Language Testing in Asia

RESEARCH Open Access

Designing and validating a potential assessment inventory for assessing ELTs' assessment literacy

Fateme Nikmard and Zohre Mohamadi Zenouzagh*

* Correspondence: Zohreh.Zenooz@gmail.com; Zohre.mohamadi@kiau.ac.ir
English Translation and Teaching Department, Karaj Branch, Islamic Azad University, Karaj, Iran

Abstract

English teachers' assessment literacy has always been considered as an important factor in their performance. However, no instrument has ever been developed to assess this construct among Iranian EFL teachers. To fill this gap, in the first phase of the present study a theoretical framework for the main four components of teacher assessment literacy, named validity, reliability, interpretability of the results, and efficiency, was developed through extensive review of the related literature and conducting interviews with PhD candidates of TEFL. In the second phase, a questionnaire was developed and piloted with 150 participants who took part in the study through the rules of convenience sampling. More specifically, the 30 items of the newly-developed "ELTs' Assessment Literacy" questionnaire were subjected to factor analysis which revealed the presence of all the four components consisting of different number of items. These phases led to the development of a questionnaire with four components and 25 items on the basis of a five point Likert scale that measured: (1) "Validity" including six items, (2) "Reliability" including ten items, (3) "Interpretability of the Results" including eight item, and (4) "Efficiency" including five items. The findings of this study may shed lights on this subject and help researchers and teaching practitioners assess EFL teachers' assessment literacy and make principled decisions as far as assessment is concerned.

Keywords: Assessment, Assessment Literacy, Validity, Reliability, Interpretability, Efficiency

Background

As Green (2014) and Herppich et al. (2018) claimed, assessment and the results drawn from it have crucial effects on the test takers' lives. It also has a straight influence on matters such as making decisions at the level of program and might even lead to organizational changes. As a result of such points, examiners, who are usually seen to be the teachers themselves, are regarded as influential factors within any assessment cycle, and especially in examining speaking and writing tests whose results are very much dependent on the examiners' subjective point of views. Of course, Alderson and Banerjee (2002) provided some solutions to such problems. They named the assessment of pair and group activities as ways of decreasing such subjectivity. Using more

Springer Open

than one examiner was another solution they named. Using the former, a variety of patterns of interaction, for instance, examiner-examinee(s) and examinee(s)-examiner, are used, and in the latter, it is claimed that the assessment is fairer with two examiners (Alderson & Banerjee, 2002). Such considerations will lead to an enhanced amount of validity which has different types. Fulcher (1997) introduced *face validity* (student awareness of the test), *content validity* (whether the test content and program content are the same), *construct validity* (related to the results checked through correlations), *concurrent validity* (self-assessment, and the assessment of tutors), and *reliability* as very important points to be taken into account in any assessment.

As a matter of fact, the essential point in Herppich et al.'s (2018) ideas is that any teacher needs to provide a reliable and valid judgment about each and every individual's achievement, performance, and characteristics. That is, competent teachers have to take into account the common judgment biases if they want to have a valid and reliable judgment. Furthermore, an important point to be careful about in having at least somehow adequate judgment is to consider assessment steps and decision points that help the examiner to recognize the necessary knowledge needed in performing these steps (Herppich et al., 2018).

Language testing which is considered as an important matter throughout the twenty-first century has caused the language testing profession to be more and more important as a result of which language assessment literacy has received more attention and prominence (Fulcher, 2012). It means, as Fulcher (2012) stated, assessment literacy plays a crucial role in the construction of the new educational programs and materials since language assessment and testing ought to change so that it can meet the new emerging needs of all involved in the field of assessment including teachers as well.

Subsequently, regarded as a prominent matter, assessment literacy was the subject of a new study carried out by Kim, Chapman, Kondo, and Wilmes (2020) who tried to examine the assessment literacy educators which need to be able to interpret score reports. Although a questionnaire consisting of just 15 items was used in the first phase of this research to gather data from the teachers, the items are merely focused on the teachers' perception of the usefulness and meaningfulness of the scores. It means it did not take the four components under study in the present research into consideration when collecting data about the way teachers assign the scores.

Other researchers such as Deygers and Malone (2019) also worked on the assessment literacy with the main goal of checking the amount of the assessment literacy of the university admission officers and policy makers in which the only instrument used was the interviews conducted with the so-called officers and policy makers. Then, the data collected was qualitatively analyzed. The point is that in the case of this inquiry as well, there was not a tool, a questionnaire for instance, to be used as an instrument for collecting more clear-cut data.

