Iranian Journal of Applied Linguistics (*IJAL*)

**Vol. 23, No. 2, September 2020, 106- 145**

# Language Assessment Courses at Iranian State Universities: Are they Comprehensive Enough to Develop Valid Language Assessment Literacy (LAL) among EFL Students?

**Masoomeh Taghizadeh, Golnar Mazdayasna*, Fatemeh Mahdavirad**

*Department of English Language and Literature, Yazd University, Yazd, Iran*

## Abstract

In the educational setting of Iran, language assessment literacy (LAL) is still an underexplored issue. This paper investigated the development of LAL among EFL students taking language assessment course at state universities in Iran. The three components of LAL (i.e., knowledge, skills, and principles) were the focus of the inquiry. To collect the required data, a questionnaire, encompassing 83 Likert items and a set of open-ended questions, was developed, and responses from 92 course instructors were collected. Teaching and assessment practices of two course instructors were also observed throughout an educational semester. SPSS (26) was used to analyze the data. Findings revealed that these courses mainly focused on knowledge and skills, overlooking the principles of assessment. Adherence to traditional assessment approaches, use of inappropriate teaching materials, and lack of practical works in assessment also characterized the investigated courses. The paper concludes with suggestions to better design language assessment courses to increase the assessment literacy of English graduates who will probably enter the teaching contexts after graduation.

**Keywords:** Language Assessment Literacy (LAL), Assessment/testing knowledge, Assessment/testing skills, Assessment/testing principles

**Article Information:**

**Received:** 29 May 2020 **Revised:** 1 August 2020 **Accepted:** 12 August 2020

**Corresponding author:** Department of English Language and Literature, Yazd University, Yazd, Iran. **Email:** gmazdayasna@yazd.ac.ir

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

## 1. Introduction

In line with changes in language teaching and learning, there have also been shifts in assessment and testing views. These changes have required teachers to assess students' broader knowledge and life skills and expected them to consider assessments as having very close relationships with instruction and learning. However, research in educational context has shown that teachers' insufficient knowledge of assessment issues has often led them to ignore quality assurance in their activities and associate their assessment practices with traditional assessment and use of poorly designed tests (Alderson, 2005; Popham, 2011; Price, 2005; Stiggins, 1991). Upon recognizing such a critical role of assessment in education, assessment specialists and educational researchers have seriously called for teachers to be assessment-literate, given that they can play a crucial role in the functionality and usefulness of assessment in education (Popham, 2011; Price, 2005; Stiggins, 1991).

However, despite such emphasis on this critical topic, there are only a few studies that have focused on the issue within the Iranian context. Therefore, the current study attempted to shed more light on this issue, with a specific focus on the nature of language assessment courses at Iranian state universities and their efficacy in developing English students' LAL.

## 2. Review of Literature

Popham (2011) asserts that for teachers to develop professional competence and to promote students' learning and prove fruitful to their institutions, assessment literacy is a critical requirement, and for this purpose, teachers require a valid knowledge base of the assessment process. The results of a study by Sikka, Nath, and Cohen (2007), investigating in-service teachers' beliefs and assessment practices, suggested a requirement for inclusion and employment of various kinds

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

of assessments in teacher education programs. Lukin, Bandalos, and Eckhout (2004) explored the role of assessment training and found that it affected teachers' assessment knowledge and skills positively and enhanced their confidence.

As in general education, there have also been calls within the language education field for EFL/ESL teachers to develop assessment literacy (e.g., Fulcher, 2012; Giraldo, 2018; Lam, 2014; Marhaeni et al., 2020; Scarino, 2013; Shim, 2009). Shim (2009) explored teachers' attitudes and practices concerning classroom-based language assessment. Results showed that teachers were familiar with the assessment principles and had developed an adequate knowledge base in testing, yet they did not transfer this knowledge into their practices. Marhaeni et al. (2020), investigating the assessment literacy of Indonesian English teachers, showed that for them, LAL was categorized as average in different aspects. With respect to pre-service teachers, Lam (2014) points out that training pre-service language teachers and enabling them to employ sound assessment in their prospective work settings has been neglected.

Consequently, there is an urgent need to consider the development of LAL among EFL/ESL teachers as a necessary element of their teacher training programs, and, as stated by Fulcher (2012), it is essential to study the extent to which language teachers have been trained to manage LAL. Nevertheless, the number of studies conducted on language teacher's assessment literacy in Iran is rare. To our knowledge, no study has examined the nature and functionality of assessment training in developing language assessment literacy within the Iranian state universities. Therefore, this study attempted to investigate the degree to which language assessment courses at the B.A. level in the field of English language have been well designed to raise prospective teachers' awareness of the essentials of assessment by examining a course instructor survey and classroom observations. In the B.A English language syllabus, defined by the Ministry of Science, Research and Technology (MSRT), a two-credit compulsory module concerning assessment and testing is presented for language students at state

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

universities. The general syllabus of the course, including the number of the sessions and time allocattion, is introduced by MSRT, however, the course instructors are the ones who decide on issues such as teaching materials and instructional and assessement approaches. Language graduates usually enter the teaching context either in private language institutes or public schools.

The conceptual framework of the current investigation into assessment literacy is driven by Davies' (2008) three components of language assessment literacy, that is, *skills, knowledge,* and *principles*. In this classification, knowledge pertains to the relevant background in language and measurement. It also involves examining different language learning models and theories in language teaching and testing. Skills refer to the relevant methodology in testing and assessment, such as item construction, doing statistics, test revision, and tactics in using the software programs required. Principles address concepts such as validity, reliability, ethics, appropriate use of language test, fairness, test impact, etc. Davies (2008) argues that there has been a movement from *skills* to *skills + knowledge* to *skills + knowledge + principles.* Furthermore, he emphasizes that *skills + knowledge* is insufficient without the inclusion of principles. Hence, as Davies (2008, p.329) rightly points out, "careful balancing of the practical (the skills) with the descriptive (the knowledge) and the theoretical (the principles) are needed. All are necessary, but one without the other(s) is likely to be misunderstood and/or trivialized". Hence, in this research, LAL is defined as acquiring knowledge, skills, and principles necessary to construct, interpret, evaluate, and use different types of tests. Based on the framework discussed, the following research questions directed the design of this study:

1. To what extent do language assessment course instructors at state universities incorporate aspects related to each LAL component into their B.A. program?
2. What kind of assessment approaches and practices do instructors employ?

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

3. What are the shortcomings, if any, of language assessment courses in developing students' LAL?

## 3. Method

To conduct the present study, it seemed rational to adopt both quantitative and qualitative research methods, as deficiencies have been attributed to the sole application of either method (Dörnyei, 2007). Two research methods -a survey and classroom observation- have been recognized as appropriate for the investigation of the research questions which drive the present study. The researchers aimed at constructing instruments that (a) reflect the multiple dimensions of assessment literacy as indicated in Davies' (2008) classification of LAL and (b) include aspects of assessment as advocated by contemporary standards (e.g., Classroom Assessment Standards (JCSEE, 2015)). However, as the assessment-related aspects underlying each component have not been fully outlined in the relevant literature, the preceding step to instruments design was to find out the aspects belonging to each overarching component. For this purpose, renowned textbooks, written by significant scholars in the field of language testing and assessment were taken into consideration, and the literature on AL/LAL was thoroughly reviewed. Experts and colleagues' consultations were also sought on placing each specific aspect of LAL into the relevant component. The results formed the base for the conduct of both the survey and the observations.

