Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 13, No. 1, March 2023
ISSN 2476-5880

Development and Validation of a Scenario-Based Teacher Language Assessment Literacy Test

Mohammad Reza Anani Sarab[1], Simindokht Rahmani[2]*

Received: 9 August 2022 Accepted: 26 September 2022

Abstract
Language testing and assessment have grown in popularity and gained significance in the last few decades, and there is a rising need for assessment literate stakeholders in the field of language education. As teachers play a major role in assessing students, there is a need to make sure they have the right level of assessment knowledge and skills to accomplish their duties as assessors. The present study sought to develop and validate an assessment literacy test. To this end, a thirty-five scenario-based Language Assessment Literacy Test (LALT) was developed based on the seven standards covered by the Standards for Teacher Competence in Educational Assessment of Students (1990). Construct validity of the test was investigated by collecting data from 168 Iranian EFL teachers. To investigate the validity of the measure, tests of reliability, item difficulty, and item discrimination were carried out. The test was then subjected to EFA (Exploratory Factor Analysis whose results showed a seven-factor solution for the test items. The implications of the study for EFL teachers, language testers, teacher trainers, and curriculum developers are discussed.

Keywords: assessment; Exploratory Factor Analysis (EFA); language assessment literacy; teachers' language assessment literacy

1. Introduction

During the past decades, the use of assessment and tests has grown not only to improve learning but also to promote accountability in different contexts (Brookhart, 2011; Coombe et al., 2020; Fulcher, 2012; Stiggins, 2010; Taylor, 2009). Due to globalization, educational systems worldwide keep being upgraded to be in line with new educational approaches and meet the changing needs of modern society. Since assessment is a central piece within these systems, there is a need to ensure that the local assessment standards reflect the internationally accepted ones, and the demand for more professionals with adequate levels of literacy in the field of assessment has increased (Taylor, 2013). Globalization has also resulted in mass immigration. With more people immigrating to other parts of the world for various reasons, such as finding better job opportunities or seeking better education, there is a growing demand for assessment and assessment literate professionals to respond to the need for evaluating immigrants' readiness to

[1] Shahid Beheshti University, Tehran, Iran, Email: anani@sbu.ac.ir
[2] Shahid Beheshti University, Tehran, Iran. (corresponding author), Email: simindokht.rahmani@gmail.com

integrate as citizens and skilled workforce. Language tests are both a vital component of citizenship tests and a requirement for education and employment opportunities. As a result, designing and administering fair language tests and having the necessary knowledge and skills in this area seem to have gained more importance than ever (Fulcher, 2012; Kunnan, 2019; Shohamy & McNamara, 2009; Taylor, 2009).

Traditionally defined as knowledge and skills different assessors should have about assessment (Stiggins, 1991, 1995), assessment literacy (AL) was alluded to for the first time in a seminal document known as *Standards for Teacher Competence in Educational Assessment of Students* (1990). The document comprises guidelines prescribing seven competency domains teachers are required to be skilled at to make judicious assessment decisions in and outside their classrooms. The competency domains can be classified into two main strands, with the first one focusing on instruction and learning and the second one dealing with the uses of tests and test results. The first strand includes choosing and developing assessment methods appropriate for instructional decisions (American Federation of Teachers et al., 1990). The second strand covers areas such as administering, scoring, and interpreting the results of both external and teacher-produced assessment methods and using the results for assessing students and evaluating instruction and school curricula; developing valid student grading procedures and communicating the results to the relevant stakeholders; and finally, being able to recognize unethical, illegal, and other inappropriate methods and uses of assessment information (AFT et al., 1990).

There are different stakeholders such as policymakers, assessment professionals, teachers, university registrars, parents, and students who need to be literate in the field of assessment (Kremmel & Harding, 2020; O'Loughlin, 2013; Pill & Harding, 2013; Taylor, 2013). Among the key players, teachers are considered the centerpiece of quality education and, therefore, have received the most attention. Assessment literate teachers can draw correct conclusions about student learning, share that information with students and other stakeholders, and modify instruction as necessary, while teachers with insufficient AL are more likely to make misdirected and ill-informed educational decisions due to their insensitivity to reliability and validity issues in assessment (Xu & Brown, 2017). The research on language assessment literacy (LAL) has mainly focused on teachers, emphasizing teachers' practices, beliefs, and needs. The findings of these studies demonstrate that many teachers do not feel sufficiently prepared and think they need more assistance with practical issues like how to use assessment concepts and techniques and make assessment-related decisions, despite the significance of student performance assessment as one of the primary roles of classroom teachers. (Ahmadi et al., 2022; Fulcher, 2012; Mertler & Campbell, 2005; Vogt & Tsagari, 2014; Yastıbaş & Takkaç, 2018). Thus, understanding and measuring teachers' current levels of assessment literacy is critical to promoting research on both AL and the development of AL training modules for teachers (Gardner & Rea-Dickins, 2001; Xu & Brown, 2017).

1.1. Language Assessment Literacy

The extension of assessment literacy (AL) to the field of language testing is called Language Assessment Literacy (LAL) (Inbar-Lourie, 2008a). In an oft-cited definition of the term, Inbar-Lourie (2008a) distinguishes LAL from AL on the basis of the addition of language-specific competencies which make it a multifaceted and complex construct. This intricacy stems from the particular difficulties in measuring "linguistic skills, knowledge, and communicative competence" (Harding & Kremmel, 2016, p. 414). That is why providing a clear-cut definition has proved to be controversial (Giraldo, 2021a; Stabler-Havener, 2018). The boundaries of the concept of AL, in general, and LAL, in particular, are still debated by Scholars in the field (e.g., Fulcher, 2012; Inbar-Lourie, 2013a; Jeong, 2013; Malone, 2013; Scarino, 2013; Taylor, 2013). LAL is conceptualized in the form of a five-component model by Brindley (2001), a skills, knowledge, and principles model by Davies (2008) and Inbar-Lourie (2008a), a practices, principles, and contexts model developed by Fulcher (2012), and LAL stakeholder profile model developed by Pill and Harding (2013) and Taylor (2013).

Brindley (2001) is the first language tester who takes the standards as a starting point to propose the components of a teacher professional development program for assessment. The model is modular in the sense that, depending on the differing needs of the professionals, three components are considered core and two optional. From these components, one can infer the core elements of language assessment literacy or the type of assessment competence teachers need to develop. The first core unit is called "the social context of assessment." This core unit is intended to help teachers understand the role and purpose of assessment in their context. The second core unit, labeled as "defining and describing proficiency," is intended to help teachers understand the theoretical basis of testing and assessment as a basis for their ability to evaluate assessment procedures. The third ancillary unit, labeled as "constructing and evaluating tests," aims to enable teachers to construct valid and reliable assessments and evaluate them critically. The next ancillary unit which deals with assessment in the language curriculum helps teachers develop an understanding of the concept of objectives and benchmarks as well as the way criterion-referenced procedures can be applied in assessing progress and achievement. The fifth unit which deals with putting assessment into practice is not relevant to LAL competencies. According to Brindley (2001), depending on the nature and extent of an individual's involvement in assessment, each person might require a different level of LAL.

Next, Davies (2008) brings us one step closer to the abstract competence level of LAL. His three-dimensional model distills the topical components of AL into knowledge, skills, and principles. Knowledge provides the background for setting the context of assessment, understanding models of language description, evaluating tests, and interpreting the assessment results. Skills cover the practice of language assessment, including item writing and the use of statistics and item analysis. Principles refer to the ethics of test use or the responsibilities of test developers and users in making ethical choices. In his view, the first two components are directly related to assessing language, while principles relating to standards of practice, fairness, ethics, and assessment outcomes can universally be applied to both AL and LAL. As for the emphasis

these components have received in language testing textbooks, research by Bailey and Brown (1995) and by Brown and Bailey (2008) showed that more emphasis had been put on the first two components. The same trend was observed by Jeong (2013) and Jin (2010) who studied language testing courses and found that the main focus was on knowledge and skills, not principles and consequences of assessment. Despite the varying degrees of emphasis put on the three components, they have consistently been discussed in theoretical discussions regarding LAL (Davies, 2008; Fulcher, 2012).

Taking a broader view, Inbar-Lourie (2008) investigates the components of AL in general education and moves on to the possible implications for LAL. Her approach to AL is socio-constructivist in the sense that the knowledge base of AL should include both testing and assessment for learning. In this sense, AL assumes social aspects which move beyond the technicalities of test development and use. Because of its social aspects, AL would apply to different groups of stakeholders with different needs. She categorizes the five components introduced by Brindley (2001) into three dimensions of "what, how, and why." According to Inbar-Lourie, knowing the "what" (which is about the second core unit proposed by Brindley) together with operating the "how" (which includes the third, fourth, and fifth units) are the prerequisites to understanding the "why" (which comprises of the first core unit). She considers the fourth component of Brindley's model (i.e., assessment in the language curriculum) as something which goes beyond the purpose, content, and techniques of language testing and assessment to deal with language assessment reforms and initiatives. Although Brindley intended this component to be of use to language professionals in specific contexts, Inbare-Lourie viewed it as applicable to teachers as well. Inbar-Lourie's classification of the core units also corresponds with Davies' (2008) three-dimensional model.

Fulcher (2012) proposes a three-tier hierarchical LAL model based on Davies' (2008) topical components and attempts to offer an expanded definition of LAL. As the term "hierarchical" suggests, Fulcher's model comprises three layers. The bottom layer (practices) includes the knowledge, skills, and abilities needed to create and evaluate various tests. The next layer (principles) guides practice to be ethical and fair. The top layer (context) is about putting the first two layers in historical, philosophical, social, and political frameworks to evaluate the impact and role of testing clearly.

