trait scoring)
7. Reliability (test-retest, parallel forms, split-halves, Kuder Richardson formulae, Cronbach's alpha, rater reliability)
8. Validity (predictive, concurrent, content, construct, face, response)
9. Using different types of interpretation (norm referenced and criterion referenced interpretation)
10. Administering and scoring and analyzing paper and pencil, computerized or internet-based testing (through paper and pencil or computer software programs)

11. Alternative assessment
12. Practicality (ease of administration, ease of scoring, ease of interpretation, availability of resources)
13. Proper use of test (correct interpretation of test results) and consequences of test (social, educational, political) in society with regard to the limitation of test result interpretation (indirectness, incompleteness, imprecision, subjectivity, relativeness)
14. Authenticity (situationally authentic tests, interactionally authentic tests)
15. Interactiveness (interaction between test takers' characteristic and test tasks)
16. Test administration in standard setting (either classroom assessment or large-scale testing)
17. Teacher accountability (obligation of instructors to accept responsibility for students' performance)
18. Using and interpreting descriptive statistics including measurement of central tendency and variability
19. Acquaintance with multicultural assessment approaches
20. Avoid bias in grading (calculate final score according to relative importance of assessment range such as short quiz, midterm and final)
21. Using and interpreting inferential statistics (parametric vs nonparametric)
22. Using and interpreting advanced statistics (classical true score theory, generalizability theory, item response theory, structural equation modeling, path analysis) and more modern statistical tests (multilevel modeling, Rasch)
23. Doing planning (determining the content of test)
24. Compiling table of test specifications and blue prints
25. Knowing different types of feedback (constructive, on time, meaningful, verbal, written, individual, negative)
26. Fairness and ethical consideration (provide security) in testing or assessment
27. Informing and justifying students on how their work was derived
28. Reviewing items (modification and improvement of the quality)
29. Doing pretest (item facility, item discrimination, choice distribution)
30. Using scale of measurement (nominal, ordinal, interval, ratio scale)
31. Research methods in setting up experiments in testing (quantitative, qualitative, mixed-method approaches)
32. Washback and impact
33. Preparing items

34. Eliminate various types of bias in testing such as: cultural, content, linguistic, ethnic, sex, background knowledge, and native language bias
35. Having deep cultural awareness which informs test creation, dissemination and evaluation
36. Recognizing test function (achievement, proficiency, aptitude, selection, placement, diagnosis)
37. Developing and using recognition type, suppletion type and personal type (check list, journal, audiotapes) assessment
38. Recognizing test distinction (formal versus informal, traditional versus alternative tests, low stake versus high stake tests, teacher made versus standardized tests)

The Third Research Question

The third question aimed at investigating the factorial structure of the newly developed AL questionnaire. The Principal Component Analysis (PCA), as a part of Exploratory Factor Analysis (EFA) was run to extract the factors (i.e., components). Before running EFA, its assumptions were checked.

The first thing that needed to be considered was normality. It is believed that a sample size of 300 is desirable for factor analysis. The sample size of this study (N = 386) was considered suitable for EFA. In the correlation matrix, if few correlations exceed |0.30|, EFA may not be suitable. The correlation matrix was checked and multiple correlations were observed that were above |0.30|, justifying the use of EFA. Due to the lack of space and the large size of the table, only a part of correlations is displayed in Table 4.