As there are a range of different classroom instructional practices from individual work of doing a task to group discussions or to the use of media such as video in the classrooms all of which can be used as the materials for assessing learners, it is of utmost importance for each and all EFL teachers to have a fair amount of information about such matters (Hougen, 2015). Considering the matter of assessment literacy from the instructional practices point of view, Salimi and Farsi (2018) worked on two groups of EFL native and non-native English teachers' viewpoint about classroom assessment literacy. At the end of their research, they found out that there is a considerable difference between the two groups' perspective towards the classroom assessment literacy. However, such a difference was not observed between male and female teachers. That is, no difference whether a teacher is a man or a woman, their standpoints are not very much different.

Consequently, as Herppich et al. (2018) asserted, the existing gap in language teachers' literacy is that there is no valid criteria to measure teachers' assessment competence, i.e., their judgment accuracy, which makes the job of a researcher a piece of work open to question. As they believe, although there are some ways to make sure of the accuracy of the judgments, such as checking the correlation between teachers' judgments of the students' characteristics and students' outcomes in a standardized test, it is a better idea to have a validity measure for evaluating teachers' assessment competence. In fact, most of the studies conducted so far have focused on reliability and validity issues and not the other dimensions of assessments. It is also stated that the criteria based on which validity is judged is still unclear (Herppich et al., 2018).

Furthermore, most of the inquiries carried out in the field of assessment literacy so far have benefitted from quantitative measures (Coombe, Vafadar, & Mohebbi, 2020) and with no direct reference to the four main aspects of assessment literacy which are the focus of this research. As a result, because there was not any specific criteria developed for estimating the knowledge of EL teachers' regarding validity, reliability, interpretability of the results, and efficiency, and to fill the gap provided earlier, the researchers of the present study through of developing a questionnaire focusing on the so-called four main components of assessment literacy. Some detailed items were written for each of these four concepts based on the data derived from some discussions as well as the literature available on the issues.

Literature review

Assessment

Having the knowledge of assessment is an indispensable part of being an EFL teacher since they always need to be sound and fair in decisions they make about the learners' progress and achievement (Farhady & Tavassoli, 2018). As Brown and Abeywickrama (2018) claimed, in the recent decades, assessment has changed a lot from the traditional one-shot testing, and it has gained many fans as a kind of ongoing procedure of gathering information on the learners' performance based on which their achievement is evaluated. Such an influential trend in the process of assessment has caused many changes since almost all stakeholders, including teachers, need to diverge from the traditional kind of testing approach and try to advance their assessment knowledge to be prepare enough to act as an updated teacher (Farhady & Tavassoli, 2018). Accordingly, Xu (2018) asserted that assessment, as a vital means of maintaining student learning, is on the way of obtaining recognition and gaining momentum in the research literature whose main aims is, in fact, to encourage and help teachers learn more and more about prominent matters of language assessment. As Xu (2018) highlighted the point, since teachers have the critical role of using assessment throughout their classes, they have to know more about assessment (i.e., assessment literacy).

Assessment literacy

There are a lot of different aspects in language assessment one of which is assessment literacy which is a very important one, in fact. According to the definition provided by Ng, Xie, and Wang (2018), assessment literacy is the extent to which educators understand the rules and regulations along with the practices of an acceptable assessment which is an influential prerequisite in the system of education for it has a considerable effect on the students' learning. Put it simply, Weideman (2019) described assessment literacy as the language teachers' consciousness, awareness, and knowledge of assessment. It is also used by Fulcher (2012) to refer to all the knowledge and skills stakeholders need to be able to deal with the matters related to the assessment world. However, Ng et al. (2018)claimed that teachers in general, both those in the pre-service phase of their work and the in-service ones, have a weak knowledge of assessment literacy and therefore, they cannot implement it effectively in their classes. That is to say, assessment is a key component in teaching practices, and it is in fact an effective and influential factor (Razavipour, 2013). He further stated that accountability, values, ethics, and policies in assessment are the very first important lessons included in assessment literacy knowing which is crucial to be literate, and the second is to do with their meaning and the way they are to be used. In the case of this first issue, McNamara (2006) has introduced the authors and journals that work on the two important matters of ethical issues and washback. He also explains that ethical testing practice includes three core domains of responsibility one of which is *accountability*, that is, a sense of responsibility to the people directly influenced by the test as well as those who use the information it offers. A second issue is about the effects testing has on teaching or washback. Finally, the third matter is related to the impact of a test outside the classroom. He is also asserted that those who believe language testing can be an ethical activity take either a broader or more restricted view of the ethics of testing.