Later, Pill and Harding (2013) underscore the differential assessment needs of stakeholders. They borrowed a model of proficiency from the fields of Science and Mathematics to estimate the levels of LAL among non-practitioners. Their approach shows a shift from LAL as a competence model to LAL as a proficiency model with five levels. They propose 5 levels of LAL, ranging from "illiteracy" to "normal literacy", "functional literacy", "procedural and conceptual literacy", and "multidimensional literacy" (p.383). Stabler-Havener, (2018) commend Pill and Harding (2013) for their creativity in using scientific literacy models to explore LAL, while Harding and Kremmel (2016) take issue with this model's overemphasis on procedural and theoretical knowledge and its lack of due attention to social, political, and ethical competencies.

They also refer to its limitation of failing to determine the appropriate level of LAL for various stakeholders.

Taylor's (2013) model of LAL, which is based on discussions of AL/LAL by Jeong (2013), Malone (2013), O'Loughlin (2013), Pill and Harding (2013), and Scarino (2013) (Bøhn & Tsagari, 2021), encompasses eight dimensions of knowledge of theory, technical skills, principles and concepts, language pedagogy, sociocultural values, local practices, personal beliefs/attitudes, and scores and decision making. The most important feature of this model, which was absent in other models, is the inclusion of teachers' personal beliefs/attitudes in the model and the importance of this facet is due to the fact that the extent to which teachers are willing to implement new educational policies is influenced by their attitudes and beliefs about assessment (Stabler-Havener, 2018). Moreover, this model differs from other models in differentiating between the LAL needs of different groups of stakeholders. Different LAL profiles were created by Taylor (2013) for different groups of stakeholders, including test writers, university administrators, classroom teachers, and professional testers. However, according to (Kremmel & Harding, 2020), the notable void in these profiles is their hypothetical nature and the fact that they are not accessible to a broader range of stakeholders.

As Kremmel and Harding (2020) put it quite aptly, different models and conceptualizations of AL (e.g., Brindley, 2001; Davies, 2008; Fulcher, 2012; Pill & Harding, 2013) have not provided a diagnostic approach to identifying a variety of profiles. These models are speculative and theoretical, which is why there is no clear understanding of the levels of LAL needed by different stakeholders.

1.2. Language Assessment Literacy Instruments

A number of instruments were developed to measure the LAL of teachers and to investigate their needs. Teacher Language Assessment Literacy Questionnaire (TLALQ) adapted from EALTA is a scale originally designed and developed by Hasselgreen et al. (2004) as part of an EALTA project called ENLTA Activity 5 to conduct a European survey of needs in language testing and assessment (LTA). The original survey was designed to target three types of stakeholders in LTA (language teachers, language teacher trainers, and experts).

Later, a Teachers' Language Assessment Literacy (TLAL) scale was developed and validated by Fulcher (2012). This survey instrument contributed to the field significantly by providing new information regarding assessment literacy and some solutions for solving problems associated with other survey techniques. It was an attempt to investigate the training needs of language teachers in assessment. It consists of four factors: “test design and development”, “validity and reliability”, “large-scale standardized testing”, and “classroom testing and washback” (p.120).

A multiple-choice test, titled The Classroom Assessment Knowledge (CAK), was designed and validated by Tao (2014) to measure EFL teachers' assessment knowledge in a classroom-based assessment context. The test consists of nine subscales, with seven scales covering the Seven standards and two standards added to expand the relatively narrow view of the original standards. The aim was to respond to the criticism in the literature indicating that the seven standards are not

entirely compatible with the new trends in AL, particularly considering classroom assessment (Brookhart, 2011). The nine strands forming the basis of this scale are "(1) choosing appropriate assessment methods, (2) developing assessment, (3) administering, scoring, and interpreting assessment results, (4) developing valid grading procedures, (5) using assessment results for decision making, (6) recognizing unethical assessment practices (7) keeping accurate records of assessment information, (8) ensuring quality management of assessment practices, and (9) communicating assessment results" (p. 108). Out of the 27 multiple-choice items of the scale, 11 were adapted from Mertler and Campbell's (2005) Assessment Literacy Inventory (ALI), and the rest were developed based on an in-depth review of the relevant literature. Findings indicated satisfactory measurement characteristics for the scale. The results also pointed to the low level of Colombian EFL teachers' classroom assessment literacy.

The LAL Survey questionnaire developed by Kremmel and Harding (2020) is a comprehensive instrument which can be used for research, self-assessment, needs analysis, and reflective practice. This measure aims at investigating LAL through an empirical analysis of various stakeholders' needs in the field rather than simply representing the prescriptive views of LAL researchers through theoretical models (Kremmel & Harding 2020). The instrument development went through multiple stages starting in 2015 with focusing on simple definitions and continuing with elaborated definitions, expert review 1, pretest 1 aiming to gather quantitative and qualitative feedback from 62 participants, further refinement to wording, pretest 2 with 25 participants, expert review 2, and the creation of the final version in 2017. Using this instrument, the survey was conducted in several stages targeting stakeholders such as teachers, assessment developers, examiners, researchers, score users, learners, parents, and policymakers.

More recently, Mohammadkhah et al. (2022) developed a self-assessment LAL scale to guage EFL teachers' conceptions of assessment. Drawing upon Xu and Brown's (2016) model and Fulcher's extended definition of LAL, they focused on the contextual factors which impact on teachers' beliefs about assessment. In doing so, they created a grid through reviewing the literature and conducting thematic analysis. The scale was validated by performing Exploratory (EFA) and Confirmatory Factor Analysis (CFA). Results of EFA indicated five factors: (1) assessment in language pedagogy, (2) disciplinary knowledge, (3) social and local values, (4) assessment theories, and (5) personal beliefs and attitudes. The emerging factors of their study are in line with Xu and Brown's (2016) model and the dimensions proposed by Taylor (2013).

In another study, Ölmezer-Öztürk and Aydin (2018) developed a Language Assessment Knowledge Scale (LAKS) including 60 items and four constructs (assessing reading, writing, speaking, and listening) which was intended to measure EFL teachers' knowledge of language assessment. The item pool was based on the most recurrent knowledge elements of the assessment of the four language skills drawn out of language and testing books from 1969 to 2012. The researchers validated their test through four stages: (1) holding individual meetings with ten teachers to ensure the comprehensibility of the items, (2) asking the opinions of 11 testing and assessment experts as to the necessity of the items, and (3) presenting the instrument to 18 practitioners and holding meetings with them to talk about the validity, comprehensibility, and

compatibility of the items, and (4) piloting the final version with a total of 50 teachers. Despite the low factor-loading and item total correlations in the two constructs of assessing speaking and assessing writing- which was reported as a shortcoming by the authors, second-order confirmatory factor analysis findings indicated a generally acceptable model fit.

Farhady and Tavassoli (2018) developed the Language Assessment Knowledge (LAK) test based on the systematic 12-step model put forth by Downing (2006) and Lane et al. (2016). What sets this scale apart from others is its data-driven nature. The test was constructed based on data from EFL teachers In Iran using Fulcher's (2012) needs assessment questionnaire. After revisions were made to the test with the help of experts, it was piloted. While the piloting results revealed an acceptable reliability figure (0.71), there is no mention of any validity check. The final scenario-based scale includes 33 items. The findings of the tests mentioned above revealed that contrary to the results of teachers' needs assessment indicating a high level of language assessment knowledge, most of the teachers had a low level of language assessment knowledge

A Comprehensive review of the literature shows that although there has been considerable research into the construction and validation of LAL scales, most of the measures developed are self-report surveys which fall short of providing an accurate picture of teachers' levels of LAL (Weng & Shen, 2022). Moreover, as was mentioned above, the very few objective tests (e.g., multiple choice or multiple matching) lack optimal psychometric qualities. To add to the current LAL objective tests with improved psychometric properties, the present study aimed to develop and validate a test of LAL to measure the level of EFL language teachers' assessment literacy.

2. The Present Study

Although the literature shows a dearth of AL and LAL instruments which are representative of the more recent changes in assessment (Gotch & French, 2014) and the current measures in the field of language do not reflect the true nature of LAL as they are mostly based on the seven Standards, there is no alternative but to develop an instrument based on the seven standards, for the true nature of LAL, which according to Inbar-Lourie (2017) is differential and context-based, is a future trend in language assessment. The seven standards designate the assessment targets for teachers. How these targets are achieved in practice would inevitably depend on the nature of the subject, which is, in this case, language comprehension and production. In other words, they are general and applicable to all fields within educational assessment. Therefore, if a test developed for teachers teaching content subjects is given to EFL teachers, it is their general assessment knowledge and not their language-related assessment knowledge that will be measured (Ölmezer-Öztürk & Aydin, 2018). That is why we found it imperative to contextualize the items within the framework of LAL to measure LAL of EFL teachers.

To this end, the construct of Language Assessment Literacy (LAL) was inferred from the literature comparing the influential theoretical models of LAL, test specifications were determined, items were developed, and piloting and field testing were carried out to validate the Language Assessment Literacy Test (LALT).

2.1. Construct of LAL

The seven standards have formed the basis for endeavors to specify what teachers need to know and be able to do (later referred to as Assessment Literacy) both in general and language education. Built upon these standards, LAL has evolved through two major phases.

The development of LAL starts with a more componential view of the concept, with Brindley (2001), Inbar-Lourie (2008a), and Davies (2008) proposing various models (Kremmel & Harding, 2020; Wheng & Shen, 2022) with different terminologies which seem to express the same framework and core components. Inbar-Lourie's model is an extension of Brindley's (2001) five-component model, and Davies's (2008) model parallels Inbar-Lourie's (2008). Also, in the three models, due attention is paid to the socio-economic context, ethics, and fairness, which highlights the more recent concerns in assessment in both general education and ELT (English Language Teaching). One main difference between Inbar-Lourie's model and the other two is that she does not simply underscore the language unit in LAL; she brings it to the center and proposes a core language knowledge base for the newly developed concept of LAL.