Table 4

Correlation Matrix

	i03	i04	i05	i06	i07	i09	i10	i11	i12	i15	i25	i26	i27	i28	i29	i30
1	.488	.40	.26	.26	.32	.12	.42	.41	.29	-.18	-.11	-.03	.03	.01	-.07	.03
3	1.0	.48	.47	.43	.56	.51	.43	.45	.36	.38	-.13	-.06	-.03	.00	-.07	-.06
4		1.0	.38	.24	.27	.24	.33	.35	.29	.31	-.2	-.08	-.04	-.04	-.16	-.12
5			1.0	.37	.51	.42	.32	.42	.56	.58	-.17	-.03	-.06	-.06	-.13	-.06
6				1.0	.66	.35	.39	.53	.26	.28	-.07	-.19	.00	-.02	.00	-.09
7					1.0	.40	.45	.62	.36	.46	-.02	-.08	-.08	-.10	-.00	-.10
9						.45	.28	.33	.10	-.06	-.05	-.11	-.10	.07	-.20	-.11
10						1.0	.49	.25	.36	.38	.00	-.05	-.11	-.01	-.07	-.09
11							1.0	.47	.29	.31	.01	-.03	-.08	.01	-.07	-.02
12								1.0	.56	.58	-.09	-.06	-.06	-.10	-.03	-.05
14									.36	.46	-.06	-.09	-.07	.06	.05	-.01
16									.41	.36	-.13	-.11	-.17	-.11	-.05	-.06
18									.36	.39	-.09	-.11	-.07	-.15	-.09	-.08
19									1.0	.52	-.28	-.36	-.32	-.37	-.30	-.35
22										1.0	.30	.52	.50	.19	.18	.46
23											.35	.53	.48	.23	.34	.43
24											.54	.22	.19	.10	.61	.27
26												1.0	.18	.16	.59	.42
27													1.0	.30	.39	.55
28														1.0	.25	.40
29															1.0	.41

Then, the results of Kaiser-Meyer-Olkin (KMO) and Bartlett's test of sphericity were checked to further ensure the suitability of data for factor analysis. Table 5 shows the results.

Table 5

KMO and Bartlett's Test for AL Questionnaire

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.856
Bartlett's Test of Sphericity	Approx. Chi-Square	5059.307
	df	703
	Sig.	.000

The Bartlett's test of sphericity confirmed that the matrix of correlations deviates significantly from an identity matrix ($\chi^2 = 5059.307$, $p < .001$), suggesting that a common shared variance accounted for the intercorrelations among the items. Moreover, the KMO index was 0.856, higher than .60, further supporting the factorability of data.

To decide how many factors to retain, the researchers relied on three criteria: eigenvalues, the scree plot and parallel analysis. Table 6 shows the eigenvalues of the variables before and after extraction. To save space only eigenvalues above 1 are reported.

Table 6

Eigenvalues after Extraction, and Rotation

Component	Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.222	16.374	16.374	3.398	8.943	8.943
2	5.632	14.821	31.195	3.154	8.300	17.243
3	1.949	5.130	36.325	2.526	6.648	23.891
4	1.854	4.878	41.203	2.447	6.439	30.330
5	1.483	3.904	45.107	2.425	6.381	36.711
6	1.353	3.561	48.668	2.228	5.862	42.573
7	1.274	3.353	52.021	2.186	5.754	48.327
8	1.188	3.127	55.149	1.860	4.895	53.222
9	1.153	3.035	58.184	1.527	4.019	57.241
10	1.021	2.687	60.871	1.380	3.630	60.871

Extraction Method: Principal Component Analysis.

As is clear in Table 6, eigenvalue for ten factors is above 1, and only these factors should be kept for further analysis. However, a prerequisite assumption for the reliability of this criterion is that the communality of all variables should be greater than 0.7. Therefore, it was necessary to control these values which are displayed in Table 7.

Table 7

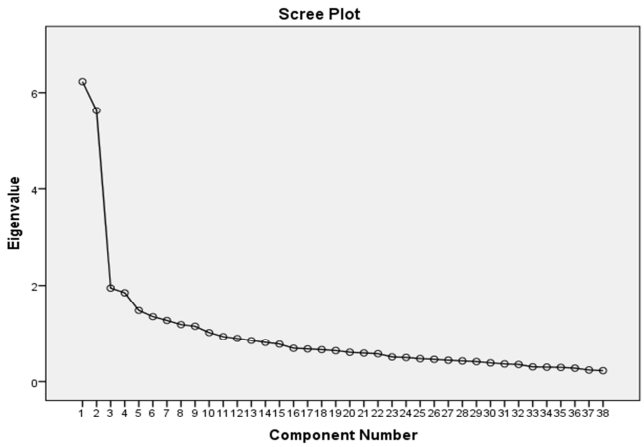
Communalities Before and After Extraction for AL Questionnaire