### *3.1. Survey*

#### *3.1.1. Instrument*

Questionnaire has been identified as an appropriate tool for this particular research agenda as it provides a self-report research instrument which also makes the generalizability of the findings more feasible (Dörnyei, 2007). For the purpose

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

of this study, Brown and Bailey's (2008) questionnaire on the description of language assessment courses seemed inspiring; however, not complete enough as the questionnaire mainly focused on traditional testing issues. Besides, the questionnaire mainly covered issues related to the skills and knowledge of assessment, and issues related to the principles of assessment were not adequately covered. Hence, it was decided to add the new issues in assessment into the questionnaire and incorporate the three components in a more balanced way. Specifically, the questionnaire was developed to examine the teaching content, teaching materials of the course, instructors' approaches towards assessment, and the problems faced in the course. The questionnaire included three sections. The initial questions were related to the demographic information of the participants and the course. The content section of the questionnaire consisted of 83 five-point Likert items, ranging from 0, indicating *none,* to 4, meaning *extensively*, representing the estimated amount of time working on each particular aspect. Three open-ended questions also explored assessment procedures, teaching materials, and the practical problems usually faced in the course.

### *3.1.2. Participants*

Ninety-two language assessment course instructors participated in the data collection process.

### *3.1.3. Procedure*

After reviewing the relevant literature, the preliminary version of the questionnaire was prepared. The validity of the draft version was verified by two experts based on the two criteria motioned. Although there were some comments concerning the time-consuming nature of the questionnaire, it could not be avoided as the researcher wished to construct a comprehensive questionnaire. Finally, the questionnaire was piloted among 30 instructors. Cronbach's analysis

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

indicated inter-item reliability indices of (.73), (.70), (.75) for the three sections of knowledge, skills, and principles, respectively.

The starting point to find potential participants for the study was to search through university websites to find programs presenting a language assessment course in their B.A. syllabus. The questionnaire was delivered by e-mail or through personal visits. The instructors were uninformed about the exact purpose of the study; however, they were assured about the confidentiality of their responses. Some weeks after the initial sending of the questionnaire, instructors received a reminder. All survey responses were dated between April and June, 2019. Among the 121 instructors surveyed, 92 responded, representing a response rate of 76%. To analyze the data, SPSS 26.0. was employed. Frequencies and percentages were calculated for questions in sections I and III and means and standard deviations for items in section II. T-test analysis was also used to determine the significance of the differences, if any.

### *3.2. Observation*

Questionnaires are good devices to get a general understanding of the efficacy of language assessment courses in developing LAL among students; however, in order to get a deeper insight into what really happens in such courses, field study was required. Classroom observation was recognized as a suitable choice as it gives the researcher the opportunity to see closely what teachers are doing in the classroom rather than having to rely on what they say they do (Dörnyei, 2007).

#### *3.2.1. Observational tools*

For the purpose of the study, an observational scheme was required to be designed. According to Dörnyei (2007), observation schemes ease up the objective and systematic description of classroom events and behaviors, facilitating cross-study comparisons in different contexts and increasing the generalizability of research. The designed observation scheme included two main

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

parts. A pre-observation form which sought specific information about the session to be observed such as the topic, lesson plan, learning outcomes, teaching material and activities, and instructors' evaluation of their adherence to the course syllabus. This form was filled out by the instructors before each session started. However, findings based on this form are not of consideration in this paper due to length-limits. Instead, the focus is on data based on the observation checklist which included three sections. The first section involved information such as class and session number, starting and ending time, number of students, and so on. The second section investigated instructors' component focus (i.e., knowledge, skills, and principles) based on the subtopics taught. Section three, which explored instructors' assessment approaches and practices, consisted of a set of Likert-scale items for which the observer needed to tick whether the aspect mentioned was observed extensively, moderately, a little, or not at all. For all three sections, some space was provided for taking field notes on processes, situations, interactions, and tasks/activities. Classroom events were also audio-recorded for further checks and analyses.

### *3.2.2. Participants*

Teaching practices of two language assessment course instructors, holding Ph.D. in TEFL, were observed throughout one academic semester.

### *3.2.3. Procedure*

The preliminary task for the observation of the classes was to design a suitable, fit-for-purpose observation scheme. Two experts confirmed the relevance of the designed scheme to the reseatch agenda. The courses were surveyed for 15 sessions at a 2019 semester course in language assessment at two universities in Kerman and Rafsanjan which were accessible to one of the researchers assigned as the observer. The classes were held once a week, and a total of 83 students were enrolled in the course. The researcher's assent with data protection, ethical

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

considerations, and guarantee of anonymity were assured. During the class time, the observation checklist was filled out by the observer, notes were taken, and sessions were audio-recorded for further checks. A post-hoc rating scale coding procedure was used after the observation session, through which decisions were made on the frequency of each event/behavior along a scale ranging from "extensively" to "not at all". To address the reliability issue, with the instructors' permission, sessions were audio- recorded, and inter-rater reliability was carried out to determine whether the coders make the same coding decisions. When the coding procedure was completed, 85% coding agreement was achieved. In qualitative studies, validity is related to accuracy of the information obtained through the data collection processes and analyses (Dörnyei, 2007). Non-participatory role was taken and the instructors were kept uninformed of the purpose of the research to avoid personal presuppositions influencing the results of the study and to minimize bias and enhance validity. For the validity of the decisions, two experts' consultations were also sought. In addition to the padagogical prcatices, formal assessments (e.g., quizzes, mid-term/ final exam sheets) were also analyzed to determine the degree to which LAL requirements were reflected in teachers' assessment practices.

In the analysis phase, decisions were made on the component focus and the extent to which each variable related to assessment practices was observed. Descriptive statistics were presented for time through its transformation into 5-minute units. In the next step, univariate analysis and multiple comparisons for the component focus and assessment practices were performed within and across the classes.

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

## 4. Results

### *4.1. Survey*

#### *4.1.1. General Information about the Course*

All classes were held for about sixteen sessions, each lasting for about 90 minutes. Concerning class size, 33% were classes of fewer than 30 students, 58% between 30 to 40 students, and 9% more than 40 students.

#### *4.1.2. Teaching Content*

Section II, including 83 Likert items, investigated the content of teaching with respect to the three components of LAL. Reliability was checked for the consistency of the responses through Cronbach's analysis which indicated satisfactory alpha levels of .73, .88, and .77 for each component of knowledge, skills, and principles, respectively. For each dimension, sub-dimension, and variable, the Mean (S.D.) and Median (IQR) are reported in tables 1, 2, and 3. In the text, the mean and the standard deviation (Mean (S.D.)) are provided to ease the average response interpretations and to compare them across different items.

##### *4.1.2.1. Knowledge of testing and assessment*

The first 26 items elicited the amount of focus devoted to the knowledge aspects of assessment and testing, including issues related to basic concepts in testing, history of testing, approaches to testing/assessment, and different functions and types of tests. As shown in table 1, a total mean of 2.48 (017) was reported for this component of LAL. Basic concepts in testing received a relatively high mean of 3.03 (0.64), and history of testing got a mean coverage of 2.44 (0.44). In this regard, the course mainly focused on providing information on psychometric-structuralist (3.32 (0.63)) stage in testing, compared to the sociolinguistic-pragmatic stage (1.70 (0.78)). Likewise, for approaches to testing/assessment, with the total mean of 2.22 (0.43), the mean ratings for the discrete-point approach (3.26 (0.68)) and integrative approach (2.87 (0.80)) seemed relatively

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

high, compared to the communicative (1.64 (0.76)) and performance-based (1.10 (0.80)) approaches. Among the different functions that tests can perform, with the total mean of 2.67 (0.37), it seemed that achievement tests (3.33 (0.8)), proficiency tests (2.85 (0.74)), and selection tests (2.90 (0.61) were the focused ones, unlike diagnostic tests (1.80 (0.76)). Regarding types and classifications of tests, with the total mean of 2.43 (0.21), summative assessment (3.24 (0.64)), norm-referenced tests (3.04 (0.80)) and large-scale testing (3.09 (0.71)) were covered almost highly, with formative assessment (1.35 (0.87)), alternative assessment (1.24 (0.84)), and computer adaptive tests (0.78 (0.75)) receiving the lowest means.