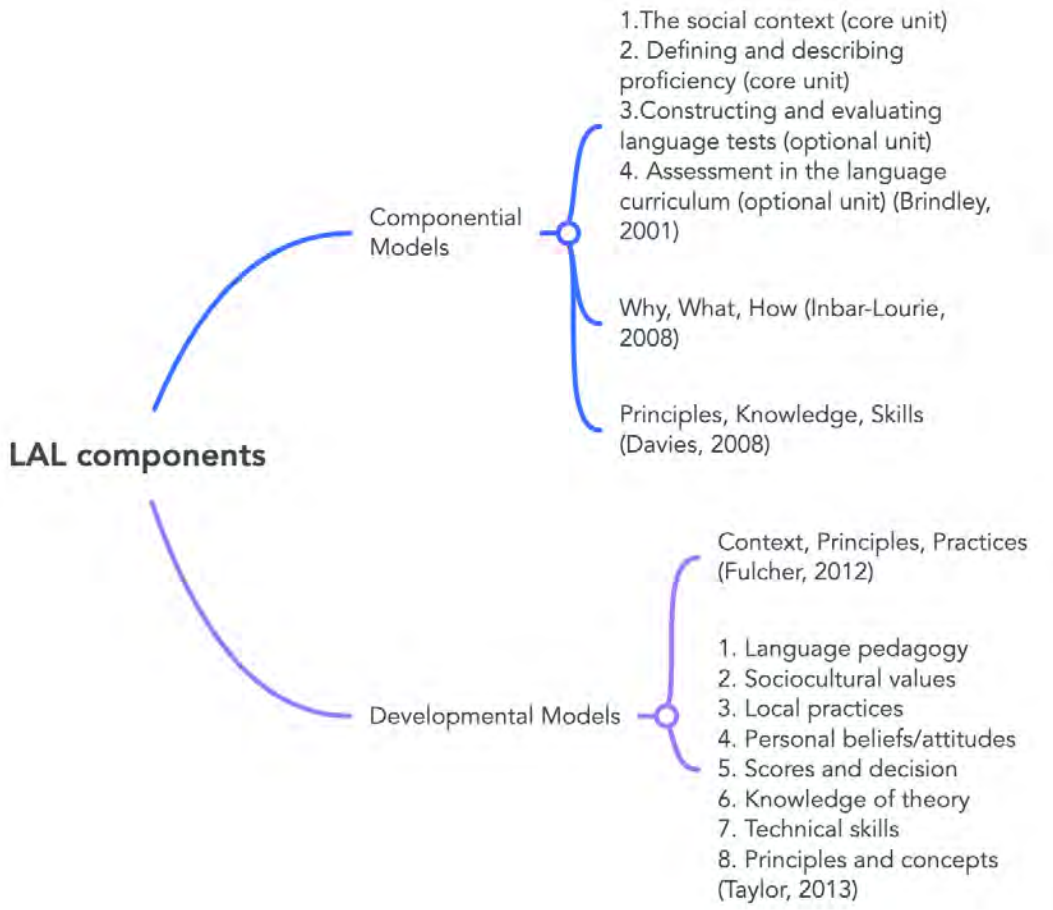
A shift can be seen in the efforts made by the assessment scholars from components to their hierarchical organization with the aim of identifying the levels of AL and the needs of different stakeholders based on their degree and nature of involvement in assessment activities (Kremmel & Harding, 2020). As an example, Davies's pyramidal model consisting of skills, knowledge, and principles as layers organized hierarchically from the bottom to the top shows that the more extended skills component is of more relevance to the everyday practices of teachers whose knowledge base and their awareness of the body of assessment principles would determine the quality of their practice. In a similar way, the three layers of Fulcher's (2008) model include practices, principles, and context. It is worth noting that the "principles" component in Fulcher's model bears no similarity to the "principles" component in Davies' (2008) model. While in the latter model, principles refer to the social role of assessment, in Fulcher's model, they encompass guidance for assessment in practice. Fulcher, who takes a procedural approach to AL, proposes that, in practice, the principles and assessment skills are integrated into the process of developing and delivering language tests and assessments. Although this proposition implies that teachers may draw on all three layers in their assessment practices, Fulcher is careful enough to recognize that "not all of these components will be essential for all stakeholders" (p. 13).

In their conceptualization of LAL, Phil and Harding (2013) are more explicit in recognizing LAL proficiency levels. In moving away from dichotomizing literacy (stakeholders being literate or illiterate), they propose a ladder-like organization for assessment literacy. Taylor (2013) took the literacy ladder proposed by Phil and Harding (2013) and the components of LAL from the literature (e.g., Brindley 2001; Davies, 2004; Fulcher, 2012; Inbar-Lourie, 2008) to shape them into five levels (see Phil & Harding, 2013) and eight dimensions. The model identifies various stakeholders and implies that different stakeholders (depending on their role in testing) need different levels of LAL. Each group of stakeholders falls into one profile which mandates a certain level of LAL in each dimension. Although Taylor's (2013) model seems more comprehensive than the previous models, according to Bøhn and Tsagari (2021), it fails to portray the LAL components

 ISSN 2476-5880

in great detail, which might result in confusion as to what precisely each component entails. For instance, the "scores and decisions" component overlaps with "language pedagogy" since formative assessment can include both decision-making and pedagogy. As another example, the two components of "knowledge of theory" and "principles and concepts" overlap since both theory and concepts can refer to the same notion. While Taylor's model has been used in different studies (e.g., Baker & Riches, 2018; Bøhn and Tsagari, 2021; Yan et al., 2018), the only empirical study which has attempted to operationalize this model is that of Kremmel and Harding (2019). In their study, the results of Exploratory Factor Analysis show a nine-factor model. Figure 1outlines a diagrammatic sketch of the components of LAL in the models discussed in this section.

Figure 1
LAL Models and Components



2.2. Test Specifications

Although the literature on LAL is maturing and scholars have developed different models of LAL (e.g., Brindley, 2001; Davies, 2008, Fulcher, 2012, Inbar-Lourie, 2008, and Taylor, 2013), since the construct is still in its infancy and has a multi-layered nature (Inbar-Lourie, 2013b; 2017), there is no consensus over a particular framework which can conceptualize LAL in a straightforward and detailed manner. Yet, as depicted above, the recurrent models of conceptualization of LAL have focused on knowledge, skills, and principles.

The seven standards which, despite the various models proposed in the literature, remain to be the most widely used framework (DeLuca et al., 2016), were employed as the base model for our test (Appendix A illustrates the seven components and the competencies specified for each). To contextualize the components of AL proposed in the form of seven standards, we made use of the examples of the language-specific elements of the three main components of LAL (knowledge, skills, and principles) provided by Giraldo (2021b) based on a comprehensive review of literature on LAL. The elements related to the "knowledge" component include "(a) models describing language ability, (b) frameworks for doing language assessment, (c) purposes and theoretical concepts, and (d) relevant theories in second language acquisition and language teaching pedagogy" (Giraldo, 2021b; p. 267).

As for the test method, we adhered to the scenario-based multiple-choice item format of the test developed by Mertler and Campbell (2005). There are a number of good reasons why multiple-choice questions were selected to complement the scenarios. Multiple-choice questions have certain advantages like being quick and easy to score, giving the test developer more room to test a broader range of higher-order thinking skills, and covering more content areas that can be answered in a limited time span. In addition, recent research has revealed that the presence of competitive alternatives as answer choices can induce respondents to think deeply about why an alternative is correct and the others (the distractors) are not (Little et al., 2012). Similar to Mertler and Campbell's test (2005), we defined five scenarios featuring EFL teachers in language-classroom contexts and attempted to create an item pool for each scenario based on the seven standards.

With regards to the delivery specification, for practical reasons, the test was planned to be presented in the form of a computer-based test. Since variations in test presentation can cause construct-irrelevant fluctuations in scores (Fulcher, 2011), we ensured to maintain the elements related to the chosen platform, details of test administration, and timing unchanged across all respondents.

2.3. Item Design

Based on the conceptual and application skills that define our test's content domain, a decision was made on the number of items and scenarios. For reliability reasons and the need for extra items deemed necessary, considering that some items might be put aside in the validation process, we decided to develop seven items for each scenario, that is, 49 items in total.

The items were designed based on the content domain specified for each standard. The test was organized in two sections; section one items elicit information from the respondents about their relevant personal features and their background in testing and assessment, and the items in section two present the seven scenarios with multiple-choice items (for a sample of items see Appendix B).

2.4. Pilot One

We started the piloting stage with in-house alpha testing with invited experts (Fulcher & Davidson, 2007). To determine if the content and form of the test are adequate and spot design faults, a group of 10 TEFL Ph.D. candidates and graduates who had been selected through opportunity sampling from accredited universities in Iran were asked to take the test. It is important to note that in Iran, during post-graduate years (both MA and Ph.D.) in the field of TEFL, students pass courses on assessment and testing which are supposed to give them an in-depth understanding of testing and assessment content area. The participants were asked to evaluate the test and recommend improvements regarding the content and form of the test in the Microsoft Word file they were sent through (a) making changes (using tracking and reviewing features of Microsoft Word so that we can trace them) and (b) leaving their comments about the difficulty of the items (i.e., particularly easy/ difficult items), confusing questions (in terms of their wording and choices), the clarity of the scenarios (whether each scenario provides the information the test takers need to answer the questions following it), and the order of the scenarios and questions. We decided to take this stage further by inviting five of the participants to attend a semi-structured interview due to the ambiguities in the feedback provided by them in the written format. To deal with the problems raised at this stage, we removed four of the targeted questions (which left us with 45 items) or made minor or major revisions to some others to make sure that the items are testing the construct. For instance, one participant found one of the questions rather vague as she did not know what the term "the standardized achievement test" used in the question referred to. We made a minor change to the question by adding "referred to above" to highlight the connectedness of the questions in the scenario and remove the ambiguity. As another example, the same participant drew our attention to the "lack of transparency" of another item. To tackle the problem, we changed the focus of the scenario to "writing assessment task" and modified the chocies to make them more relevant the the new focus and in this way make the item more transparent. Table 1 illustrates the changes made to this item.