Item	Initial	Extraction	Item	Initial	Extraction
q01	1.00	.696	q20	1.00	.561
q02	1.00	.583	q21	1.00	.621
q03	1.00	.608	q22	1.00	.526
q04	1.00	.518	q23	1.00	.628
q05	1.00	.580	q24	1.00	.624
q06	1.00	.653	q25	1.00	.757
q07	1.00	.764	q26	1.00	.644
q08	1.00	.786	q27	1.00	.727
q09	1.00	.522	q28	1.00	.606
q10	1.00	.500	q29	1.00	.529
q11	1.00	.609	q30	1.00	.455
q12	1.00	.693	q31	1.00	.525
q13	1.00	.643	q32	1.00	.691
q14	1.00	.495	q33	1.00	.634
q15	1.00	.549	q34	1.00	.433
q16	1.00	.697	q35	1.00	.545
q17	1.00	.653	q36	1.00	.673
q18	1.00	.631	q37	1.00	.694
q19	1.00	.496	q38	1.00	.581

As shown in Table 7, most of the communalities are not above 0.7. This indicates that the criterion of retaining factors with eigenvalues above 1 could not be quite reliable and making decisions on the number of factors to be retained needs more care. Therefore, the researchers decided to look at the scree plot of the variables to decide how many factors to retain. Figure 1 displays the scree plot of the items in the AL questionnaire.

Figure 1

Scree plot of the variables in the AL questionnaire



The scree plot shows that the last sharp bend occurs from the eleventh factor and only ten factors are qualified to be retained. In order to consolidate the decision regarding the number of factors to retain, the researchers ran parallel analysis. Table 8 shows the random data eigenvalues resulting from parallel analysis.

Table 8

Random Data Eigenvalues

Root	Mean	Percentile	Root	Mean	Percentile
1.	1.65	1.74	20.	0.72	0.74
2.	1.57	1.62	21.	0.69	0.72
3.	1.51	1.56	22.	0.67	0.692
4.	1.46	1.51	23.	0.65	0.67
5.	1.42	1.46	24.	0.62	0.65
6.	1.29	1.33	25.	0.60	0.62
7.	1.19	1.23	26.	0.57	0.60
8.	1.12	1.15	27.	0.55	0.57
9.	1.09	1.12	28.	0.53	0.55
10.	0.98	1.01	29.	0.50	0.53
11.	0.97	0.98	30.	0.48	0.50
12.	0.93	0.96	31.	0.46	0.48
13.	0.91	0.93	32.	0.43	0.46
14.	0.88	0.90	33.	0.41	0.43
15.	0.85	0.88	34.	0.38	0.41
16.	0.82	0.84	35.	0.359	0.38
17.	0.79	0.82	36.	0.33	0.35
18.	0.77	0.79	37.	0.30	0.33
19.	0.74	0.77	38.	0.26	0.29

In parallel analysis, only factors that have eigenvalues higher than the random eigenvalues are retained. The comparison of the values in Table 8 with the eigenvalues in Table 6 indicates that for the first 10 items, the eigenvalues in Table 6 are higher than their corresponding random eigenvalues in Table 8. Accordingly, the results of the parallel analysis also confirmed that only 10 factors should be retained.

The factor rotation procedure started with an orthogonal rotation of the factors. The loadings of each item on different factors are displayed in Table 9; loadings below 0.4 are not shown.

Table 9

Rotated Component Matrix for AL Questionnaire

	Component									
	1	2	3	4	5	6	7	8	9	10
q23	.77									
q33	.73									
q24	.72									
q28	.71									
q29	.62									
q16		.80								
q06		.74								
q13		.68								
q10		.64								
q09		.58								
q01			.80							
q11			.69							
q04			.65							
q03			.54							
q25				.81						
q27				.74						
q32				.74						
q26					.74					
q35					.66					
q19					.64					
q20					.54					
q34					.54					
q30										
q07						.83				
q08						.78				
q12						.68				
q18							.73			
q21							.65			
q22							.64			
q31							.51			
q37								.82		
q36								.80		
q38								.68		
q02									-.72	
q15									.62	
q14									-.55	
q17										.78
q05										.67

Table 9 shows that all items except Item 30 loaded properly onto one of the ten extracted factors with a loading value above 0.4. This means that 37 items in the AL questionnaire were qualified to represent the underling factors of AL. However, Item 5 and Item 17, which were loaded onto the 10th factor, were also removed from the final AL questionnaire since it was not proper to have a factor with only two representative items. Based on the content of the items that loaded onto factors, the researchers gave a name to each factor. Table 10 displays the name of each factor.