Table 1. *Descriptive analysis for variables related to knowledge*

| | | | Variables | | Sub-component | | Main component | |
|---|---|---|---|---|---|---|---|---|
| Main component | Sub-component | Variables | Mean (SD) | Median (IQR) | Mean (SD) | Median (IQR) | Mean (SD) | Median (IQR) |
| Knowledge | | basic concepts in testing(test, measurement, assessment ,etc.) | 3.03 (0.64) | 3.00 (3.00 - 3.00) | ----- | ----- | 2.48 (0.17) | 2.50 (2.38 - 2.58) |
| | History of testing/assessment | pre-scientific stage | 2.30 (0.75) | 2.00 (2.00 - 3.00) | 2.44 (0.44) | 2.33 (2.00 - 2.67) | | |
| | | psychometric-structuralist stage | 3.32 (0.63) | 3.00 (3.00 - 4.00) | | | | |
| | | sociolinguistic-pragmatic stage | 1.70 (0.78) | 2.00 (1.00 - 2.00) | | | | |
| | Approaches to testing/Assessment | discrete-point approach to testing | 3.26 (0.68) | 3.00 (3.00 - 4.00) | 2.22 (0.43) | 2.25 (2.00 - 2.50) | | |
| | | integrative approach to testing | 2.87 (0.80) | 3.00 (2.00 - 3.00) | | | | |
| | | communicative approach to testing | 1.64 (0.76) | 2.00 (1.00 - 2.00) | | | | |
| | | performance-based approach to testing | 1.10 (0.80) | 1.00 (1.00 - 2.00) | | | | |
| | o n | aptitude tests | 2.48 (0.93) | 3.00 (2.00 - | 2.67 | 2.67 | | |

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

| | | | | | |
|---|---|---|---|---|---|
| | | | 3.00) | (0.37) | (2.50 - 3.00) |
| Types of Test | diagnostic tests | 1.80 (0.76) | 2.00 (1.00 - 2.00) | 2.43 (0.21) | 2.42 (2.25 - 2.58) |
| | placement tests | 2.68 (0.84) | 3.00 (2.00 - 3.00) | | |
| | proficiency tests | 2.85 (0.74) | 3.00 (2.00 - 3.00) | | |
| | achievement tests | 3.33 (0.58) | 3.00 (3.00 - 4.00) | | |
| | selection tests | 2.90 (0.61) | 3.00 (3.00 - 3.00) | | |
| | objective tests | 3.40 (0.49) | 3.00 (3.00 - 4.00) | | |
| | subjective tests | 2.99 (0.65) | 3.00 (3.00 - 3.00) | | |
| | norm-referenced tests | 3.04 (0.80) | 3.00 (2.00 - 4.00) | | |
| | criterion-referenced tests | 2.58 (0.77) | 3.00 (2.00 - 3.00) | | |
| | summative assessment | 3.24 (0.64) | 3.00 (3.00 - 4.00) | | |
| | formative assessment | 1.35 (0.87) | 1.00 (1.00 - 2.00) | | |
| | alternative assessment | 1.24 (0.84) | 1.00 (1.00 - 2.00) | | |
| | large-scale testing | 3.09 (0.71) | 3.00 (3.00 - 4.00) | | |
| | classroom testing | 2.47 (0.78) | 3.00 (2.00 - 3.00) | | |
| | high-stakes tests | 2.86 (0.67) | 3.00 (2.00 - 3.00) | | |
| | low-stakes tests | 2.18 (0.80) | 2.00 (2.00 - 3.00) | | |
| | computer adaptive tests | 0.78 (0.75) | 1.00 (0.00 - 1.00) | | |

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

### *4.1.2.2. Skills of testing and assessment*

The next 36 items in this section tapped the amount of coverage devoted to teaching and practicing testing and assessment skills (see table 2 for the results). A mean rating of 2.13 (0.17) was obtained for this component of LAL.

A mean of 2.53 (0.71) was reported for the test design sub-component. In this area, principles and practice of item writing received a mean of 3.13(0.79); however, the construction of test syllabuses/item specification was not adequately covered (1.99 (0.82)). There seemed not to be much attempt in developing students' skills in using alternative assessment procedures (1.70 (0.77)). Practice in testing language skills/sub-skills (2.43 (0.40)) was mostly devoted to testing grammar and structure (3.35 (0.65)), testing vocabulary (2.95 (0.78)) and testing reading comprehension (2.66 (0.89)). The sub-component of item/test analysis received a mean of 2.07(0.88). Analyzing item characteristics seemed to be of average consideration (2.47(1.02)). However, focus on test revision (1.82 (0.74)) and test critiquing (1.93 (0.78)) seemed to be low. Issues in test administration (1.59 (0.90)), test scoring (1.78 (0.81)), and using different types of interpretation (NR/ CR) (2.47 (0.78)) were not so much focused. Administering and using computer-/internet-based testing received means of less than 1. In addition, moderate coverage of 2.13 (0.57) and 2.30 (0.38) was found for strategies to estimate test reliability and validity, respectively. Regarding instruction in using statistics, descriptive analysis (2.23 (0.61)) seemed to be covered to some extent; whereas, doing inferential analysis seemed to be the skipped part of all classes, receiving a mean of zero.

Table 2. *Descriptive analysis for variables related to skills*

| | | **Variables** | | **Sub-component** | | **Main component** | |
|---|---|---|---|---|---|---|---|
| Sub-component | Variables | Mean (SD) | Median (IQR) | Mean (SD) | Median (IQR) | Mean (SD) | Median (IQR) |

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Skill | Test design | construction of test syllabuses | 1.99 (0.82) | 2.00 (1.00 - 3.00) | 2.53(0.71) | 2.50 (2.25-2.74) | 2.13 (0.17) | 2.11 (2.00 - 2.25) |
| | | principles and practice of item writing | 3.13 (0.79) | 3.00 (3.00 - 4.00) | | | | |
| | | developing selected-response items | 3.32 (0.51) | 3.00 (3.00 - 4.00) | | | | |
| | | developing constructed-response items | 3.08 (0.60) | 3.00 (3.00 - 3.00) | | | | |
| | | developing alternative assessment | 1.70 (0.77) | 2.00 (1.00 - 2.00) | | | | |
| | | Developing assessment at different levels | 2.00 (0.77) | 2.00 (1.25 - 3.00) | | | | |
| | Testing language skills /subskills | testing listening | 1.86 (0.76) | 2.00 (1.00 - 2.00) | | | | |
| | | testing reading | 2.66 (0.89) | 3.00 (2.00 - 3.00) | | | | |
| | | testing speaking | 1.73 (0.98) | 2.00 (1.00 - 2.00) | | | | |
| | | testing writing | 2.07 (0.85) | 2.00 (1.00 - 3.00) | | | | |
| | | testing vocabulary | 2.95 (0.78) | 3.00 (2.00 - 4.00) | | | | |
| | | testing grammar | 3.35 (0.65) | 3.00 (3.00 - 4.00) | | | | |
| | Analyzing items/test | analyzing item characteristics | 2.47 (1.02) | 3.00 (2.00 - 3.00) | | | | |
| | | test item revision | 2.09(0.98) | 3.00 (2.00 - 3.00) | | | | |
| | | test paper revision | 1.82 (0.74) | 2.00 (1.00 - 2.00) | | | | |
| | | test critiquing | 1.93 (0.78) | 2.00 (1.00 - 2.00) | | | | |
| | | test administration | 1.59 (0.90) | 2.00 (1.00 - 2.00) | | | | |
| | | test scoring | 1.78 (0.81) | 2.00 (1.00 - | | | | |