 ISSN 2476-5880

Table 1
Example of an Item Before and After Being Revised

Version	Example item
Before	Ms. Yegane is planning to design an integrated task to assess her students. She needs to a) involve more than one skill in the task. b) include multiple questions in the task. c) design a task which requires some collaborative work on the part of test takers. d) use pictures, diagrams, or charts in the task.
After	Ms. Yegane is planning to design an integrated skills assessment task for writing. She needs to integrate: **Select ALL the answers you think are correct.** a) reading and writing b) listening and writing c) written skills d) oral skills

2.5. Pilot Two

In the second piloting, we trialed the items with a larger group which was supposed to be as representative of the target population as possible (Fulcher, 2011). Therefore, the test revised based on the results of the alpha testing stage was administered to 30 EFL teachers from different backgrounds (age, gender, years of teaching experience, assessment qualification, and so on). This stage aimed to investigate whether the items were at the appropriate level of difficulty and ensure that they discriminated well through statistical evidence.

The participants of this stage were EFL teachers teaching in language institutes and public schools, undergraduates, MA, and Ph.D. students from different fields, including TEFL, English language and literature, Linguistics, and Translation studies. Based on the results of item analysis, 36 items (80%) stood within the acceptable range (0.3 to 0.7) of the item difficulty index; and 37 items (82%) demonstrated an acceptable level of item discrimination, with 25 items (67.5%) showing a very good level of discrimination power (value of 0.4 and greater), 10 items (20%) indicating a "reasonably good but possibly open to improvement" level (value of 0.3 to 0.39), and two items (5.4%) indicating a "not very good level and in need of some revision" level (value of 0.20 to 0.29). Finally, the Cronbach alpha coefficient for the test was 0.89, indicating a high internal consistency level.

Based on the results, we decided to remove six items (with 39 items remaining). These items were still poor after multiple revisions in item development and alpha testing stages. We also made some modifications to the other items. In one item, for example, two of the distractors

were still dysfunctional since no respondent had picked them. There were also other questions and options we rewrote partly to avoid misunderstanding.

3. Field Testing

3.1. Participants

A group of 168 Iranian EFL teachers teaching in several public schools and private language institutes participated in the validation study. In terms of age, 20.8% of the participants were 18 to 30 years old, 58.9% were 31 to 40 years old, and 20.2% were 41 to 60 years old. With regard to gender, 63.1% of the teachers were female, and 36.9% were male. Regarding their workplace, 64.9 of the respondents worked in private language institutes in Iran, and 35.1% worked in private or public schools. Most participants (72.6) held a Master degree, with Ph.D. holders and undergraduates comprising 10.7% and 16.7% of the teachers, respectively. Concerning their field of study, the great majority had studied English Language Teaching (67.5%), 11.3% Translation Studies, 9.5% Linguistics, 8.3% English Language and Literature, and only 2.9% of the language teachers came from other fields of study

3.2. Data Collection Procedures

To administer the TLALS, 360 Iranian EFL teachers were selected through convenience sampling. The link to the test was shared with the teachers through emails and on social media (Telegram, in particular). The test was given online on the Google form platform. The assumption behind using Google form was that due to the English language of the test and types of questions (more than one option should be selected for some of the questions), as a free platform, Google form was considered a more suitable platform compared to other platforms such as SurveyMonkey, Typeform, and Porsline. A total of 168 teachers took the TLALS which was both constructed and administered in English. The test's link consisted of clear instructions on how to take the test in both participants' native language (Persian) and English.

3.3. Data Analysis

The test items were scored as correct or incorrect; each correct response was given one score, and each false answer was given a zero score. All data were exported to SPSS software (version 26) to conduct the necessary analyses. At first, an item analysis was conducted under the Classical Test Theory to measure item difficulty and item discrimination. Subsequent to performing the item analysis, an exploratory Factor Analysis (EFA) using Mplus software (version 8.3) was run to validate the newly-developed test, and finally, the reliability of the test was measured.

4. Results and Discussion

4.1. Item Difficulty and Item Discrimination

Item difficulty, which is simply defined as the proportion of test takers who answer a question correctly, has an acceptable range of about 0.3 to 0.7 and is given as a *p* value (Henning, 1988).

The item analysis results demonstrated 4 items outside the acceptable range of item difficulty (questions 3, 15, 21, and 35 with the difficulty coefficients of 0.86, 0.89, 0.91, and 0.85, respectively). To determine the degree to which items could discriminate between higher ability and lower ability respondents, item discrimination index was also calculated. Based on the results, 34 out of 39 items showed an acceptable level, with 29 items demonstrating a very good level of discrimination (value of 0.4 and greater), 3 items (1, 13, and 36) displaying a "reasonably good but possibly open to improvement" level (value of 0.3 to 0.39), and 2 items (18 and 31) indicating a "not very good level and in need of some revision" level (value of 0.20 to 0.29). The remaining 5 items (3, 15, 21, 26, and 35) were considered poor items (value of 0.19 and lower) which meant they needed to be majorly revised or eliminated (Ebel & Frisbie, 1986).

Based on the figures reported above, the researchers decided to keep items 3 and 35. Both these items had indicated rather poor results (both were too easy, and neither had the right level of discrimination power). The reason for this decision was that it is sometimes encouraged in test design to keep some of the too-easy items, especially at the beginning of the test, as a warm-up, or at the end of the test when the test-takers might be rather tired (Gajjar et al., 2014). The other two items (13 and 25) which were both non-discriminatory and too easy were removed from the test since they both stood in the middle of the test and could not be revised. Items 18 and 31, which did not have a very good item discrimination index, went under some minor revisions and were also retained in the final version of the test.

4.2. Construct Validation

Exploratory Factor Analysis (EFA) stands out as a common statistical technique used in scale development and scale adaptation studies in social sciences to determine the underlying latent factors (Bandalos & Finney, 2010; Brown, 2006) and explain the existing structure (Hayton et al., 2004). EFA was used in the present study to investigate the number of factors in our instrument and to find out the items which loaded on each factor. In order to determine the underlying factors of a construct, EFA is an essential step. Confirmatory Factor Analysis (CFA) is also commonly used either apriori for the test of the model or after EFA to test the possible confirmation of the model (Malhotra et al., 2007). It is, however, recommended to carry out a cross-validation study to run CFA (Brown, 2006) since using an independent sample can yield more reliable results (Izquierdo et al., 2014). Therefore, we decided to limit our current work to EFA to explore the dimensions of LAL and leave CFA for further studies with new samples. While using EFA, researchers need to make certain critical decisions (Costello & Osborne, 2005; Schmitt, 2011). They need to determine which method of estimation to use, what criteria to use to determine the number of variables (factors), and whether rotation is required. The following paragraphs discuss the process of EFA conducted and the reasons for the choices made along the way.

The 37 items of the TLALS were subjected to Exploratory Factor Analysis (EFA) using Mplus (version 8.3). Prior to performing EFA, the descriptive model fit statistics were carried out. There are several tests to determine how well the model fits the observed data. For instance, $\chi 2$ is a classic goodness-of-fit measure which is not recommended in the literature due to its

disadvantages, such as being sensitive to sample size (see Brown, 2006 and Jöreskog, 1969 for a detailed discussion of the drawbacks). Among the alternative methods suggested in the literature, Root Mean Square Error of Approximation (RMSEA), which is a measure of fit introduced by Steiger and Lind (1980) and is not sensitive to sample size (Brown 2006); CFI (Comparative Fit Index)/TLI (Tucker–Lewis Index); and SRMR (Standardized Root Mean Square Residual) were assessed. As can be seen in Table 2, RMSEA value is less than 0.05 (0.04<0.05), and CFI /TLI value is 0.98 which is more than 0.95 (as it should be more than 0.95). Moreover, as an absolute measure of model fit criterion, SRMR value is 0.04, which is less than 0.10 or 0.08, and is considered a good fit (Hu & Bentler, 1999). It was, therefore, concluded that the model had a reasonably good fit.

Table 2
Model Fit Information

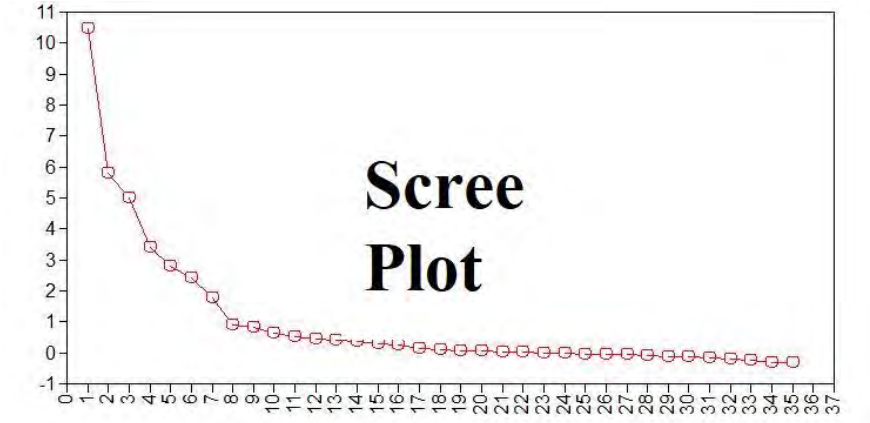
Factor numbers	RMSEA	SRMR	CFI	TLI
7	0.047	0.048	0.987	0.980

Note. RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; TLI = Tucker–Lewis Index.

After extraction, the number of factors to be retained needs to be determined. It is important to note that over-extraction or under-extraction of factors kept for rotation can have a marked effect on the results, which is why choosing a strong method to do so is of utmost importance (Costello & Osborne, 2005). Retaining all factors with eigenvalues greater than 1.0 is reported as one of the least accurate methods (Velicer & Jackson, 1990). The scree test, parallel analysis, and Velicer's MAP criteria are other methods to determine the number of factors. According to Velicer and Jackson (1990), while these three methods are all known to be accurate and user-friendly, the scree test is the best option since the other two methods are not available in the most frequently employed software. Therefore, to determine the number of factors extracted, the scree test was used. The scree plot examines the graph of the eigenvalues and attempts to spot the natural bend in the data where the curve flattens out. The number of datapoints above the "break" is usually the number of factors to retain. Figure 2 shows a clear break after the seventh factor (the point at which the break occurs should not be included). Therefore, using Cattell's (1966) scree test, it was decided to retain seven factors for further investigations.