Table 10

Emerging Factors in AL Questionnaire

Factor	Given Name
1	Test Construction
2	Administering, Rating, and Interpreting Test
3	Alternative and Digital-based Assessment
4	Giving Feedback in Assessment
5	Ethical and Cultural Considerations in Assessment
6	Psychometric Properties of a Test
7	Using and Interpreting Statistics
8	Recognizing Test Type, Distinction and Function
9	Authenticity

The Fourth Research Question

The aim of this question was to determine which of the components of the newly developed questionnaire is a better predictor of AL. To answer this question, a multiple regression analysis was run. The first step in running this analysis was checking its assumptions.

The first assumption is independence of residuals, which was checked using the Durbin-Watson statistic. This assumption was not applicable to our data since the independent and dependent variables were not measured independently, and the independent variables were the composing elements of the dependent variable.

The second assumption is multicollinearity. To test this assumption, the correlation between each pair of AL components were checked, the results of which are presented in Table 11. As the coefficients of correlation indicate, there is no high correlation between any pair of the AL components. Thus, the assumption of lack of multicollinearity has been met.

Table 11

Analysis of Multicollinearity between Each Pair of Independent Variables

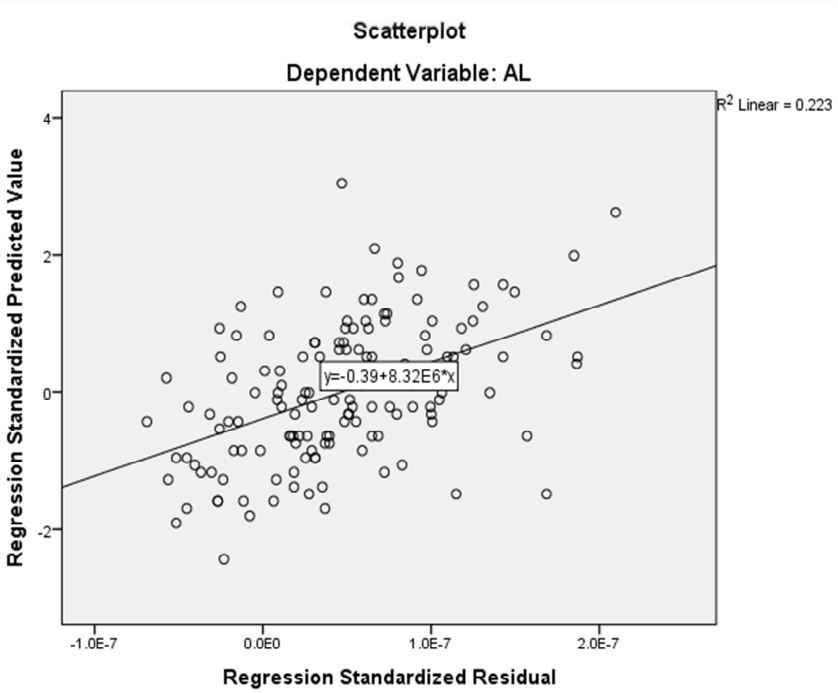
	F1	F2	F3	F4	F5	F6	F7	F8	F9
F1	1	-.102	-.017	.448*	-.417*	-.084	.447*	-.124	-.046
F2			.573*	-.174*	-.018	.551*	-.060	.060	-.129
F3			1	-.178*	-.139	.599*	.050	.085	-.040
F4				1	-.217**	-.073	.271*	-.044	.055
F5					1	-.009	-.404*	.065	-.027
F6						1	.056	.104	.041
F7							1	-.019	.097
F8								1	-.070
F9									1

The third assumption is homoscedasticity, according to which the variance of the residuals at each level of the predictors should be the same. In other words, there should be no pattern in the scatterplot of the Regression Standard Residual and

Regression Standardized Predicted Value. The lack of such a pattern is indicated in Figure 2.

Figure 2

Checking the Homoscedasticity for Multiple Regression Analysis



Having made sure that the assumptions were met, the researchers ran this analysis using the standard method. Table 12 contains the results of ANOVA which assesses the overall significance of the multiple regression model.