Accordingly, a distinction can be made. In this distinction, the first class is called the *social responsibility view* and the second the *professional responsibility view*. Socially, responsible language testing concerns the social consequences of test use and says they are the responsibility of the test developer. According to the other approach, language testers should take responsibility for developing quality language tests (McNamara, 2006).

Assessment literacy is also defined by Weideman (2019a) in this way: it is the language teachers' knowledge of assessment and their attentiveness towards the point. In fact, he claims that in language testing, there has never been one comprehensive definition for the term validity.

Validity

Weideman (2019a) says that as the time passes, new ideas and concepts are added to the previous ones. Some of such concepts are those of "usefulness," "fairness," and "meaningfulness" all of which defined validity as relating to the interpretation of test scores.

However, there are wider meanings of validity according to what is claimed by Weideman (2019a). He introduced ethical considerations in language testing, such as treating those taking a test fairly, with care and empathy, and with due respect for the consequences of making the outcomes of the test which is known as another aspect of validity to be taken into account. The point led to the introduction of the concept of consequential validity. This concept in Guerrero'd (2000) idea can be best shown using an inclusive evaluative judgment of the measure.

In his articles, Weideman (2019a) as well as its follow-up (Weideman, 2019b), he claimed that his major aim was to discover one more alternative which is related to the several additions to and further interpretations of validity. Such a characterization of validity is a new one which needs a strong theoretical and conceptual framework since it has not yet been adequately developed in applied linguistics. His final word in his article is that it is likely that the disagreement about the concept of validity in the initial view which says it has influenced so many other disciplines is still rooted in the views of Messick and his followers.

The validity theory of Samuel Messick, according to McNamara (2006), is related to what he thought of as having great effects on language testing especially in two major ways: (1) it focuses on the ways inferences made have to be challenged, and (2) it concentrates on the consequences of test use. This latter point caused some debates on ethics, impact, accountability, and washback in language testing. In fact, Messick has situated his theory in the field of values. McNamara (2006) declared that in his validity framework, Messick proposed that testing by its nature can never be straight and direct.

Messick's framework contains explanation of test paradigm and test criterion. In this framework, it is suggested that an essential point in testing is the division between the test and the criterion domain. In the criterion domain, the relevant domain of behavior, knowledge, or skills related to which candidates' point of view can be established are referred to a lot (Bachman, Lyle, & Palmer, 1996).

McNamara (2006) further introduced Bachman as a great fan of Messick's framework whose 1990 book was deeply influenced by Messick. Such influence appeared to be rooted in three main points: (1) the criterion domain is obviously treated as a construct; (2) the test construct (i.e., what has to be stated about the individual test taker) is explained so that its relationship to the criterion construct is made clear; and (3) test method is preserved as a feature of the test content.

McNamara (2006) also talked about values and test constructs such as policy and performativity which some of the important points of each are provided below. Related to the policy aspect of values and test constructs, some implications are:

(1) The construct, which is considered as the heart of the test and the approaches of Messick's and Bachman's, is the outcome of political forces and not academic debates.
(2) The criterion construct is the basis for the definition of the individual ability.
(3) An essential emphasis is on the wording and rewording of the scale descriptors. That is, face validity is also a significant aspect of validity which has often been dismissed in discussions of validity.
(4) The role of the test instrument is neglected.

Furthermore, performativity, which is said to be derived from the word "perform" with the meaning of "Action" as the noun form, suggests that producing an utterance is the matter of performing an action. Performativity (i.e., the creation of a sense of something inner by certain acts) is distinguished from expression (the external exhibition of something inner). That is to say, expression goes from the inside to the outside which means that some inner principle is expressed. It is asserted in the same paper that the notion of performativity has several implications in the field of language testing. The first is that test constructs (e.g., communicative language ability) are performative accomplishments. Another implication is that it obliges us to ask what political governing and disciplinary performances are attained by the practice of testing (McNamara, 2006).

However, validity is an issue capable of being calculated. There are a variety of ways through which it can be estimated. Some of such methods are introduced by Fulcher (1997) which are listed below:

- “Correlation and principal components analysis” used to check the construct validity.
- “Analysing the cut scores across referred and non-referred students” to be able to establish a final cut score.
- Evaluating “concurrent validity” using the data gathered from 33 students.
- Asking subject specialists to comment on the questions set in the test for the sake of checking “content validity”.
- “Feedback from students” since the tests has consequences for the test takers and for the institutions whose decisions are based on the scores the give to their students.