 ISSN 2476-5880

Figure 2
Scree Plot



Once factors are extracted, factor rotation is conducted to minimize the complexity of the factor loadings and increase their interpretability. Factor rotation can be orthogonal or oblique, both of which are commonly used. The difference, however, lies in their underlying assumption. In orthogonal rotation, the assumption is that there are no intercorrelations between the factors, while oblique rotation assumes that factors are correlated. In our study, factor rotation was done using Geomin (oblique rotation) since it is encouraged to consider a certain degree of correlation among the factors to avoid losing important information and have more accurate results (Costello & Osborne, 2005). Moreover, even if the factors are not correlated, both rotation types render identical results. Once the variables were loaded on each factor, the rotated solution indicated seven factors indicating some strong loadings (see Tables 3 to 9).

The results of factor rotation provided the factor loadings of the items which demonstrate the correlation coefficient of each item and the factors. Muthén and Muthén (2012) cite 0.3 as a minimum loading of an item (the figure is 0.32 according to Tabachnick & Fidell, 2001), 0.3 to 0.6 as moderate factor loadings, and 0.6 or higher as strong factor loadings. Accordingly, two of the items in the LALT were deleted as their factor loadings were not within the acceptable range. The results showed that the first factor consisted of seven items, the second factor five items, the third factor four items, the fourth factor four items, the fifth factor seven items, the sixth factor four items, and the last factor four items. There were 37 items left in the test.

Table 3 demonstrates the rotated factor loadings of the first factor, all of which stand above 0.6 and are, therefore, considered acceptable and representative of the first standard in the seven standards as the latent variable. That means all seven items under this factor assess language teachers' knowledge and skills in "choosing appropriate assessment to make instructional decisions".

Table 3
EFA Theoretical Model for Latent Factor 1: Standard 1

Items	Rotated factor loadings
F1T1	0.931
F1T2	0.958
F1T3	0.881
F1T4	0.933
F1T5	0.707
F1T6	0.968
F1T7	0.751

Table 4 demonstrates the rotated factor loadings of the second factor, all of which stand above 0.6 and are, therefore, considered acceptable and representative of the second standard in the seven standards as the latent variable. That means all seven items under this factor assess language teachers' knowledge and skills in "developing appropriate assessment methods to make instructional decisions ".

Table 4
EFA Theoretical Model for Latent Factor 2: Standard 2

Items	Rotated factor loadings
F2T1	0.799
F2T2	0.888
F2T3	0.647
F2T4	0.615
F2T5	0.878
F2T6	0.648
F2T7	0.820

Table 5 demonstrates the rotated factor loadings of the third factor, all of which stand above 0.6 and are, therefore, considered acceptable and representative of the third standard in the seven standards as the latent variable. That means all the four items under this factor assess language teachers' knowledge and skills in "administering, scoring, and explaining the results of assessment."

 ISSN 2476-5880

Table 5
EFA Theoretical Model for Latent Factor 3: Standard 3

Items	Rotated factor loadings
F3T1	0.830
F3T2	0.869
F3T3	0.814
F3T4	0.831

Table 6 demonstrates the rotated factor loadings of the fourth factor, all of which stand above 0.6 and are, therefore, considered acceptable and representative of the fourth standard in the seven standards as the latent variable. That means all four items under this factor assess language teachers' knowledge and skills in "using assessment results for decision-making about students."

Table 6
EFA Theoretical Model for Latent Factor 4: Standard 4

Items	Rotated factor loadings
F4T1	0.697
F4T2	0.946
F4T3	0.944
F4T4	0.877

Table 7 demonstrates the rotated factor loadings of the fifth factor, all of which stand above 0.6 and are, therefore, considered acceptable and representative of the fifth standard in the seven standards as the latent variable. That means all four items under this factor assess language teachers' knowledge and skills in "developing valid scoring procedures."

Table 7
EFA Theoretical Model for Latent Factor 5: Standard 5

Items	Rotated factor loadings
F5T1	0.812
F5T2	0.767
F5T3	0.844
F5T4	0.753

Table 8 demonstrates the rotated factor loadings of the sixth factor, all of which stand above 0.6 and are, therefore, considered acceptable and representative of the sixth standard in the seven standards as the latent variable. That means all four items under this factor assess language teachers' knowledge and skills in "communicating the assessment results to different stakeholders."

 ISSN 2476-5880

Table 8
EFA Theoretical Model for Latent Factor 6: Standard 6

Items	Rotated factor loadings
F6T1	0.918
F6T2	0.904
F6T3	0.833
F6T4	0.988

Table 9 demonstrates the rotated factor loadings of the seventh factor, all of which stand above 0.6 and are, therefore, considered acceptable and representative of the seventh standard in the seven standards as the latent variable. That means all the five items under this factor assess language teachers' knowledge and skills in "specifying illegal and inappropriate methods."

Table 9
EFA Theoretical Model for Latent Factor 7: Standard 7

Items	Rotated factor loadings
F7T1	0.918
F7T2	0.700
F7T3	0.700
F7T4	0.964
F7T5	0.976

The final 35-item test is a scenario-based test consisting of five scenarios measuring Iranian EFL teachers' language assessment literacy (see Appendix C).

4.3. Internal Consistency

The Cronbach alpha coefficient for the test with 39 items was 0.93. Since this value is greater than 0.70, it indicates a high level of internal consistency (Nunnally, 1978). Thus, the newly developed scale (TLALS) is a reliable measure of EFL teachers' assessment literacy.

5. Conclusion

Over the past two decades, there has been a push toward developing measures with a major focus on LAL due to the peculiarity of the language element which could add further complexities to the concepts of assessment and assessment literacy. Hence, the current study was undertaken to develop a test to measure EFL teachers' levels of assessment literacy. In this connection, a 35-item LALT was developed and validated through investigating the construct validity of the LALT by collecting data from 168 Iranian EFL teachers. In order to confirm the factorability of the data, the descriptive model fit statistics were assessed. The results of EFA showed a seven-factor solution for the test items. The test is a scenario-based test with five scenarios that measure Iranian EFL

teachers' language assessment literacy. In line with the 1990 Standards, some of the items are meant to gauge basic notions about assessment and testing, such as using assessment activities for assigning student grades and informing students and parents of the results of assessments; others are related to knowledge of standardized testing, and the remaining items revolve around classroom assessment. The results showed that the first factor included seven items, the second factor five items, the third factor four items, the fourth factor four items, the fifth factor seven items, the sixth factor four items, and the seventh factor four items.

The present study has some limitations which need to be considered in interpreting the outcomes. For one thing, we would caution researchers to keep in mind that Exploratory Factor Analysis (EFA) is the first step in test development or adaptation, and leaping to firm conclusions based on the analysis carried out at this stage is hardly acceptable. As Izquierdo et al. (2014) put it, using a Confirmatory Factor Analysis (CFA) on the same sample after carrying out EFA is rather "sneaky" (p. 396) since the element of chance can unjustly contribute to the results of CFA. It is, therefore, recommended to test the validation of the instrument using CFA in other independent samples (Brown, 2006). Hence, to provide more informative analyses, the next step after this study could be to conduct CFA using a different sample group to determine whether the instrument has the same structure across various populations.

For another thing, since the participants in this study were not selected randomly, there is an element of sampling bias that should be considered, and additional research can play a part in further validation of the instrument. For instance, due to the challenges of the data collection process, the sample consisted of predominantly female participants, a great majority of MA students and graduates come from the field of EFL, and language institute EFL teachers significantly outnumber school EFL teachers. It is, therefore, recommended to focus on samples more representative of EFL teachers working in public and private schools as well as EFL teachers coming from different backgrounds and levels of education in future studies. Moreover, the online administration of the test would mean a rather restricted degree of control over the respondents. We attempted to reduce this predicted limitation through alpha and beta testing stages, as recommended by Fulcher (2011) and Fulcher and Davidson (2007), which can contribute to the validation process.

Finally, one other limitation of the study is that the instrument is based on the selected topics from the seven standards which, despite their ongoing popularity in the field of general education, are not enclusive enough to cover the more recent notions in assessment or the elements in LAL. Yet, as discussed earlier in the present study, the multifaceted nature of LAL and the complexities it involves due to the addition of linguistic knowledge and skills make the new concept of LAL intricate and still growing (Giraldo, 2021b). Therefore, it is important for further studies to, for example, explore teachers' LAL competencies through taking a skill-specific view and focusing on different language skills.

As a concluding remark, despite the aforementioned shortcomings, this study can have important implications for EFL teachers, language testers, curriculum developers, and teacher trainers. On a broad scale, the instrument can help shed light on our overall understanding of the

state of LAL among EFL teachers. More specifically, if teacher trainers are well-informed about EFL teachers' levels of assessment literacy, they can develop teacher training courses tailored to teachers' assessment literacy levels and needs, enabling them to instruct teachers on different facets of their assessment literacy, which can, in turn, help EFL teachers improve their assessment knowledge and practice. We also hope this study can provide a preliminary model for developing and adapting other measures, which will help further expand our understanding of the field.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Ahmadi, S., Ghaffary, S., & Shafaghi, M. (2022). Examining teacher assessment literacy and instructional improvement of iranian high school teachers on various fields of study. *International Journal of Language Testing*, *12*(1), 1-25.

American Federation of Teachers, National Council on Measurement in Education & National Education Association (1990). Standards for teacher competence in educational assessment of students. *Educational Measurement: Issues and Practice*, *9*(4), 30-32.

Baker, B. A., & Riches, C. (2018). The development of EFL examinations in Haiti: Collaboration and language assessment literacy development. *Language Testing*, *35*(4), 557-581.

Bandalos, D. L., & Finney, S. J. (2010). Factor analysis: Exploratory and confirmatory. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 93-114). Routledge.

Bailey, K. M., & Brown, J. D. (1995). Language testing courses: What are they? In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 236–256). Multilingual Matters.