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

| | | | | | |
|---|---|---|---|---|---|
| Test administration Scoring /interpretation | developing and using a scoring rubrics | 0.82 (0.63) | 1.00 (0.00 - 1.00) | 1.35(0.69) | 1.40 (0.84-0.93) |
| | administering and scoring computer- | 0.12 (0.33) | 0.00 (0.00 - 0.00) | | |
| | using different types of interpretations( | 2.47 (0.78) | 2.00 (2.00 - 3.00) | | |
| Establishing Reliability | internal consistency | 2.59 (0.61) | 3.00 (2.00 - 3.00) | 2.13 (0.57) | 2.00 (1.75 - 2.50) |
| | test-retest reliability | 2.57 (0.67) | 3.00 (2.00 - 3.00) | | |
| | parallel form reliability | 2.41 (0.70) | 2.00 (2.00 - 3.00) | | |
| | split half reliability | 2.40 (0.70) | 2.00 (2.00 - 3.00) | | |
| | K-R 20 | 1.89 (0.79) | 2.00 (1.00 - | | |
| | K-R 21 | 1.93 (0.78) | 2.00 (1.00 - | | |
| | intra-rater reliability | 1.61 (0.78) | 2.00 (1.00 - 2.00) | | |
| | inter-rater reliability | 1.62 (0.77) | 2.00 (1.00 - 2.00) | | |
| Establishing Validity | content validity | 2.55 (0.65) | 3.00 (2.00 - | 2.30 (0.38) | 2.33 (2.00 - 2.50) |
| | construct validity | 2.38 (0.71) | 2.00 (2.00 - 3.00) | | |
| | face validity | 2.34 (0.67) | 2.00 (2.00 - | | |
| | criterion-related validity | 2.01 (0.88) | 2.00 (1.00 - 3.00) | | |
| | concurrent validity | 2.34 (0.72) | 2.00 (2.00 - 3.00) | | |
| | predictive validity | 2.16 (0.70) | 2.00 (2.00 - 3.00) | | |
| Doing statistics | practice in doing descriptive statistics | 2.23 (0.61) | 3.00 (3.00 - 4.00) | 1.11(0.30) | 1.00 (0.74-1.02) |
| | practice in doing inferential | 0.00 (0.00) | 0.00 (0.00 - 0.00) | | |

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

### *4.1.2.3. Principles of testing and assessment*

The last 22 items estimated the amount of focus on the principles of assessment and testing. Based on the results (see table 3), all the aspects investigated received a total mean coverage of 1.81 (0.4). Considerations of test reliability (2.42 (0.79)) and test validity (2.04 (0.63)) were the focused aspects. An average mean of 2.17 (0.76) was devoted to standard-setting in language testing. Ethical practices in students' preparation for assessment (2.03 (0.70)), test administration (2.03 (0.70), test scoring (2.13 (0.70)), and test scores interpretation (2.11 (0.72)) were the next considered issues. Making sound decisions based on the test/assessment (2.14 (0.79)) received average attention. Other aspects such as considerations of test washback on teaching and learning/ test impact on society, fairness in testing, influence of societal/cultural values in testing, authenticity in testing, doing assessment based on multiple sources of evidence, critical approaches to language testing, importance of incorporating formative/alternative assessment, importance of incorporating learner autonomy and self-assessment in testing, and language program evaluation received mean coverage of less than 2.

Table 3. *Descriptive analysis for variables related to principles*

| | | **Variables** | | **Main Dimension** | |
|---|---|---|---|---|---|
| Main Dimension | Variables | **Mean (SD)** | **Median (IQR)** | **Mean (SD)** | **Median (IQR)** |
| Principles | standard setting in language testing | 2.17 (0.76) | 2.00 (2.00 - 3.00) | 1.81 (0.54) | 1.95 (1.32 - 2.09) |
| | ethical practices in students' preparation for assessment | 2.03 (0.70) | 2.00 (2.00 - 3.00) | | |
| | ethical practices in test administration | 2.03 (0.70) | 2.00 (2.00 - 3.00) | | |
| | ethical practices in test scoring | 2.13 (0.70) | 2.00 (2.00 - 3.00) | | |
| | ethical practices in test scores interpretation | 2.11 (0.72) | 2.00 (2.00 - 3.00) | | |
| | making sound decisions based on the test/assessment | 2.14 (0.79) | 2.00 (1.25 - 3.00) | | |
| | considerations of test | 2.42 (0.79) | 2.00 (2.00 - 3.00) | | |

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

| | | |
|---|---|---|
| reliability | | |
| considerations of test validity | 2.04 (0.63) | 2.00 (2.00 - 2.00) |
| Considerations of test washback(on teaching and learning) | 1.84 (0.77) | 2.00 (1.00 - 2.00) |
| Considerations of test impact(on society) | 1.58 (0.82) | 2.00 (1.00 - 2.00) |
| Considerations of test authenticity | 0.99 (0.70) | 1.00 (1.00 - 1.00) |
| Considerations of fairness in assessment/ testing | 1.99 (0.65) | 2.00 (2.00 - 2.00) |
| Doing assessment based on multiple sources of evidence | 1.47 (0.80) | 1.00 (1.00 - 2.00) |
| language program evaluation | 1.73 (0.66) | 2.00 (1.00 - 2.00) |
| Importance of incorporating formative assessment | 1.63 (0.75) | 2.00 (1.00 - 2.00) |
| Importance of incorporating alternative assessment | 1.66 (0.75) | 2.00 (1.00 - 2.00) |
| Importance of incorporating learner autonomy and self-assessment in testing | 1.48 (0.70) | 1.00 (1.00 - 2.00) |
| The influence of societal values on testing practices | 1.51 (0.69) | 2.00 (1.00 - 2.00) |
| Critical approaches to language testing | 1.64 (0.81) | 2.00 (1.00 - 2.00) |
| Importance of considering cultural aspects in assessment | 1.55 (0.70) | 2.00 (1.00 - 2.00) |
| Testing in relationship to curriculum | 1.68 (0.61) | 2.00 (1.00 - 2.00) |

As indicated by the above results, in these courses, knowledge, skills, and principles of assessment are not covered equally. Tables 4 shows that the mean differences are significant using the Tukey HSD test (p-value<0.001).

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

Table 4. *The Tukey HSD for comparing*

| (I) Group | (J) Group | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval | |
|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound |
| Knowledge | skill | .35* | .050 | .000 | .23 | .46 |
| | principles | .69* | .050 | .000 | .55 | .79 |
| Skill | knowledge | -.35* | .050 | .000 | -.46 | -.23 |
| | principles | .32* | .050 | .000 | .20 | .44 |
| Principles | knowledge | -.67* | .050 | .000 | -.79 | -.55 |
| | skill | -.32* | .050 | .000 | -.44 | -.20 |

*. The mean difference is significant at the 0.05 level.