Here are some of the scholars who worked on validity using various forms. In Youn’s (2020) study, separate consecutive organizations and interactional characteristics found in examinees’ levels functioned as a piece of evidence for critical validity in assessing learners’ interactional competence. Recently, “argument-based validity approach” has been a common way of assessing the validity of different tests using which Klebanov, Ramineni, Kaufer, Yeoh, and Ishizaki (2019) advanced an inquiry to check the validity of standardized writing tests, and Darabi Bazvand, Khorram, and Mirsalari (2018) carried out a comprehensive research in an attempt to develop and validate a collocational behavior test (CBT).

Reliability

Tommerdahl and Kilpatrick (2014) believed that reliability of the language samples, or the matter of how generally trustworthy data is, is considered as a vital point. They go on claiming that if a particular measure is said to be reliable, then it should provide similar results again and again when used in similar circumstances.

Reliability, or technical consistency of a language test in Weideman’s (2019a) word, is an integral part of validity, and some examples of its measures are Cronbach’s alpha and greatest lower bound which are introduced. Furthermore, reliability under classical test theory (CTT) is introduced by Tommerdahl and Kilpatrick (2014) as the most precise way of checking reliability. CTT is comprised of three major elements of true ability, measurement error, and the actual observed score of the participant for any given score, i.e., X = T + E. The observed score of the participant, called X, refers to the real score that a participant receives each time a test is administered. This score is composed of the true score and error score. The true score, named T, is the supposed mean score that the participant would get if they took the same test an endless number of times.

Reliability of a test can also be checked through *test-retest methodology*, in which individuals are tested twice under the same situations the results of which are then compared to define the degree of agreement between them. Two common types of agreement are relative and absolute. The former is defined as a way of ranking the participants in the same way whose scores may vary while in the latter, the test, and retest scores of an individual participant agree. The point is that it is this absolute agreement that reliability is interested in (Tommerdahl & Kilpatrick, 2014).

Intra-class correlation coefficient (ICC) is then presented as an accurate measure of the reliability between the language samples. As the main causes of random error in CTT, inconsistencies between raters, inconsistencies across different versions of an assessment or across difficulty levels of different items on an assessment, inconsistencies of occasion, or differences in time and place, are named. In fact, in the case of language samples, the same factors as mentioned earlier are said to determine whether a particular language sample is representative of a child's actual language competence or not (Tommerdahl & Kilpatrick, 2014).

In their paper, Schils, van Der Poel, and Weltens (1991) addressed some misconceptions related to test reliability which are:

1) It is shown that the significant differences which were observed among groups in the acceptable tests do not have the problem of low reliability coefficients.
2) It is discussed that test reliability should be determined in advance (or a priori, i.e., before a test is actually used in a research). That is to say, post-hoc calculations of test reliability are redundant and may even be ambiguous.
3) The results of computer replications show that reliability coefficients such as Cronbach's alpha is dependent to a large extent on the heterogeneity of the sample of participants and the range of item difficulties which causes serious limitations imposed on the usefulness of Cronbach's alpha as a reliability measure, especially when computed after the test is used in reality. However, Cronbach's alpha seems to be an inappropriate scalability measure.

Schils et al. (1991) also agreed with the point that there are various ways of calculating reliability. They went on stating that in some cases, we need to estimate the reliability of the test while in some others, the "reliability" of the difference between the two groups is needed. They also asserted that even in the case of test reliability, there are different ways of calculating the point which are test-retest correlation, parallel-forms correlation, corrected split-half correlation, or internal consistency-based approaches such as KR-20, KR-21, or Cronbach's alpha. Cronbach's alpha in their word is a scalability measure as well. That is, after the test taker tests the participants and finds out that there are implicational relationships among the test items, then the items can be scalable. Scalability is the extent to which a set of items is scalable (Schils et al., 1991).

To sum up, if one is keen on finding out something about the scalability of a set of items, it is better for them to use a simple measure like the average rank correlation than Cronbach's alpha. Besides, although more complicated, there are several scale models that can answer the question of scalability in a very useful way.

As a conclusion for the reliability issue, which is an important aspect of assessment literacy, it is considered as a prominent factor in the present research and taken into account as one of the four main components of teachers' assessment literacy for which some items are developed.

Interpretability of the results

An important point Weideman (2019a) talks about is the interpretability of the scores. He stated that without interpretation with a clear reference to the language ability being measured, language test scores are meaningless instead of meaningful. Therefore, tests that do not measure that ability would clearly have a strong basis. In support of their statement, Schils et al. (1991) asserted that the most central task of statistics in behavioral investigation is to take care of the correctness of the conclusions drawn against possible biases resulting from sampling and measurement error. Moreover, Tommerdahl and Kilpatrick (2014) declared that reliability information is a very important matter in the interpretation of test results.