Bøhn, H., & Tsagari, D. (2021). Teacher educators' conceptions of language assessment literacy in Norway. *Journal of Language Teaching and Research*, *12*(2), 222-233.

Brindley, G. (2001). Language assessment and professional development. In C. Elder, A. Brown, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honor of Alan Davie*s (pp. 126–36). Cambridge University Press.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 30*(1), 3-12.

Brown, T. A. (2006). Confirmatory factor analysis for applied research. Guilford Press.

Brown, J. D., & Bailey, K. M. (2008). Language testing courses: What are they in 2007? *Language Testing, 25*(3), 349–384.

Campbell, C., Murphy, J. A., & Holt, J. K. (2002). *Psychometric analysis of an assessment literacy instrument: Applicability to pre-service teachers*. Paper presented at the annual meeting of the Mid-Western Educational Research Association, Columbus, OH.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, *1*(2), 245-276.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation, 10*(7), 1-8.

Coombe, C., Vafadar, H., & Mohebbi, H. (2020). Language assessment literacy: What do we need to learn, unlearn, and relearn?. *Language Testing in Asia, 10*(1), 1-16.

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

DeLuca, C., Lapointe-Mcewan, D., & Luhanga, U. (2016). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability*, *28*(3), 251-272.

Downing, S.M. (2006). Twelve steps for effective test development. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp.3-25). Lawrence Erlbaum Associates.

Ebel, R. L. & Frisbie, D. A. (1986). *Essentials of educational measurement*. Prentice-Hall.

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.

Fulcher, G. (2011). *Practical language testing*. Routledge.

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

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. Routledge.

Gajjar, S., Sharma, R., Kumar, P., & Rana, M. (2014). Item and test analysis to identify quality multiple choice questions (MCQs) from an assessment of medical students of Ahmedabad, Gujarat. *Indian Journal of Community Medicine: Official Publication of Indian Association of Preventive & Social Medicine*, *39*(1), 17-20.

Gardner, S., & Rea-Dickins, P. (2001). Conglomeration or chameleon? Teachers' representations of language in the assessment of learners with English as an additional language. *Language Awareness*, *10*(2-3), 161-177.

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

Giraldo, F. (2021a). Language assessment literacy and teachers' professional development: A review of the literature. *Profile: Issues in Teachers' Professional Development*, *23*(2), 265–279.

Giraldo, F. (2021b). A reflection on initiatives for teachers' professional development through language assessment literacy. *Profile: Issues in Teachers' Professional Development, 23*(1), 197–213.

Gotch, C. M., & French, B. F. (2014). A systematic review of assessment literacy measures. *Educational Measurement: Issues and Practice, 33*(2), 14–18.

Harding, L., & Kremmel, B. (2016). Teacher assessment literacy and professional development. In *Handbook of second language assessment* (pp. 413-428). De Gruyter Mouton.

Hasselgreen, A., Carlsen, C., & Helness, H. (2004). *European survey of language testing and assessment needs*. (Research Report: Part one: General findings). http://www.ealta.eu.org/documents/resources/survey-report-pt1.pdf

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*(2), 191-205.

Henning, G. (1988). The influence of test and sample dimensionality on latent trait person ability and item difficulty calibrations. *Language Testing*, *5*(1), 83-99.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55.

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

Inbar-Lourie, O. (2008b). Language assessment culture. In E. Shohamy, & N. Hornberger (Eds.), *Encyclopedia of language and education: Language testing and assessment* (pp. 285-300). Springer.

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

Inbar-Lourie, O. (2013b, November). *Language assessment literacy: What are the ingredients?* Paper presented at the 4th CBLA SIG Symposium Programme, University of Cyprus.

Inbar-Lourie, O. (2017). Language assessment literacy. In *Language Testing and Assessment* (pp. 257-270). Springer.

Izquierdo Alfaro, I., Olea Díaz, J., & Abad García, F. J. (2014). Exploratory factor analysis in validation studies: Uses and recommendations. *Psicothema, 26*(3), 395-400.

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

Jin, Y. (2010). The place of language testing and assessment in the professional preparation of foreign language teachers in China. *Language Testing, 27*(4), 555-584.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183-202.

Kremmel, B., & Harding, L. (2020). Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy survey. *Language Assessment Quarterly*, *17*(1), 100-120.

Kunnan, A. J. (2019). An Agenda for Language Assessment Development, Research and Pedagogy. *Journal of Asia TEFL*, *16*(1), 327.

Lane, S., Raymond, M.R., Haladyna, T.M., & Downing, S.M. (2016). Test development process. In S. Lane, M.R. Raymond, & T.M. Haladyna (Eds.), *Handbook of test development* (pp.3-18). Routledge.

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological science*, *23*(11), 1337-1344.

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

Mertler, C. A. (2003). *Pre-service versus in-service teachers' assessment literacy: Does classroom experience make a difference?* Paper presented at Annual meeting of the Mid-Western Educational Research Association, Columbus, OH.

Mertler, C. A. (2004). Secondary teachers' assesment literacy: Does classroom experience make a difference? *American Secondary Eduation, 33*(1), 49-64.

Mertler, C. A., & Campbell, C. S. (2005). *Measuring teachers' knowledge and application of classroom assessment concepts: Development of the Assessment Literacy Inventory.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Mohammadkhah, E., Kiany, G. R., Tajeddin, S. Z., & ShayesteFar, P. (2022). Teachers' conceptions of language assessment: Theoretical knowledge and attitudinal dimensions of language assessment literacy model. *International Journal of Language Testing*, *12*(1), 82-102.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

National Council on Measurement in Education. (1990). *Standards for teacher competence in educational assessment of students*. ERIC Clearinghouse.

Nunnally, C. J. (1978). *Psychometric theory*. McGraw-Hill.

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.

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

Pastore, S., & Andrade, H. L. (2019). Teacher assessment literacy: A three-dimensional model. *Teaching and Teacher Education, 84*, 128-138.

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

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.

Schmitt, R. S. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment, 29*(4), 304-321.

Shohamy, E., & McNamara. T. (2009). Language tests for citizenship, immigration, and asylum. *Language Assessment Quarterly, (6)*1, 1–5.

Stabler-Havener, M. L. (2018). Defining, conceptualizing, problematizing, and assessing language teacher assessment literacy. *Working Papers in Applied Linguistics and TESOL, 18*(1), 1–22.

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

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan, 77*(3), 238-245.

Stiggins, R. (2010). Essential formative assessment competencies for teachers and school leaders. In H. L. Andrade, & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp.233-250). Taylor & Francis.

Tao, N. (2014). *Development and validation of classroom assessment literacy scales: English as a Foreign Language (EFL) instructors in a Cambodian higher education setting* (Doctoral dissertation, Victoria University).

Tabachnick, B. G. , & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). HarperCollins.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics, 29*, 21–36.

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

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate behavioral research*, *25*(1), 1-28.

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

Weng, F., & Shen, B. (2022). Language Assessment Literacy of Teachers. *Frontiers in Psychology*, *13*.

Xu, Y., & Brown, G. T. (2017). University English teacher assessment literacy: A survey-test report from China. *Papers in Language Testing and Assessment, 6*(1), 133-158.

Yan, X., Zhang, C., & Fan, J. J. (2018). "Assessment knowledge is important, but…": How contextual and experiential factors mediate assessment practice and training needs of language teachers. *System*, *74*, 158-168.

Yastıbaş, A. E., & Takkaç, M. (2018). Understanding language assessment literacy: Developing language assessments. *Journal of Language and Linguistic Studies*, *14*(1), 178-193.

Appendix A
Content Domain of Assessment Literacy (AFT, NCME, & NEA, 1990)

Assessment literacy components	Competencies (conceptual and application skills)
Standard 1: Choosing assessment methods	• Using the concepts of assessment error and validity when selecting approaches to classroom assessment • Understanding how valid assessment data can help instructional activities • Understanding the impact of invalid information on instructional decisions • Using and evaluating assessment choices while taking into account, among other things, students' backgrounds. • Being aware that certain assessment approaches can conflict with specific instructional objectives

 ISSN 2476-5880

Standard 2: Developing assessment methods	• Planning the collection of information that assist in decision-making • Knowing and following proper principles for creating and using appropriate assessment methods in teaching • Analyzing student data to evaluate each assessment method's effectiveness
Standard 3: Administering, scoring, and interpreting the results	• Interpreting informal and formal teacher-produced assessment results • Using stencils for scoring response-choice questions, guides for marking essay questions and projects, and rubrics for rating performance assessments • Administering standardized achievement tests and interpreting the commonly reported scores • Having a conceptual understanding of the summary indexes frequently reported with assessment results • Applying the score and summary index principles in ways that improve how assessments are used • Analyzing assessment results to determine students' strengths and weaknesses • Using assessment methods to support students' educational development
Standard 4: Using assessment results	• Using assessment information to prepare a suitable instructional plan to aid students' educational progress • Interpreting the results correctly and avoiding common misinterpretations • Knowing about the results of local, regional, state, and national assessments and how to use them to enhance education
Standard 5: Developing valid grading procedures	• Devising, implementing, and explaining a system for developing grades • Understanding and articulating why the grades given are accurate • Recognizing and avoiding improper methods of grading, such as utilizing grades as a form of punishment • Improving the validity of their interpretations by evaluating and modifying their grading procedures
Standard 6: Communicating the results	• Giving appropriate explanations of how students' background must moderate how their assessments are interpreted • Explaining that assessment results do not suggest that background factors do not limit students' educational development • Communicating to students and parents how progress is assessed • Understanding and explaining the importance of taking measurement errors into account when making decisions about individual students • Explaining the limitations of formal and informal assessment methods. • Explaining printed reports of the results of student assessments
Standard 7: Recognizing unethical assessment methods	• Knowing the laws and case decisions that affect their assessment practices • Knowing that various assessment procedures can be misused or overused, having harmful consequences