Table 12

ANOVA for Significance of the Multiple Regression Model

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13063.459	9	1451.495	.	.[b]
	Residual	.000	136	.000		
	Total	13063.459	145			

a. Dependent Variable: AL

b. Predictors: (Constant), F9, F5, F6, F8, F4, F7, F2, F1, F3

As shown in Table 13, the p-value is not reported in SPSS Output. The reason is that the components of AL are the all-and-only independent variables in the multiple regression analysis. Therefore, no inference can be made regarding the significance of the model. However, the predictability of the AL from its components was checked. Table 13 presents the results of this analysis.

Table 13

Standardized Coefficients for Multiple Regression Analysis

Model		Standardized		Coefficients
		Beta	t	Sig
1	(Constant)	.000	1.000	
	F1	.403	170941847.075	.000
	F2	.432	177983747.364	.000
	F3	.301	117976880.749	.000
	F4	.211	98978544.352	.000
	F5	.400	186480105.350	.000
	F6	.240	96256273.744	.000
	F7	.258	118480856.639	.000
	F8	.262	139755469.115	.000
	F9	.169	88723546.834	.000

a. Dependent Variable: AL

Based on Table 13, it could be concluded that all components of AL ($p < 0.01$) were significant predictors of AL. However, since the aim of the fourth research question was identifying the best predictor, it can be concluded that F2 (i.e., Administering, Rating, and Interpreting Test), which yielded a Beta coefficient of 0.432, was the best predictor in comparison to other components.

Discussion

This study attempted to find the main items of questionnaires developed in previous research. The items identified in this study cover both the theoretical, sociocultural, and psychological aspects of language assessment. They are claimed to be more comprehensive than previous studies. Previous studies have not covered all aspects of AL. For instance, Fulcher (2012) has considered only the theoretical aspects of AL. On the other hand, some studies have mainly focused on the sociocultural and / or psychological aspects. For instance, Moradan and Pourasadollah (2014) have focused on the psychological and emotional aspect of AL. It seems that the dominance of the theoretical aspect of testing is gradually diminishing. This can be attributed to the movement towards the cognitive and / or sociocultural aspects of teaching and testing. The problem of focusing on either the theoretical or the sociocultural / psychological aspect exists in questionnaires developed in the context of Iran as well. Many studies conducted in Iran have mainly focused on the theoretical aspect. For instance, the questionnaire developed and validated by Khanjani et al. (2017) focused on the theoretical aspect. Also, the

one developed by Mellati and Khademi (2018) considered issues such as assessment knowledge. Therefore, in spite of improvements in the AL of Iranian teachers, much more needs to be done to assist them to consider all aspects of second language assessment.

Generally, there were similarities between the previously developed questionnaires. For example, they did not appear to be comprehensive. Moreover, most of them did not consider technological issues. Also, they did not focus on teachers' knowledge related to both external and internal expectations of learners. This has been acknowledged in the relevant theories. For instance, as Fulcher (2012) has argued, some questionnaires do not attempt to explore teachers' ability to develop sociocultural and / or political bias-free assessment tools. As a result, these tools may not be able to assess learners comprehensively and miss some aspects. This can lead to a situation in which an important aspect of teachers' knowledge is neglected (Lan & Fan, 2019; Pastore & Andrade, 2019).

Also, this study found some new themes to be included in the newly developed questionnaire. The first one, using dynamic assessment, has not been explored in previous empirical research. The findings of this study showed that this theme should be considered and included in the newly developed questionnaire. In the process of the development of the questionnaire, the researchers attempted to consider the issue that tasks and materials should be selected and analyzed in a way that enables second language teachers and / or assessors to predict the sorts of problems which second language learners may face. The second theme was multicultural assessment approaches and subjects related to cultural, content, and linguistic bias in language testing. Not only language assessment, but also other associated aspects – such as curriculum development and language teaching itself – are influenced by multiculturalism (Sleeter & Carmona, 2017). Due to the expansion of global communication as well as mobility – especially in EFL contexts such as Iran – second language assessment should consider multiculturalism. Students coming from diverse cultural backgrounds need to be assessed with respect to the cultural issues, and cultural differences that might influence their performance should be considered. The third theme (computer-based testing, WBT, CALT and technology-based assessment) can be considered a rather new theme in comparison to the other themes. It may be argued that assessment through technology has various advantages over traditional ones. Especially due to the Coronavirus pandemic and the development of different websites and software / applications for designing and administering tests, it can be of great help. "Different types of feedback in assessment" is the last newly identified theme. Different types of feedback and their effectiveness need to be explored. However, previous research has not paid enough attention to this issue. As Hattie and Timperley (2007) have argued, different types of feedback may lead to various results. Students' perceptions and whether they reject, accept, or modify feedback is important and can demonstrate the effectiveness of feedback.