Going back to Messick, he believed that testing can never be direct. It is, in fact, a procedure for deriving inferences about something which is not observable, and it is inevitably uncertain and indirect. However, the outcome is something obvious, and it is not the way only Messick thought of. That is to say, tests are processes for collecting evidence both for and against the interpretation that can be made tangible through scores (McNamara, 2006). Doing such a thing in his viewpoint is, in fact, a matter of validity checking which can help ensuring about the defensibility and justice in interpretations based on test performance in which the procedure is of utmost importance. That is to say, if the procedure is faulty, the inferences made about the individuals are not correct and sound.

On the other hand, the notion of consequential validity that has attracted a lot of attention and dispute because of its obvious practical implications is also defined as a matter of the interpretation of the results and the effects it has on the test takers' lives (McNamara, 2006).

Furthermore, interpretability of the scores, which is an important terminology related to the results of the tests, is described by Weideman (2019) as the degree of the easiness of the test which can be checked through interpreting the results obtained from a test when comparing with other administrations of the same test. As an instance, an average score of 50% is meaningless as a sort of magical "pass" due to the fact that any score needs interpretation. The point is that such numbers can give more information only if they are interpreted correctly and with reference to a fixed rule.

There is an important argument emphasized by Weideman (2019). He emphasized that diagnostic information backs instructional design while at the same time helps the teacher to recognize what should be highlighted in the succeeding language teaching sessions. The importance of this point is that such information will only be found if the assessment standard or interpretability is applied in the right way.

Efficiency

Stated by De Corte (2000), classification decisions relate to circumstances in which predictor tests are used to assign subjects to a number of different assignment positions. They are said to be important matters in all assessments. The problem, however, is to select the specific tests that make the most efficient classification possible, and this is the matter called efficiency of a test.

Allocation average, used for continuously measurable assignment criteria, shows the anticipated criterion score of the optimally assigned individuals. Therefore, the arrangement efficiency of a test battery is in fact the matter of the maximum possible allocation average that can be attained by using the battery. It also shows how the allocation average can be decided upon analytically when the predictors of inter-correlations and validities are identified. This calculation method is based on a formulation of the classification efficiency problem (De Corte, 2000).

The current study

The present study's basic focus was on the four main components of assessment literacy introduced here (i.e., validity, reliability, interpretability, and efficiency). In addition, the details elaborated on under each one were used to come up with the main points included in each component. The elaborations were essential since it was necessary to make both the researcher and subsequently the readers aware of the meaning of each component and the specific points each referred to.

Method

As it was mentioned earlier, the present research was carried out using a sequential exploratory mixed methods design. Accordingly, to be able to collect the necessary data for carrying out the present research within such a design, the following points were taken into accounts.

Participants

This study had two groups of participants. The first group was consisted of five female PhD candidates of TEFL studying and teaching in Karaj Islamic Azad University, Iran, with the following demographic features (Table 1).

The other group of participants (i.e., 150 ELTs) was also English teachers who are teaching in different institutes and universities. They are either students or graduates of different educational levels of English majors from BA to PhD. That is, they were six English translation BA graduates, seven MA students of TEFL, 125 MA graduates of TEFL, and 12 PhD candidates of TEFL. The age range of this second group was

Table 1 Demographic information of the first group of the participants

Participants	Age	Degree	Teaching experience
Participant 1	35	PhD candidate in TEFL	11 years
Participant 2	35	PhD candidate in TEFL	9 years
Participant 3	31	PhD candidate in TEFL	7 years
Participant 4	32	PhD candidate in TEFL	15 years
Participant 5	38	PhD candidate in TEFL	15 years

between 22 and 42 years old who had teaching experienced of 1 to 17 years. Another point which is worth mentioning is the sample size for the present research which was 150 in the piloting phase of the investigation. For the sake of running factor analysis, a sample size larger than 100 would suffice (Dornyei, 2007).

Materials and instruments

The instruments used to collect the necessary data were four questions at the first stage and a questionnaire consisted of 30 items in the second phase, which was developed by the researcher, and are both explained in detail in the following sections.

The five main questions

At the start of the study and at the same time as the researcher was busy reading the related literature, she also talked with five PhD candidates from each she asked the following five questions:

1. What is validity in your idea? Could you give me some examples of a valid test?
2. What is reliability in your opinion? Why you call a test reliable?
3. Is interpreting the results of a test important? Why?
4. How do you usually make sure that your interpretation of the tests is valid and reliable?
5. How do you make sure of the efficiency of the tests?