Appendix B
Sample Items for Each Standard

Standard	Item
1	Read the scenario first and select the best choice. Ms. Farhang needs an assessment that can help her give the students feedback on the relative strengths and weaknesses of their reading and writing skills. To this end, she needs to use a) a grammar and vocabulary test b) an integrated skills test c) an oral performance test d) a test with separate sections on reading and writing
2	Ms. Yegane is planning to design an integrated skills assessment task for writing. She needs to integrate: **Select ALL the answers you think are correct.** a) reading and writing b) listening and writing c) speaking and writing d) written skills
3	Ms. Adib, a B2-level English instructor, needs to know which of the following specifications is relevant to the "coherence and cohesion" criterion in the rating scale for writing. a) topic development b) the effective use of conventions of the communicative task c) unity of ideas and unity of structural elements of the text d) the range of vocabulary items used in the text
4	Ms. Farhang has recently assigned his class a writing task. Her assessment of the grammar used in the students' writing shows that the students have frequent mistakes in using the grammar taught in the recent unit. She decides to prepare a remedial lesson. In her remedial lesson, she should focus on all of the followings EXCEPT a) form b) function c) use d) morphosyntax
5	The only student work Ms. Toosi grades for the current rating period is the multiple-choice test. What is the main objection to this practice? a) The test, and therefore the grades, reflect a constrained curricular focus. b) These grades are likely to be biased against some students because they are solely based on test results. c) The scores of multiple-choice tests are greatly affected by guesswork. d) The assessment results are based on students' performance on one assessment task.
6	Nader is a student in Ms. Farhang's class. He receives a raw score of 12 items answered correctly out of a possible 15 on the reading section of a standard test. The raw score is equal to a percentile rank of 45. His parents do not understand how he could answer so many items correctly and still have such a low percentile rank. They approach Ms. Farhang in search of a possible explanation. Which of the following can be an appropriate explanation offered to his parents? a) "I'm not sure… there may be a problem with how the test company figured the scores." b) "Although Nader successfully answered 12 items correctly, numerous students did better."

c) "Raw scores do not show the relative stance of a score in a distribution; we need percentile rank as a form of norm-referenced scoring to tell us about the relative stance."
d) "The mean score of this class may be quite less than Nader's score."

7 Based on students' grades from her previous B1 course, Ms. Toosi believes that some of her students with poor test scores are more intelligent than their test scores imply. Based on this argument, she decides to increase their marks by adding extra points to their exam results. What ethical issue can be raised against Ms. Toosi's decision?

a) using information not relevant to the current test
b) adjusting grades in her course
c) using previous grades to adjust current grades
d) adjusting some students' grades and not others

Scoring Key

1. d
2. a, b, c
3. c
4. d
5. d
6. b
7. d

Appendix C
Language Assessment Literacy Test

Dear colleague,

The following scenario-based test consists of five scenarios aiming at measuring Iranian EFL teachers' language assessment literacy. Some of the items are intended to measure general concepts related to testing and assessment, including the use of assessment activities for assigning student grades and communicating the results of assessments to students and parents; other items are related to knowledge of standardized testing, and the remaining items are related to classroom assessment.

Directions:

Read each scenario followed by each item carefully; select the response you think is the best one and mark your response. Even if you are unsure of your choice, ***mark the response you believe to be the best****.*

If you need a break, ***please do so after completing a scenario****.*

Your time and effort are truly valued.

Section 1: Demographic Information

1. What gender do you identify as?

A. Male
B. Female

2. How old are you?

A. 18- 30 years old
B. 31- 40 years old
C. 41-60 years old
D. 60+

3. Please select the type of university degree(s) you have received so far and specify the degree discipline/field of study
   A. Undergraduate degree (…….)
   B. Master's degree (…….)
   C. PhD degree (……)

4. Please select your last field of study at university
   A. English Language Teaching
   B. Linguistics
   C. Translation studies
   D. English literature
   E. Others

5. Please select the type(s) of institution where you teach
   A. Public school (primary, lower secondary, upper secondary)
   B. Private language institute or academy

6. How many years of Experience do you have?
   A. Less than 5 years
   B. More than 5 years

7. Please specify the training course(s) you've attended to improve your practice as an EFL teacher.
   Training course title ……
   Duration ……

8. During your pre-service or in-service teacher training, have you learned something about language testing and assessment (theory and practice)?
   If yes, please select the content item(s) you think were covered in the course(s) you passed.
   A. Test specifications/Item writing
   B. Test administration
   C. Test scoring (e.g., transform numeric scores into letter grades)
   D. Standardized testing (interpreting, analyzing)
   E. Test critiquing
   F. Test-taking skills or strategies
   G. Test theory (e.g., validity, reliability)
   H. Basic statistics (e.g., mean, percentile, bell curve)
   I. Advanced statistics (e.g., Item Response Theory)
   J. Test ethics
   K. History of language testing

L. Classroom assessment (developing)
M. Alternative/performance assessment
N. Rubric development (analytic and holistic)
O. Rater training
P. Test accommodation
Q. Other(s), please specify: ……..

9. Please select the content items of your language assessment course that you think were **most relevant** to your teaching?

A. Test specifications/Item writing
B. Test administration
C. Test scoring (e.g., transform numeric scores into letter grades)
D. Standardized testing (interpreting, analyzing)
E. Test critiquing
F. Test-taking skills or strategies
G. Test theory (e.g., validity, reliability)
H. Basic statistics (e.g., mean, percentile, bell curve)
I. Advanced statistics (e.g., Item Response Theory)
J. Test ethics
K. History of language testing
L. Classroom assessment (developing)
M. Alternative/performance assessment
N. Rubric development (analytic and holistic)
O. Rater training
P. Test accommodation
Q. Other(s), please specify: …….

10. Please select the content items that you still **need to** develop?
A. Test specifications/Item writing
B. Test administration
C. Test scoring (e.g. transform numeric scores into letter grades)
D. Standardized testing (interpreting, analyzing)
E. Test critiquing
F. Test-taking skills or strategies
G. Test theory (e.g. validity, reliability)
H. Basic statistics (e.g. mean, percentile, bell curve)
I. Advanced statistics (e.g. Item Response Theory)
J. Test ethics
K. History of language testing
L. Classroom assessment (developing)
M. Alternative/performance assessment
N. Rubric development (analytic and holistic)
O. Rater training
P. Test accommodation
Q. Other(s), please specify

Section 2: LALT Scenarios and Items
Scenario #1: Ms. Ahmadi is an EFL teacher teaching an intermediate-level class.

1. Ms. Ahmadi questions how well her students can apply what they have learned in the class to situations encountered in their everyday lives. Although the teacher's manual contains numerous items to test mastery of the skills and sub-skills taught, she is not convinced that giving a paper-and-pencil test is the best method for determining what she wants to know. The type of assessment that would best answer Ms. Ahmadi's question is called a/an

a) performance assessment
b) diagnostic assessment
c) formative assessment
d) alternative assessment

2. Ms. Ahmadi is trying to recreate a real-life speaking situation in an assessment. She needs to consider all the following key factors **EXCEPT**

a) the assessment instructions
b) students' willingness to communicate in speaking
c) different types of speaking tasks
d) the purpose of the speaking task

3. To make sure that her students' performance on an external standard achievement test is truly above the level of students' performance on the same test at a national level, Ms. Ahmadi is recommended to do all the following **EXCEPT**

a) obtain the mean of her students and the national mean
b) estimate the standard error of measurement
c) run a z-test on the mean of her students and the national mean
d) check the results of the z-test to find out the magnitude of mean differences.

4. Which of the following is an appropriate use of the results from the external standard achievement test referred to above?
Select ALL the answers you think are correct.

a) estimating the students' ability to learn a second language
b) measuring students' knowledge and skills based on a set curriculum
c) determining students' strengths and weaknesses
d) estimating the students' level of the learning outcomes

5. During a meeting with language learners' parents, one of the parents asks Ms. Ahmadi what she means by saying that her daughter has scored in the 80th percentile in an external multiple-choice standard test. Which of the following provides the best explanation of the student's score?

a) She got 80% of the items on the English test correct.
b) She is likely to earn a grade of "80" in her English class.
c) She is demonstrating above-grade-level performance in English.
d) She scored the same or better than 80% of the norm group.

6. To assess her students' progress during the term, Ms. Ahmadi uses a range of different assessment tasks. These assessments range from giving short quizzes following each unit of instruction, and checking assignments, to administering an end-of-the-term final exam. To improve the validity of this grading procedure, Ms. Admadi should

a) use the same scale for grading all assessments.
b) consider the results of each assessment task independent of all others.
c) weight assessments according to their relative importance.
d) take into consideration each student's degree of hard work in assigning the final grades.

7. Which of the following is **NOT** an appropriate use of Ms. Ahmadi's assessment information?
a) giving priority to the end-of-the-term exam when making decisions about students' learning
b) using quizzes to diagnose the students' problems and plan remedial instruction
c) using the results of all assessment tasks to come up with an aggregate score
d) returning the quiz papers to the students and providing the correct answers through question and answer

Scenario #2: Mr. Parsa, an EFL teacher, knows that his students will be sitting a standard achievement test at the end of the term.

8. Mr. Parsa's learning activities for this term will focus on oral and written comprehension and production. He wants to assess the students to determine if any reinstruction will be necessary prior to the standard achievement test that the students will take at the end of the term. Which of the following assessment strategies would be the most appropriate choice?
a) He can use a mock standard achievement test to diagnose the students' gaps in comprehension and production.
b) He can use a short oral interview followed by a composition test.
c) He can use the tests included in the teachers' manual.
d) He can compose a test by selecting items from sample tests available on the internet.

9. Mr. Parsa wants to give his students feedback and help them identify their specific strengths and weaknesses in writing. He decides to use a rating scale that gives separate scores for different aspects of a writing sample. What kind of scale and assessment does he need to use?
a) holistic rating scale/ formative
b) holistic rating scale/ summative
c) analytic rating scale/ formative
d) analytic rating scale/ summative

10. At the end of each class period, Mr. Parsa does a quick "check-in" with his students to get an impression of their understanding. In this example, the primary purpose of conducting formative assessment is to
a) determine the final grades of students.
b) determine content for the final exam.
c) plan classroom instruction.
d) evaluate curriculum appropriateness.