This study also found nine factors (components) of AL including (test construction, administering, rating and interpreting test, alternative and digital based assessment, giving feedback in assessment, ethical and cultural consideration in

assessment, psychometric properties of a test, using and interpreting statistics, recognizing test type, distinction and function and authenticity) as the main components of AL. These nine factors seem to provide a more comprehensive account than previous studies. As discussed earlier, previous research has usually paid attention to either the theoretical aspects (e.g., test construction, administering the test) or the social ones (e.g., ethical and cultural considerations) (Wright & Pandey, 2008). For instance, Popham (2009) considered only principles of fairness and principles of justice. On the other hand, Mertler (2003) focused merely on the theoretical aspect. It is fair to assume that, ignoring one aspect leads to an incomprehensive exploration of AL. Despite these differences, there are similarities between the findings of this study and those of previous ones. For instance, Fulcher (2012) suggested that teacher assessment includes knowledge, skills and abilities required to design, develop, maintain or evaluate large-scale standardized and / or classroom-based tests; familiarity with test processes; and an awareness of the principles and concepts that guide practice, including ethics and codes of practice. However, the components suggested by Fulcher seem to be too general, and it does not seem easy to operationalize them.

Another finding of this study was associated with administering, rating and interpreting tests as the best predictor of AL in comparison to the other eight components. They have traditionally been considered as one of the main issues in language testing. Consistent with previous research, the present study found that although all components of AL were significant predictors, "administering, rating and interpreting test" was the most important one. Plake and Impara (1993), similar to this study, found that knowledge of standardized testing was the most important factor. This finding is also supported by associated theories. Many theoretical studies have argued that AL should involve all aspects. For example, Davies (2008) has argued that AL involves knowledge of applied linguistics, theory and concepts and teachers' own language assessment context; knowledge as well as actions with regard to important issues in language assessment; and skills for instruction, design, educational measurement, and technological skills. Hence, all aspects are important and should be considered in language assessment.

Some of the previous studies have provided too general predictors for AL. For instance, Davies (2008) introduced skills, knowledge, and principles. Similarly, Inbar-Lourie (2008) suggested aspects such as why, what, and how assessment should be conducted. The reason for the difference between the findings of the present study and those of the previous ones seems to be associated with the idea that previously, studies mainly focused on theoretical issues – especially, as AL was a rather new concept (Black & Wiliam, 1998; Poehner & Lantolf, 2005). Gradually, researchers began to pay attention to more practical and more specific aspects.

Conclusion and implications

From the findings of this study, it can be concluded that the components that were found in this study as the underlying components of AL generally include factors related to cultural, social, and psychological aspects of AL. Also, teachers need to pay attention to factors related to the test itself. Moreover, there were some

repetitive components. Previously developed instruments have shown that various variables tend to affect assessment literacy. However, sometimes researchers seem to ignore some of these variables and focus only on some of them. In other words, they have not considered all aspects of AL. As previous studies have pointed out, determining these components is a challenge (Inbar-Lourie, 2008; Rea-Dickins, 2008). Since most of the teacher training courses, especially in the context of Iran, tend to focus only on language teaching itself and do not pay enough attention to assessment, this new questionnaire can help teachers to get more familiar with their own strengths and weaknesses regarding AL so that they can improve their knowledge and empowering themselves. One may conclude that teacher training courses need to be informed with respect to AL. More specifically, the AL of teachers should be specified in teacher training courses.

The findings of this study can be used by language teachers to get more familiar with the important aspects of language assessment. These findings include expert opinion; therefore, they can be useful particularly for novice assessors. Teachers can try to design and interpret tests which consider all aspects. Moreover, they can humanize their ways of assessment, that is, they can consider sociocultural and psychological aspects as well. Also, authorities can use the findings of the present study in order to design teacher training programs which attempt to familiarize teachers with all aspects, including both the traditional ones and the ones found by the present research study. Second language learners themselves can use these findings so that they can challenge and criticize language tests which do not consider cultural differences.