The questions were asked in a discussion-like session which the researcher held with the five participants which was also recorded. The researcher then listened to the recordings carefully and took notes of the key ideas and sentences within their talks.

The initial version of the questionnaire

After going through the literature related to the four main domains of assessment literacy, which are validity, reliability, interpretability of the results, and efficiency, and putting the outcomes derived from the literature together with the discussions the researcher had with the five PhD candidates regarding the same four matters, she came up with a questionnaire consisting of 30 items which were related to all the four fundamental components which were under the focus.

It has to be pointed out that some of the items derived from the PhD candidates' answers to the abovementioned questions. As an example, item 12, which is related to the reliability aspect of assessment literacy (i.e., a language test needs to be consistent, that is, it should provide similar results time and time again when used in similar circumstances), was the exact sentence uttered by the second interviewee. Moreover, some other items like that of 17 and 18 are derivations from the interviewees' comments and ideas provided on the basis of the questions asked.

Some other items, like item 27 which states "Tests are useful tools to make the most efficient classification possible" has the exact wording of the point found in literature while some others, item 8, for example, "Test method is a characteristic of test content", is a kind of simplified statement derived from the literature. The first version of the questionnaire was then developed containing four main components of validity, reliability, interpretability of the results, and efficiency each one with a number of items; that is, validity had 11 items, reliability had 5 items, interpretability of the results had 8 items, and efficiency had 6 items.

Furthermore, for the sake of validating the instrument, the researcher first asked three TEFL university lecturers to review the items to make sure they are appropriately worded and stated concisely. Then, to check the matter of internal reliability of the questionnaire just developed, the questionnaire was piloted with 70 participants, and the estimated reliability calculated through Cronbach's alpha level that was α = .62 which was a sign of an acceptable level of internal consistency of the test items.

Finally, the correction guidance based on which the instrument was scored is stated at the bottom of the questionnaire. That is, since the participants' answers to the items were numbered, the only thing the researcher was supposed to do was adding up the numbers the participants obtained through ticking in each box. The final score then was an illustration of the participants' assessment literacy (i.e., the higher the score, the more literate the participants were).

Procedure

First of all, a theoretical framework based on the four basic elements of reliability, validity, interpretability of results, and efficiency involved in assessment literacy and its components was developed through reviewing the literature and some discussions conducted with five TEFL PhD candidates. That is, at the beginning of the study, the researcher started reading the literature to find out the details related to the four main components of the study. She paid her utmost attention to different important points a literate teacher needed to know about assessment all through the time she was reading the literature and made detailed notes on them. A panel of experts who were associate professors of TEFL with an expertise in the field of testing, measurement, and assessment was consulted to establish a technical consensus on the constructs under study.

On the other hand, she conducted a kind of discussion with five PhD students who were all studying at the same university as the researcher asking them five influential questions regarding the main four components of the study to come up with other pieces of information which are important in their idea. These candidates were chosen due to the fact that they all passed very comprehensive language assessment and research courses in which almost all the validity, reliability, interpretability, and efficiency issues were covered and had good amount of information about these points.

Putting all the information obtained through the previous two phases, the researcher then developed 30 items on the basis of either the literature or the results of the points the interviewees talked about. It is also worth mentioning that in some items, the exact words provided in the literature or uttered by the interviewees were used as an item (Additional file 1). In the next stage, to validate the items of the newly developed assessment literacy scale, it was piloted with 100 ELTs, and the data gathered was subjected to principal component analysis which could reveal the number of components available in the questionnaire.

Referring back to the designs introduced by Best and Khan (2006), the current study was a sequential exploratory mixed-methods research that used both quantitative and qualitative ways of gathering and analyzing data the process for which can be summarized in the form of the following schematic representation derived from Creswell (2009):

According to Creswell (2009), he stated that a sequential exploratory design is a well-known design in mixed-methods design. He announced that although the two sets of data are separated, they are connected as the data in the first phase which informs the data in the second (Fig. 1).

Equational factor analysis, which is used in the quantitative phase of this study to analyze the data, is usually used to determine how many factors or constructs lie beneath a set of test scores and the extent to what extent these factors are correlated (Ockey, 2014). He stated that after factors have been recognized, a content analysis of the set of items which load on each factor is conducted using EFA to define the ability which is measured by the set of items. To run EFA, data from a study which aimed to identify the factors measured by an academic questionnaire are used.