### *4.1.3. Open-ended Questions*

The final section of the questionnaire consisted of three open-ended questions, addressing the practices in the classroom. Participants were requested not to produce extended explanations, consequently, the responses provided could be more efficiently coded and grouped into categories. Categorization was done by two coders to check for the reliability of the coding. The results produced a moderately high and acceptable kappa coefficient of agreement ($k$ = .68).

#### *4.1.3.1. Teaching materials*

The first question explored the type(s) of textbook(s)/material(s) required to be read for the course. A total of six textbooks were listed as the teaching sources. Some respondents utilized more than one book. Self-designed materials were also reported by 8% of the instructors. Two textbooks seemed to dominate the courses. Farhady, Jafarpur, and Birjandi, (1994) was used with a percentage (81), larger than that of all the other textbooks reported. The other frequently used textbook was Heaton (1988), with a percentage of 45. Four other textbooks (i.e., Bachman, 1990; Brown, 2014, Bachman & Palmer, 1996; Hughes, 2003) were utilized, each with a percentage of less than 35.

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

#### *4.1.3.2. Assessment procedures*

As a prevalent approach to assess students, most respondents referred to the use of mid-term and final examinations. For the whole group, 70% administered just formal mid-term and final exams, 4% of the instructors required their students to write a final term project, and 18% used a combination of the two. Practical works and class activities were favored by 20% of the instructors.

#### *4.1.3.3. Problems faced in language assessment courses*

The fourth question sought the kind of problems usually confronted in the course. Class size stood out with a frequency of 62. Lack of time (f = 56), teaching statistics (f = 41) and concepts and terminologies of the field (f = 28), and lack of attention to assessment in the educational system (f = 16) were also reported as serious issues.

### *4.2. Observation*

#### *4.2.1. General information about the course*

Forty-five students in class A and thirty-eight in class B were taking the course. The classes were held once a week and lasted for sixteen sessions. A total of 932 minutes for class A and 1062 minutes for class B were recorded, excluding the greeting time and talks on unrelated topics. The distribution of time was approximately normal in each class, based on the Q-Q plot (Figure 1). Results showed that the mean of time in class A (5.78 ± 6.98) and class B (6.05 ± 6.67) was not significantly different (p-value=0.266). A general outline of both classes is presented in table 5.

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

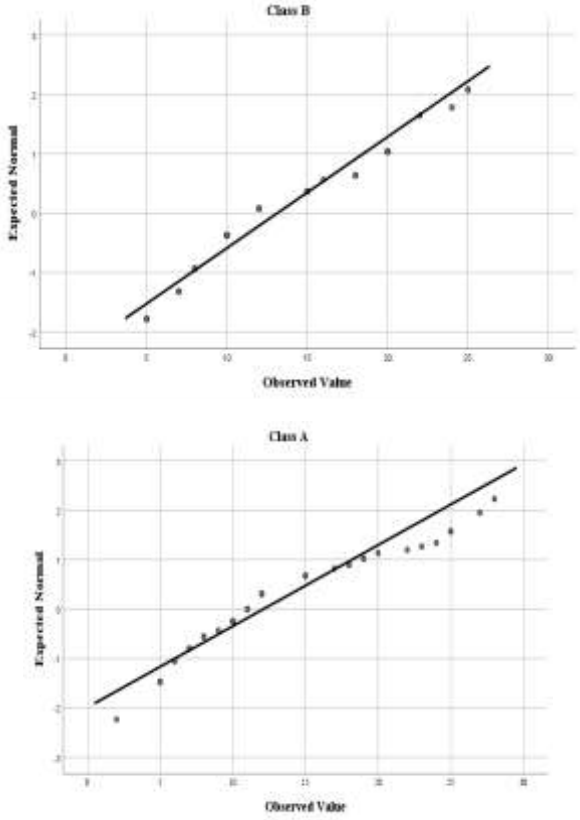



**Figure 1.** The quantile-quantile plot

Table 5. *General information concerning time allocation, class size, etc.*

| class | Time (min) | Class size | Textbooks/ materials | assessment format and scoring |
|---|---|---|---|---|
| A | 932 | 45 | Farhady, et al. (1994), chps: 1-14<br><br>Brown (2014), chps: 1,2,4,5 | Classroom attendance & participation=2 points<br>2 quizzes (paper & pencil multiple-choice format) = 3 points<br>Mid-term exam (paper & pencil open-ended items) = 5 points<br>Final exam (paper & pencil open-ended & multiple-choice items = 10 points |
| B | 1062 | 38 | Farhady, et al. (1994)<br><br>Chps:1-15 | Classroom attendance & participation=3 points<br>Mid-term exam (paper & pencil open-ended & multiple-choice items) = 6 points<br>Final exam (paper & pencil open-ended & multiple-choice items) = 11points |

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

### *4.2.2. Component Focus*

One purpose of the study was to investigate the extent to which each component of LAL was focused. Descriptive statistics for time in each component indicated mean differences for both classes (Figure 2).

Class A Class B

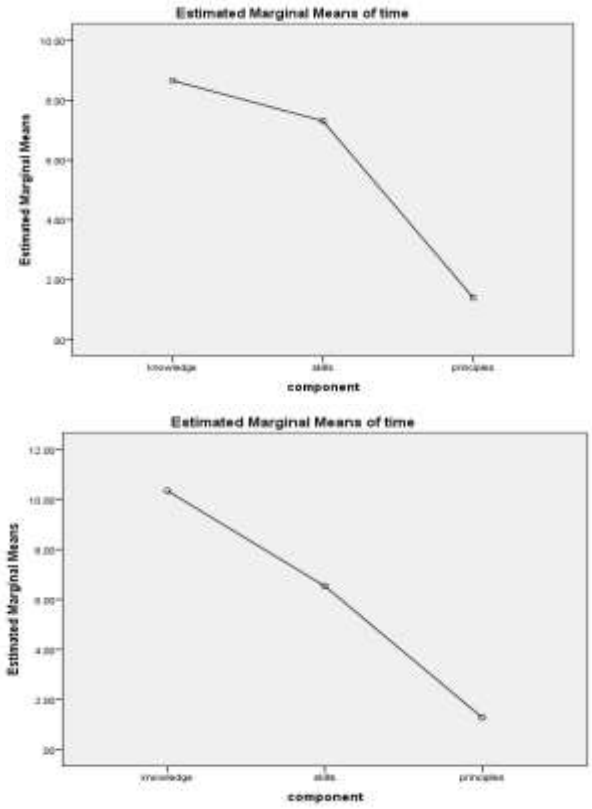



**Figure 2.** Mean plot for component by class

In the next step, univariate analysis and multiple comparisons of components were performed. For both classes, the time values in knowledge vs. principles and skills vs. principles were significantly higher. No significant differences were observed concerning knowledge vs. skills components (Table 6).