Acknowledgements

The present authors wish to acknowledge all the participants who cooperated in data collection stage and the anonymous experts and reviewers who helped us improve this work.

References

All, A., Castellar, E. P. N., & Van Looy, J. (2016). Assessing the effectiveness of digital game-based learning: Best practices. *Computers & Education*, *92*, 90-103. https://doi.org/10.1016/j.compedu.2015.10.007

Ashraf, H., & Zolfaghari, S. (2018). EFL teachers' assessment literacy and their reflective teaching. *International Journal of Instruction*, *11*(1), 425-436. https://doi.org/10.12973/iji.2018.11129a

Bangert-Drowns, R. L., Kulik, C. L. L., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213–238.

https://doi.org/10.3102%2F00346543061002213

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing*, *25*(3), 327-347. https://doi.org/10.1177%2F0265532208090156

Esfandiari, R., & Nouri, R. (2016). A mixed-methods, cross-sectional study of assessment literacy of Iranian university instructors: Implications for teachers' professional development. *Iranian Journal of Applied Linguistics*, *19* (2), 115-154. http://dx.doi.org/10.29252/ijal.19.2.115

Farhady, H., & Tavassoli, K. (2018). Developing a language assessment knowledge test for EFL teachers: A data-driven approach. *Iranian Journal of Language Teaching Research*, *6*(3), 79-94. https://dx.doi.org/10.30466/ijltr.2018.120602

Feuerstein, R., Miller, R., Rand, Y., & Jensen, M.R. (1981). Can evolving techniques better measure cognitive change? *The Journal of Special Education*, *15*(2), 201–219. https://doi.org/10.1177%2F002246698101500209

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly*, *9*(2), 113-132. https://doi.org/10.1080/15434303.2011.642041

Fulmer, G. W., Lee, I. C. H., & Tan, K. H. K. (2015). Multi-level model of contextual factors and teachers' assessment practices: An integrative review of research. *Assessment in Education: Principles, Policy, and Practice, 22*(4), 475–494. https://doi.org/10.1080/0969594X.2015.1017445

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112. https://doi.org/10.3102%2F003465430298487

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 1277-1288. https://doi.org/10.1177%2F1049732305276687

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing*, *25*(3), 385–402. https://doi.org/10.1177%2F0265532208090158

Janatifar, M., & Marandi, S. (2018). Iranian EFL teachers' language assessment literacy (LAL) under an assessing lens. *Applied Research on English Language, 7(3),* 361-382. https://dx.doi.org/10.22108/are.2018.108818.1221

Jeong, H. (2013). Defining assessment literacy: Is it different for language testers and non-language testers? *Language Testing, 30*(3), 345–362. https://doi.org/10.1177%2F0265532213480334

Khanjani, A., Vahdany, F., & Jafarigohar, M. (2017). EFL teacher education in Iran: Does it promote trainees' pedagogical content knowledge? *Journal of Research in Applied Linguistics, 8*(2), 159-186. https://dx.doi.org/10.22055/rals.2017.13096

Lam, R. (2015). Language assessment training in Hong Kong: Implications for language assessment literacy. *Language Testing, 32*(2), 169-197. https://doi.org/10.1177%2F0265532214554321

Lan, C., & Fan, S. (2019). Developing classroom-based language assessment literacy for in-service EFL teachers: The gaps. *Studies in Educational Evaluation*, *61*(3), 112-122. https://doi.org/10.1016/j.stueduc.2019.03.003

Mellati, M., & Khademi, M. (2018). Exploring teachers' assessment literacy: Impact on learners' writing achievements and implications for teacher development. *Australian Journal of Teacher Education*, *43*(6), 1-18. http://dx.doi.org/10.14221/ajte.2018v43n6.1

Malone, M. E. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing, 30*(3), 329–344. https://doi.org/10.1177%2F0265532213480129

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*(2), 200–206. https://doi.org/10.3758/BF03194052

Mertler, C. A. (2003). Preservice versus in-service teachers' assessment literacy: Does classroom experience make a difference? *Educational Resources Information Center (ERIC),*1-31.