11. Parisa, one of Mr. Parsa's students from a previous semester, received a percentile rank of 60 on the listening comprehension subtest of the standard achievement exam. This score is most appropriately interpreted as which of the following?
a) Parisa scored above average.
b) Parisa scored below average.
c) Parisa scored at the institute's/school's average.
d) Not enough information was provided for making a judgment.

12. Ali, another student in Mr. Parsa's class, receives a T-score of 47 on the reading comprehension portion of the standard achievement test. The cut-score is 50; therefore, Ali does not pass this subtest. However, the subset has a standard error of measurement equal to 1.6. Which of the following is the best decision for Mr. Parsa regarding instruction appropriate to meet Ali's needs?
a) Ali has not achieved the minimum reading comprehension level and should receive remedial reading instruction.
b) Mr. Parsa knows that Ali could have scored higher, so the test results should be ignored.

c) Ali has achieved the minimum reading comprehension level; nothing different or additional should be done.
d) Mr. Parsa knows that Ali should have scored much lower, so the test results should be ignored.

13. Sarah, another student in Mr. Parsa's class, scored 17 on the reading comprehension portion of the test with a mean of 15 and a standard deviation of 2. She scored 16 on the listening section of the test with a mean of 14 and a standard deviation of 3. Based on the above information, in comparison to her peers, which statement is correct about Sarah?
a) Sarah is better at reading than listening.
b) Sarah is better at listening than reading.
c) Sarah scored above average in both skills.
d) Sarah scored one SD above the average in reading.

14. Mr. Parsa was worried that his students would not perform well on the standard level achievement test. He did all of the following to help increase students' scores. Which was unethical?
a) He instructed students on strategies for taking multiple-choice tests, such as how to use answer sheets.
b) He planned his instruction to focus on the skills covered on the test.
c) He encouraged the students to do their best and provided them with a reward after testing was complete.
d) He allowed students to practice with items from an alternative exam form.

Scenario #3 Ms. Toosi teaches a B1 English course. She is about to finish the course and has decided to give her students a single end-of-the-term multiple choice test.

15. Ms. Toosi is supposed to develop the multiple-choice test herself and does not have the means or time to pilot her test in the same way as international test bodies. She should do all the followings **EXCEPT**
answering the questions himself.
a) reviewing the items closely to spot any possible error.
b) showing the test to a/ some colleague(s) and asking for feedback.
c) getting feedback from the students while they are taking the test.

16. Suppose Ms. Toosi had the time and means to conduct pre-testing. After pre-testing, she should conduct an item analysis on the results and examine all of the following **EXCEPT**
a) item difficulty index.
b) item discrimination index.
c) choice distribution.
d) criterion-related validity.

17. Being aware of the limitations of multiple-choice items, she uses this item format to assess all **EXCEPT**
a) comprehension skills.
b) knowledge of language elements.
c) productive use of language.
d) receptive use of language.

18. Ms. Toosi decides to score the tests using a scale of 0 to 100. Generally speaking, what is the proper interpretation of a student score of 85 on this scale?
a) The student answered 85% of the items on the test correctly.

b) The student knows 85% of the content covered by this instruction unit.
c) The student scored above 85% of other students who took this test.
d) The student scored lower than 85% of other students who took the test.

19. The only student work Ms. Toosi grades for the current rating period is the multiple-choice test. What is the major criticism of this practice?
a) The test, and therefore the grades, reflect a constrained curricular focus.
b) These grades are likely to be biased against some students because they are solely based on test results.
c) The scores of multiple-choice tests are greatly affected by guesswork.
d) The assessment results are based on students' performance on one assessment task.

20. Mr. Javid, another English instructor, bases his grades primarily on his observations of students during class. Which statement describes the type of assessment they are involved in?
a) Ms. Toosi uses formal assessment; Mr. Javid uses informal assessment.
b) The type of assessment they are undertaking is formative assessment.
c) They use the evidence collected through formative assessment for summative purposes as well
d) Ms. Toosi is involved in summative assessment, while Mr. Javid's undertaking is formative assessment.

21. Based on students' grades from her previous B1 course, Ms. Toosi believes that some of her students with poor test scores are more intelligent than their test scores imply. Based on this argument, she decides to increase their marks by adding extra points to their exam results. What ethical issue can be raised against Ms. Toosi's decision?
a) using information not relevant to the current test
b) adjusting grades in her course
c) using previous grades to adjust current ones
d) adjusting some students' grades and not others

Scenario #4. Mr. Sharif is an English instructor. Experienced in issues of learning assessment, Mr. Sharif is often asked to respond to colleagues' questions concerning best practices for evaluating student learning.

22. Ms. Adib, a B2-level English instructor, asks Mr. Sharif which of the following specifications is relevant to the "coherence and cohesion" criterion in the writing rating scale.
a) topic development
b) the effective use of conventions of the communicative task
c) unity of ideas and unity of structural elements of the text
d) the range of vocabulary items used in the text

Ms. Adib, a B2-level English instructor, needs to know which of the following specifications is relevant to the "coherence and cohesion" criterion in the rating scale for writing.

23. Ms. Yegane is planning to design an integrated skills assessment task for writing. According to Mr. Sharif, she needs to integrate:
Select ALL the answers you think are correct.
a) reading and writing
b) oral skills
c) speaking and writing
d) written skills

24. Mr. Peyvandi is planning to test a specific grammar point his class has recently worked on. He asked Mr. Sharif about the item types that suit better for assessing grammar in context. Which of the following do you think Mr. Sharif recommended Mr. Peyvandi to use?

a) standard cloze test
b) modified cloze test
c) multiple choice test
d) dictation

25. Mr. Shokooh will be teaching a new course next month. He wants to ensure the course content and assessment fit together well to achieve his teaching and learning goals. According to Mr. Sharif, when should he plan the assessment?

a) at the same time as planning the course
b) while the course is going on
c) after the course is taught
d) before the course is planned

26. Ms. Javan is not sure about the concept of "test misuse." Mr. Sharif explains the concept and uses the following example to demonstrate the concept.

"If you use any of the following tests to assess your students' achievement, you've misused the test."

Select ALL the choices you think were mentioned by Mr. Sharif as test misuse.

a) a proficiency test
b) a placement test
c) a diagnostic achievement test
d) a final achievement test

27. Recently, the institute has asked Ms. Rastegar and the other language teachers to prepare the learners for an external test to assess their language level and grant successful candidates a certificate of attainment. She asks Mr. Sharif what technical term applies to this external test.

a) standardized achievement test
b) placement test
c) proficiency test
e) selection test

28. One of the teachers is redesigning her tests to make greater use of "story problems" to check students' grammar understanding. She consults with Mr. Sharif to see what concerns she should be aware of when designing such a test. Which statement is **NOT** an appropriate recommendation when designing this type of assessment?

a) Make sure that the reading level is grade-appropriate.
b) Avoid scenarios more familiar to certain groups over others.
c) Check for clarity of sentence construction.
d) Incorporate scenarios used during instruction.

Scenario #5. Ms. Farhang is responsible for teaching a C1 level English course. Over the past couple of years, her students have struggled with their reading and writing skills, but she is unsure where the specific difficulties lie.

29. Ms. Farhang needs an assessment that can help her give the students feedback on the relative strengths and weaknesses of their reading and writing skills. To this end, she needs to use
a) a grammar and vocabulary test
b) an integrated skills test
c) an oral performance test
d) a test with separate sections on reading and writing

30. Ms. Farhang wishes to conduct an assessment to identify the difficulties her students are experiencing in reading comprehension. Which of the following would best meet her needs?
a) a diagnostic test
b) an informal assessment task
c) a standardized test
d) an authentic assessment task

31. To refine both her instruction and assessment of reading comprehension skills, Ms. Farhang conducts an item analysis on students' performance on the last term's final exam. She should discard or substantially revise a test item that
a) has a difficulty value between .50 and .75.
b) has a discrimination value equal to or above .30.
c) has a discrimination value with a negative sign.
d) has a difficulty value equal to .90.

32. Ms. Farhang has recently assigned his class a writing task. Her assessment of the grammar used in the students' writing shows that the students have frequent mistakes in using the grammar taught in the recent unit. In her remedial lesson, she should focus on all of the followings EXCEPT
a) form
b) function
c) use
d) morphosyntax

33. Ms. Farhang wants to be sure that the midterm grades she assigns to her students' performance in class reflect each student's respective level of content mastery for that unit. Which of the following grading systems would best accomplish this goal?
a) a criterion-referenced grading system
b) a norm-referenced grading system
c) a pass-fail grading system
d) a subjective rating grading system

34. Nader is a student in Ms. Farhang's class. He receives a raw score of 12 items answered correctly out of a possible 15 on the reading section of a standard test. The raw score is equal to a percentile rank of 45. His parents do not understand how he could answer so many items correctly and still have such a low percentile rank. They approach Ms. Farhang in search of a possible explanation. Which of the following can be an appropriate explanation offered to his parents?
a) "I'm not sure… there may be a problem with how the test company figured the scores."
b) "Although Nader successfully answered 12 items correctly, numerous students did better."
c) "Raw scores do not show the relative stance of a score in a distribution; we need percentile rank as a form of norm-referenced scoring to tell us about the relative stance."
d) "The mean score of this class may be quite less than Nader's score."

35. After the term ends, Ms. Farhang finds out that her students have satisfactorily mastered the necessary reading strategies. However, when her students move into higher-level courses, she hears from colleagues that some students perform poorly on items addressing these same strategies. Considering the discrepancy between students' performance in Ms. Farhang's class and their performance in her colleagues' classes, what action is most appropriate when making decisions concerning improving student learning?

a) recommend that classroom instruction be consistent among all instructors teaching the same course
b) ensure alignment between instruction and what knowledge or skills are being measured
c) design a test at a higher level of difficulty to get a better picture of students' comprehension skills
d) identify the percentage of students predicted to perform well in the higher-level course