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

Table 6. *Multiple Comparisons of component by class*

Dependent Variable: time

Tukey HSD

| class | (I) component | (J) component | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval Lower Bound | 95% Confidence Interval Upper Bound |
|---|---|---|---|---|---|---|---|
| class A | Knowledge | skills | 1.346 | 2.324 | .832 | -4.300 | 6.993 |
| | | principles | 7.253* | 2.324 | .009 | 1.606 | 12.900 |
| | Skills | knowledge | -1.346 | 2.324 | .832 | -6.993 | 4.300 |
| | | principles | 5.906* | 2.324 | .039 | .259 | 11.553 |
| | Principles | knowledge | -7.253* | 2.324 | .009 | -12.900 | -1.606 |
| | | skills | -5.906* | 2.324 | .039 | -11.553 | -.2596 |
| class B | Knowledge | skills | 3.813 | 2.060 | .071 | -.343 | 7.970 |
| | | principles | 9.067* | 2.060 | .000 | 4.910 | 13.223 |
| | Skills | knowledge | -3.813 | 2.060 | .071 | -7.970 | .343 |
| | | principles | 5.253* | 2.060 | .014 | 1.097 | 9.410 |
| | Principles | knowledge | -9.067* | 2.060 | .000 | -13.223 | -4.910 |
| | | skills | -5.253* | 2.060 | .014 | -9.410 | -1.097 |

*. The mean difference is significant at the 0.05 level.

### *4.2.3. Assessment Approaches and practices*

The purpose of this section was to investigate the assessment practices of the participating instructors. The main assessment practices were paper-and-pencil mid-term and final tests in open-ended and/or multiple-choice formats. Concerning the assessment practices employed by the instructors throughout the instruction, a set of variables were taken into consideration. Generally, the mean for class A was significantly higher than that in class B (mean difference = 1.00 ± 0.43, p-value=0.029) (Table 7). However, significant differences were limited just to two of the variables under consideration (checking students' understanding throughout the instruction and providing feedbacks based on the assessments

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

done). No significant differences were observed for other assessment-related aspects (Table 8).

Table 7. *Comparing the mean of assessment approaches between classes using independent t-test*

| Dimension | Class | N | Square Mean ± SD | Mean Difference ± Std. Error | 95% CI of mean difference | p-value |
|---|---|---|---|---|---|---|
| Assessment Approach | A | 15 | 3.53 ± 1.30 | 1.00 ± 0.43 | 0.11, 1.89 | 0.029 |
| | B | 15 | 2.53 ± 1.06 | | | |

Table 8: Comparing the distribution of variables related to assessment practices between classes

| Variables | Class | Mean ± SD | Median (Q1, Q3) | Min, Max | P-value* |
|---|---|---|---|---|---|
| Tasks are assigned effectively to reinforce | class | 0.93 ± 0.70 | 1.00 (0.00, 1.00) | 0, 1 | 0.481 |
| | class | 0.40 ± 0.51 | 0.00 (0.00, 1.00) | 0, 1 | |
| The instructor implemented diagnostic assessment | class | 0.60 ± 0.51 | 1.00 (0.00, 1.00) | 0, 1 | 0.269 |
| | class B | 0.53 ± 0.52 | 1.00 (0.00, 1.00) | 0, 1 | |
| Students' understanding is checked throughout the | class | 1.27 ± 0.56 | 1.00 (1.00, 2.00) | 0, 2 | 0.026 |
| | class B | 0.87 ± 0.74 | 1.00 (0.00, 1.00) | 0, 2 | |
| The instructor provides feedback based on the assessment he/she does | class | 1.40 ± 0.74 | 1.00 (1.00, 2.00) | 0, 2 | 0.038 |
| | class B | 0.73 ± 0.68 | 1.00 (0.00, 1.00) | 0, 2 | |
| The instructor involved alternative assessment (self/ peer/ whole class, | class | 0.60 ± 0.51 | 1.00 (0.00, 1.00) | 0, 1 | 0.153 |
| | class B | 0.33 ± 0.49 | 0.00 (0.00, 1.00) | 0, 1 | |
| The instructor employed technology | class | 0.00 ± 0.00 | 0.00 (0.00, 0.00) | 0, 0 | 1.000 |
| | class B | 0.00 ± 0.00 | 0.00 (0.00, 0.00) | 0, 0 | |

* The exact Man-Whitney test

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

One critical aim of this section was to explore the variety of assessment and feedback modes used by the instructors. Lack of diversity was observable in both classes. Based on the findings, instructors mostly embedded teacher-oriented assessment techniques into their teaching practices. The main assessment mode participating instructors used was oral questioning. In class A, oral questions were asked more frequently and as expected, the same result was observed for the provision of feedback. The use of alternative types of assessments and diagnostic assessment was lacking in both classes. Besides, designing tasks to have students practice in different aspects of assessment and applying technology, considered as an essential aspect of modern assessment, were neglected by both instructors.

## 5. Discussion

The findings of this research provided insights into three serious concerns. The issues raised bear some connection to the probable deficiencies of these courses in developing students' LAL in the studied context. These concerns are discussed below.

### *5.1. Lack of focus on certain aspects of LAL*

The first concern, which addresses the first and third research questions, points to the lack of interest on the part of the course instructors to focus on specific language assessment issues. Generally, the descriptive statistical analysis of the data revealed that for the undergraduate language assessment courses studied here, LAL is mostly a matter of knowledge and theory and to some extent, skills, with little importance given to principles. However, as Davies (2008) points out, language teachers are required to possess the knowledge, skills, and, of course, the principles of language assessment. Observation data also indicated that although the two instructors showed differences in certain aspects of their

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

pedagogy, they seemed to agree on the *what* of teaching, with both concentrating more on knowledge and skills, respectively. Instructors' choice of the textbook(s)/ teaching materials also revealed their unified instruction and their inclination toward teaching the same areas in assessment. The main textbook used by the instructors, based on data from both the questionnaire and observations, was Farhady, et al., (1994), which deals primarily with knowledge and skills of language assessment, devoting little space to assessment principles. Although the instructor of class A also used a more renowned, comprehensive textbook- Brown (2014)- he limited its use to four chapters, which still dealt mainly with knowledge and skill components. Lack of variation in the textbooks employed and the limitation of topics were noticeable in the data obtained through both the questionnaire and observation, despite the availability of many textbooks written by significant scholars in the field. Instructors might select such textbooks as they might believe that they cover what they think is of primary importance in assessment.

Interesting findings were also observed with respect to each component. Based on data from the questionnaire, regarding the knowledge dimension, trivial attention was devoted to providing learners with information about new assessment approaches such as communicative and performance-based assessment, diagnostic assessment, formative assessment, and computer-adaptive testing. In fact, the most significant interest was displayed towards providing students with information on discrete-point approach, objective tests, summative assessment, norm-referenced tests, and large-scale testing. With respect to the skills component, participating instructors showed little interest in alternative assessment procedures, practice in test administration and scoring and assessing speaking, writing, and listening skills. More importantly, instructors did not focus sufficiently on critiquing tests/assessments, in spite of emphasis on developing critical thinking skills to evaluate assessment practices (Scarino, 2013; Vogt & Tsagari, 2014). For example, Vogt and Tsagari (2014, p. 391) state that "the lack

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

of ability to evaluate tests critically represents a risk for the teachers to take over tests unquestioningly without considering their quality." In the case of testing principles, inadequate attention was paid to the ethical, societal, and cultural dimensions of assessment, as well as issues of fairness and incorporating multiple sources of evidence to make decisions, authenticity, learner autonomy, and importance of injecting new perspectives into assessment practices.

Overall, based on these findings, our central speculation is that students probably cannot have an adequate LAL level in certain dimensions/aspects.