Moradan & Pourasadollah. (2014). Teachers' thinking about their teaching: A critical study on Iranian TEFL teachers. *Procedia - Social and Behavioral Sciences, 98*, 1194 – 1203. https://doi.org/10.1016/j.sbspro.2014.03.534

Moafian, F., & Pishghadam, R. (2009). Construct validation of a questionnaire on characteristics of successful Iranian EFL teachers. *Pazhuhesh-e Zabanha-ye Khareji*, *54*, 127-142.

O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing, 30*(3), 363–380. https://doi.org/10.1177%2F0265532213480336

Ölmezer-Öztürk, E., & Aydin, B. (2018). Toward measuring language teachers' assessment knowledge: Development and validation of Language Assessment Knowledge Scale (LAKS). *Language Testing in Asia*, *8*(1), 1-15. https://doi.org/10.1186/s40468-018-0075-2

Pastore, S., & Andrade, H. L. (2019). Teacher assessment literacy: A three dimensional model. *Teaching and Teacher Education*, *84*, 128-138. https://doi.org/10.1016/j.tate.2019.05.003

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. https://doi.org/10.1111/j.1745-3992.1993.tb00548.x

Poehner, M .E., & Lantolf, J. P. (2005). Dynamic assessment in the language classroom. *Language Teaching Research, 9*(3), 233-265. https://doi.org/10.1191%2F1362168805lr166oa

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice, 48*(1), 4–11. https://doi.org/10.1080/00405840802577536

Razavipour, K., Riazi, A., & Rashidi, N. (2011). On the interaction of test washback and teacher assessment literacy: The case of Iranian EFL secondary school teachers. *English Language Teaching, 4*(1), 156-161. https://doi.org/10.5539/elt.v4n1p156

Rea-Dickins, P. (2008). Classroom-based language assessment. *Encyclopedia of Language and Education, 7*, 257-272. https://doi.org/10.1007/978-0-387-30424-3_180

Rezvani Kalajahi, S. A., & Abdullah, A. N. (2016). Assessing assessment literacy and practices among lecturers. *Pedagogika, 124*(4), 232–248. http://dx.doi.org/10.15823/p.2016.65

Roever, C., & McNamara, T. (2006). Language testing: The social dimension. *International Journal of Applied Linguistics*, *16*(2), 242-258. https://doi.org/10.1111/j.1473-4192.2006.00117.x

Savignon, S. J. (2018). Communicative competence. In J. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 1–7). Wiley. https://doi.org/10.1002/9781118784235.eelt0047

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. https://doi.org/10.1177%2F0265532213480128

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing*, *18*(4), 373-391. https://doi.org/10.1177%2F026553220101800404

Sleeter, C. E., & Carmona, J. F. (2017). *Un-Standardizing curriculum: Multicultural teaching in the standards-based classroom*. Teachers College Press.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics, 29*, 21–36. https://doi.org/10.1017/S0267190509090035

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. https://doi.org/10.2307/20466661

Wright, B. E., & Pandey, S. K. (2008). Public service motivation and the assumption of person-organization fit: Testing the mediating effect of value congruence. *Administration & Society*, *40*(5), 502-521. https://doi.org/10.1177%2F0095399708320187

Zygouris, S., & Tsolaki, M. (2015). Computerized cognitive testing for older adults: A review. *American Journal of Alzheimer's Disease & Other Dementias*, *30*(1), 13-28. https://doi.org/10.1177%2F1533317514522852

Authors' Biographies

Behnaz Rastegar is a PhD candidate of English Language Teaching (ELT) at Imam Khomeini International University in Qazvin, Iran. Her areas of interest are Second Language Learning and Acquisition, and Language Assessment.

Abbas Ali Zarei, an Associate Professor in English Language Teaching at Imam Khomeini International University in Qazvin, has had ۱۲۸ papers and 57 books published. He has also attended and presented at 75 national and international conferences. His main areas of interest include the teaching and learning of English vocabulary and its subcategories including English collocations and idioms**.**

Rajab Esfandiari is an Associate Professor of Applied Linguistics at Imam Khomeini International University in Qazvin, Iran. His areas of interest and specialization include Teaching and Assessing L2 Writing, Multifaceted Rasch Measurement, L2 Classroom Assessment, and EAP Teaching and Testing.