## *5.2. Dominance of traditional assessment approaches*

One purpose of the research (research question2) was to explore the instructors' assessment practices. The findings also provided another response to the third question by pointing out another probable deficiency of the courses in developing LAL, which is the instructors' adherence to traditional assessment practices. In our context, instructors mainly seemed to treat summative tests and the end-products as the norm for assessment. Data showed that traditional approaches, including teacher-led assessment activities and focus on paper-and-pencil tests were dominant in the courses, and these instructors did take into account changes that favor formative assessment. The same results were obtained through both the survey and observations. However, it is very vital that teacher educators model acceptable practices in assessment throughout the program. As it is echoed in the literature**,** when teachers have not undergone adequate training on how to assess students' learning efficiently, they begin to assess their learners as they were assessed throughout their education (Tsagari & Vogt, 2014). Consequently, EFL teacher candidates will focus on the course instructors' assessment practices and shape their literacy in language assessment by their own assessment experiences in teacher education programs.

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

### *5.3. Lack of practice in assessment*

Another issue, which still points out another deficiency of the courses (research question 3), concerns the lack of practice in assessment which otherwise could aid development of sound competence in assessment. As mentioned earlier, for teachers to enter their classrooms with the knowledge and the confidence required, pre-service training programs should be provided. However, as Malone (2017) asserts, mere training is inadequate for teacher candidates to respond to the language assessment needs, emphasizing that such training should include the "content for language instructors to apply what they have learned in the classroom and understand the available resources to supplement their formal training when they enter the classroom" (p. 235). It can be surmised that a course with a specific focus on practical tasks on the assessment of students' learning can contribute to teacher candidates' assessment literacy and, hence, is pivotal in teacher training programs.

Based on the results of the present study, the course instructors focused on talking about different aspects of assessment rather than having practice in assessment; that is, different LAL components were built on the theoretical ground. The courses seemed to be textbook-centered, as the critical determiner of selecting the "what" of teaching. Nevertheless, it should be mentioned that instructors' problems, such as short duration of the courses and large classes might have, to some extent, hindered the application of practical tasks by the instructors. However, we would advocate that instruction be reinforced by practicum and experience of assessing, as narrowing the gap between theory and practice should be an objective of an assessment course design.

### *5.4. Possible reasons for the course deficiencies*

Due to the inadequacies discussed, it can be concluded that these courses did not fulfill the expectations of being comprehensive and up-to-date. A set of factors may account for these deficiencies. A primary reason might be that the construct

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

of LAL has not been clearly defined by language testing scholars (Inbar-Lurie, 2008; Karagiorgi & Petridou, 2020; Malone, 2017). As Karagiorgi and Petridou (2020, p. 140) noted, due to the "lack of consensus within the professional testing community as to what constitutes assessment knowledge, defining it presents a major challenge." A secondary reason might be related to the course instructors' conceptualization of LAL. We cannot claim that language testing instructors have weak knowledge of LAL or have failed to update their knowledge in testing and assessment; however, it might be assumed that they believe that the stock of competencies needed for undergraduate English students is limited to the practical *know- what* and *know-how* of language testing. Their reliance on few outdated language testing textbooks can also provide evidence of what course instructors think shapes LAL for language students at the B.A. level. A third reason might be instructors' lack of willingness to adopt innovative methods. Instructors might avoid the complexities of new assessment forms because of feasibility issues. A fourth reason can be a lack of assessment policies and professional standards to guarantee the quality of teacher education programs. There is also an absence of LAL standards required as a part of essential competency in teacher recruitment in Iran.

## 6. Conclusion

The role of language assessment courses in EFL/ESL students' development of LAL is of high importance. This study aimed to examine the extent to which such courses at Iranian state universities help equip future teachers with the essentials of language assessment to assess learners effectively and accurately. Although this research did not establish that students lack LAL, findings raised questions about the quality of their LAL. It seemed that these courses primarily concentrated on teaching *what* and *how* of assessment, giving little importance to teaching assessment principles. Besides, instructors' classroom assessment

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

practices were not based on new trends in assessment and testing, and the course also lacked a balance between theoretical input and practical works.

In sum, the language assessment courses studied here need a radical shift, as called for by Brown and Bailey (2008). We suggest that the course teaching content and materials, practical activities, and assessment strategies be revised. A new syllabus along with updated resources encompassing all dimensions of assessment should be taken into consideration. Providing learners with both a sound theoretical ground and well-designed practical activities should be highlighted. Therefore, a more interactive and collaborative style of instruction is recommended. Discussions, workshops, and teamwork are suggested as students usually enjoy learning interactively and collaboratively and value the opportunity to discuss issues and perform practical tasks. They can write items together, critique each other's or other available tests, help each other proofread and revise the items/tests, etc. Most importantly, as mentioned, quality assessment course, in itself, entails the use of valid assessment procedures. Furthermore, teacher education programs must equip course instructors with adequate time and facilities to increase students' LAL. It is also much better that such training not be reduced to a single course at the undergraduate educational programs.

The present research is one of the few studies into LAL in Iranian EFL context and the first empirical study of LAL development via pre-service training courses in Iran. Besides, it is among the rare studies utilizing a combination of both the quantitative and qualitative methods in evaluating teachers' LAL development. The research has made a contribution to the development of two useful research instruments suitable for research in the field and evaluating language assessment courses in other contexts. Results can be beneficial to course instructors as they may disclose the strong and weak aspects of the courses and help them evaluate their preferences and practices. The findings can also furnish significant implications for policy makers to better understand the nature of language assessment courses, so that workable strategies can be employed to

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

obviate the problems associated with such courses. In addition, although this research has been conducted in a particular educational field, the results can make a special contribution to the overall understanding of assessment literacy development in other teacher education programs.

Despite the significant points having been described, there are some limitations to the present research which demand consideration. Concerning the survey, while the number of respondents was not small, a greater number could have affected the nature of the data obtained. Concerning the questionnaire itself, although the researcher attempted to include all aspects related to language assessment/testing in the survey designed, there might be elements of LAL that were missed. As far as the observation phase is concerned, one obvious limitation was linked to the use of purposive sampling procedures and sample size. In spite of richness of the data, the number of the courses observed was limited to two classes at two universities within a single province, which restricted the generalizability of its results across state universities in Iran. Hence, these limitations can provide opportunities for further studies. In line with these, future research should investigate the instructor's rationale for their practices at a deep level. Further research should also address more precisely whether students are academically ready to perform various assessment-related activities at the end of the course, and how they perceive LAL.

## 7. References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency. The interface between learning and assessment*. London: Continuum.

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

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

Bachman, L.F., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Brown, J. D. (2014). *Testing in language programs: A comprehensive guide to English language assessment*. New York: McGraw-Hill.

Brown, J.D., & Bailey, K.M. (2008). Language testing courses: What are they in 2007? *Language Testing*, 25(3), 349-383.

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

Dörnyei, Z. (2007). *Research methods in applied linguistics: quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.

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

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

iraldo, F. (2018). Language assessment literacy: Implications for language teachers. *Profile:*
*Issues in Teachers' Professional Development*, 20(1), 179-195.

Heaton, J. B. (1988). *Writing English language tests*. London: Longman.

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

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

Joint Committee on Standards for Educational Evaluation. (2015). Classroom assessment standards. Thousand Oaks, CA: Corwin.

Karagiorgi, Y. & Petridou. A. (2020). National literacy assessment and language assessment literacy. In: D. Tsagari (ed), *Language Assessment Literacy: From theory to practice* (pp.137- 159). Cambridge: Cambridge Scholars Publishing.

Lam, R. (2014). Language assessment training in Hong Kong: Implications for language assessment literacy, *Language Testing*, 1- 29.

Lukin, L. E., Bandalos, D. L., & Eckhout, T. J. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice*, 23(2), 26-32.

Marhaeni, A., Padmadewi, N., Tantra. D, Ratminingsih, N., Dewi, N., & Yudha Paramartha, A. (2020). English Teachers' Assessment literacy in Bali seen from teacher's professional development, teacher's service experience, and teacher's educational qualification. *Asian EFL Journal*, 27 (4.5), 56-65.

Malone, M.E. (2017). Training in language assessment. In: Shohamy E., Or., May S.(eds) *Language Testing and Assessment, Encyclopedia of Language and Education* (3rd ed,) (pp. 225-240). Cham: Springer.

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

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

Price, M. (2005). Assessment standards: The role of communities of practice and the scholarship of assessment. *Assessment & Evaluation in Higher Education 30*(3), 215-230.

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.

Shim, K. N. (2009). *An investigation into teachers' perceptions of classroom-based assessment of English as a foreign language in Korean primary education.* Unpublished doctoral dissertation. Exeter: University of Exeter.

Sikka, A., Nath, J. L., & Cohen, M. D. (2007). Practicing teachers' beliefs and uses of assessment. *International Journal of Case Method Research & Application*, 3, 240-253.

Stiggins, R. J. (1991). Assessment Literacy. *The Phi Delta Kappan*, 72(7), 534-539.

Vogt, K. & Tsagari, D. (2014). Assessment Literacy of Foreign Language Teachers: Findings of a European Study. *Language Assessment Quarterly, 11*(4), 374-402.

**Acknowledgements**

The authors would like to express their sincere gratitude to Prof. Mahmood Reza Atai for his academic support and encouragement and the two anonymous reviewers for their valuable comments made on the earlier version of this article.

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

**Appendix A (questionnaire)**

Greeting! This questionnaire is part of a project conducted for my Ph.D. studies. I am trying to describe what is going on in undergraduate language assessment courses at State universities. I would appreciate if you answer the questionnaire. Thank you for dedicating your precious time.

**Part I:**

Please provide the following information about the course you are teaching.

Length of the course.....

Size of the class.....

**Part II: Teaching content**

Please indicate the amount of time you spend on each of the following topics.

| | None | a little | Some | Moderate | Extensive |
|---|---|---|---|---|---|
| **basic concepts in testing (tests, measurement, evaluation)** | | | | | |
| **pre-scientific stage** | | | | | |
| **psychometric-structuralist stage** | | | | | |
| **sociolinguistic-pragmatic stage** | | | | | |
| **discrete-point approach to testing** | | | | | |
| **integrative approach to testing** | | | | | |
| **communicative approach to testing** | | | | | |
| **performance-based approach to testing** | | | | | |
| **aptitude test** | | | | | |
| **diagnostic test** | | | | | |
| **placement test** | | | | | |
| **proficiency test** | | | | | |
| **achievement tests** | | | | | |
| **selection tests** | | | | | |
| **objective tests** | | | | | |
| **subjective test** | | | | | |

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

**norm-referenced test**

**criterion-referenced test**

**Summative assessment**

**Formative assessment**

**alternative assessment**

**large-scale test**

**classroom test**

**high-stakes test**

**low-stakes test**

**computer-adapted tests and internet-based tests**

**construction of test syllabuses/item specifications**

**principles and practice of item writing**

**developing and using selected-response items**

**(true/false, multiple-choice, matching)**

**developing and using constructed-response items**

**(fill in the blank, short answer, etc)**

**developing and using alternative assessment procedures**

**(checklists, videotapes, audiotapes, journals, peer-assessment,**

**self-assessment, portfolios, etc.)**

**testing listening comprehension**

**testing reading comprehension**

**testing speaking**

**testing writing**

**testing vocabulary**

**testing grammar and structure**

**conducting item analysis (item facility,**

**discrimination index, distractor efficiency)**

**test administration**

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

**test scoring**

**developing and using scoring rubrics**

**administering and scoring computer-adapted**

**testing and internet-based testing**

**using different types of interpretations (norm-referenced**

**and criterion-reference)**

**test critique**

**test paper revision**

**practicing assessment at different levels**

**internal consistency**

**test-retest reliability**

**parallel-form consistency**

**split-half reliability**

**K_R_20**

**K_R_21**

**intra-rater reliability**

**inter-rater reliability**

**content validity**

**construct validity**

**face validity**

**criterion-related validity**

**concurrent-related validity**

**predictive validity**

**doing descriptive data analysis using SPSS**

**doing inferential data analysis using SPSS**

**standard setting in language testing**

**ethical practices in scoring/ test administration**

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

**ethical practices in scoring**

**ethical practices in test administration**

**ethical practices in test scores interpretation**

**making sound decisions based on test/assessment results**

**considerations of test reliability**

**considerations of test validity**

**considerations of test washback (on teaching and learning)**

**considerations of test impact (on society)**

**considerations of fairness and bias in assessment**

**considerations of test reliability**

**doing assessment based on multiple sources of evidence**

**language program evaluation**

**importance of incorporating formative assessment**

**importance of incorporating alternative assessment**

**importance of incorporating learner autonomy assessment**

**consideration of the influence of societal values on testing practices**

**consideration of cultural aspects in assessment/testing**

**critical approaches to language testing**

**testing in relationship to curriculum**

**Part III**

Please answer briefly the following questions about the course you are teaching.

1. What textbook(s) / materials do you require your students to read for this course?

2. Explain briefly your method of assessing students' learning.

3. What problems do you usually face in your assessment classes?

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

**Thank you very much**

**Appendix B (Observation Scheme)**

Class No :…….

Date of Observation :…….. Session No:…….

Start Time:…….. Ending Time:……. General Topic:……

Textbooks/materials:……

| Sub-topic | **LAL component / content focused** | | | **Notes** |
|---|---|---|---|---|
| | Knowledge | Skills | principles | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

**Evaluation Methods**

| Measure | extensive | some | A little | none | Notes |
|---|---|---|---|---|---|
| Tasks are assigned effectively to reinforce and extend learning | | | | | |
| The instructor implemented diagnostic assessment at the beginning or end to adjust subsequent instruction | | | | | |
| Students' understanding is checked throughout the instruction | | | | | |
| The instructor provides feedback based on the assessment he/she does | | | | | |
| The instructor involved alternative assessment (self/ peer/ whole class, portfolios, etc., in evaluation process | | | | | |
| The instructor employed technology in assessing students' learning | | | | | |

***Notes on Contributors:***

**Masoomeh Taghizadeh** is a PhD candidate of TEFL at Yazd University, Yazd, Iran. Her research interests include language testing and assessment, teacher education, second language acquisition, syllabus design, and materials development.

**Golnar Mazdayasna** is an Associate professor of Applied Linguistics at Yazd University, Yazd, Iran. Her research interests include English for Specific Purposes, materials designing, teacher language awareness, and writing skills. She

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]

has published articles in national and international journals. Correspondingly, she has participated and presented several papers at national and international conferences.

**Fatemeh Mahdavirad** is an assistant professor at the English Language and Literature Department of Yazd University. Her research interests include discourse analysis, text linguistics, task-based language teaching, critical discourse analysis, language acquisition, syllabus design, and materials development.

[ Downloaded from c4i2016.khu.ac.ir on 2026-11-04 ]