Results

As it was mentioned earlier, the purpose of this study was to develop an instrument capable of assessing Iranian ELT's assessment literacy. The instrument consisted of the four components of validity, reliability, interpretability of the results, and efficiency was developed in two phases explained in detail in the "Procedure" section. In the initial development phase of the study, 11 items were generated for the first component, five items for the second, eight items for the third component, and six items for the last component. Therefore, there were 30 items included in the first draft of the instrument.

The so-called first draft was then piloted with 150 ELTs, and their answers were given to SPSS software to be analyzed subsequently. However, before going into factor analysis, which was the main analysis part of the research, the reliability of the questioner was checked through Cronbach's Alpha whose results show that the questionnaire bore an initial good reliability amount as the value reported for the point is .78 in Table 2 below.

The next formula to be run was that of factors analysis on the 30 items of the questionnaire considering the four components. The following three tables as well as Fig. 2 represent the upshots.

The Kaiser-Meyer-Olkin (KMO) test is a useful measure to decide about whether a data set is plausible to be used for a factor analysis or not. According to (Hinton, McMurray, & Brownlow, 2014), the rule here is that KMO values of .5 or higher mean that doing factor analysis is a suitable choice to be run. Keeping the explanation in mind, the data set collected in the current research is suitable since the KMO value is

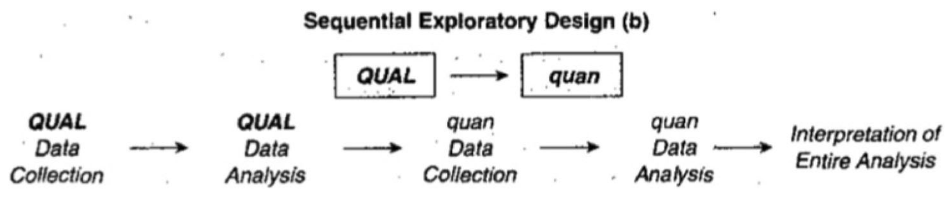

Fig. 1 Exploratory sequential MM design (Creswell, 2009, p. 209)

Table 2 Cronbach's alpha of the ELTs' Assessment Literacy Questionnaire

Reliability statistics		
Components of the Assessment Literacy Questionnaire	Cronbach's alpha	Number of items
ELTs' Assessment Literacy Questionnaire	.78	30

.56 which is higher than .5. Moreover, checking the Bartlett test's value, that is .00 and below the significant value of .05, makes it clear that running factor analysis is an appropriate way of analyzing data (Table 3).

As Hinton et al. (2014) stated, an eigenvalue of 1 means that the factor can explain as much variability in the data as a single original variable. As a result, the most plausible rule that can be used for determining if a factor is essential which is to only take those factors into account that have an eigenvalue of 1 or above. Therefore, it can be seen from the upshots provided in Table 4 that all the four components/factors have created eigenvalues bigger the so-called amount.

There is always a point on the scree plots where the eigenvalues stop fluctuating a lot. Then, factors up to this point are considered as significant factors and those after the point as not essential (Hinton et al., 2014). In the present study, four components with eigenvalues exceeding the similar criterion values were found.

As a rule of thumb, it is often said that a variable makes a significant contribution to a factor if the loading is 0.3 or greater (Hinton et al., 2014). Table 5 above shows the items loadings on the four abovementioned factors with six items loading above .3 on component 1, ten items loading above .3 on component 2, eight items loading above .3 on component 3, five items loading on component 4, and one item lower than this amount which is considered as an unsuitable item (i.e., item 20).

Based on the results obtained, the final ELT assessment literacy included the following four components and related items:

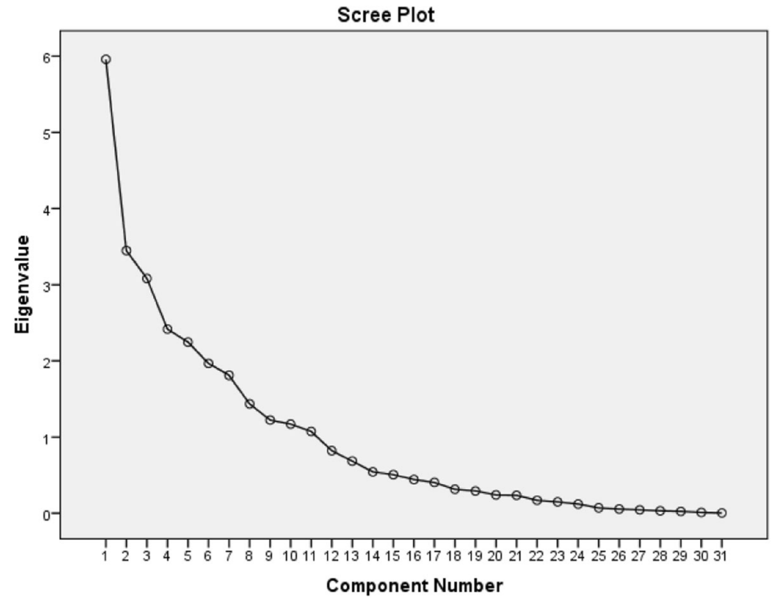

Fig. 2 Scree plot of the eigenvalues and the items of the questionnaire

Table 3 KMO and Bartlett's Test

Kaiser-Meyer-Olkin measure of sampling adequacy		.56
Bartlett's test of sphericity	Approx. chi-square	4524.73
	df	465
	Sig.	.00*

Component one: Validity which accounted for 19.21 of the total variance. This factor includes six items (4, 8, 12, 13, 19, and 28).

Component two: Reliability which accounted for 30.34 of the total variance. This factor includes ten items (5, 9, 10, 21, 22, 23, 25, 26, 27, and 30).

Component three: Interpretability of the results which accounted for 40.28 of the total variance and includes eight items (2, 11, 15, 16, 17, 18, 24, and 29).

Finally, component four: efficiency which accounted for 48.08 of the total variance which includes five items (1, 3, 6, 7, and 14).

Discussion

To start the discussion section, the authors' idea about the five questions asked about the four main constructs under the study in the first phase of data collection is presented. Regarding the first question which asked about the meaning of the validity and an example of a valid test, the researchers believed that for a test to be valid, it has to check the construct it has initially developed to check. If not, it cannot be called a valid test. IELTS, for instance, is such a valid test since taking the test, a test taker will be aware of his/ her true proficiency level in the four skills of listening, reading, writing, and speaking which is in fact the major reason for a person to take such a test. In response to the second question asking about reliability, in the authors' viewpoint, a test is called reliable in case it bears the same results if administered more than once with the same test taker since the examiner needs to be sure of the results obtained. The authors also believe that interpretation of the test results is a prominent point since it may have a great effect on the test takers' both present and future life, their emotions, their position in college, school, at work, etc. which is a matter related to the third question. Putting the two authors "ideas" together about how to make sure of the interpretations, they usually check them with another examiner who is somehow familiar with the test takers, or check the test takers' previous performance results to ensure of their validity and reliability. Finally, considering the last question about the efficiency of the tests, the authors

Table 4 Total variance explained

Total variance explained

Component	Initial eigenvalues			Extraction sums of squared loadings			Rotation rums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.95	19.21	19.21	5.95	19.21	19.21	4.47	14.44	14.44
2	3.44	11.12	30.34	3.44	11.12	30.34	4.37	14.12	28.57
3	3.08	9.94	40.28	3.08	9.94	40.28	3.37	10.87	39.44
4	2.41	7.79	48.08	2.41	7.79	48.08	2.67	8.64	48.08

Table 5 Rotated component matrix of the factor analysis of the items in the questionnaire

	Component			
	Validity	Reliability	Interpretability of results	Efficiency
Q4	.652			
Q8	.648			
Q12	.816			
Q13	.370			
Q19	.833			
Q28	.800			
Q5		.370		
Q9		.472		
Q10		.492		
Q21		.685		
Q22		.734		
Q23		.521		
Q25		.628		
Q26		.646		
Q27		.533		
Q30		.548		
Q2			.302	
Q11			.514	
Q15			.483	
Q16			.459	
Q17			.757	
Q18			.678	
Q24			.586	
Q29			.723	
Q1				− .679
Q3				.493
Q6				.367
Q7				.455
Q14				.713

try to adopt, adapt, or even develop different test based on the specific goals of any test that have the greatest classification ability.

Unfortunately, there is not enough work on the field of assessment literacy questionnaires which was the main gap existed in the literature that made the researcher to go through such investigations. However, following is just a few of the studies carried out which are somehow related to the same field, and their upshots can be compared in some way with the present one:

The only recent inventory developed in the field of assessment and testing is that of Beaudrie, Amezcua, and Loza (2019) who worked on developing a questionnaire with satisfactory psychometric properties to measure critical language awareness (CLA) in the Spanish heritage language (SHL) context since they believed it is indispensable for the students to receive instruction on both the heritage language as well as the contextual factors that affect the Spanish-English sociopolitical relationship. The final product of their study was a reliable and valid questionnaire, as the statistics showed, consisting of 19 items which are very useful in identifying change in the CLA of students in a class where CLA was taught.

An older questionnaire is that of Mertler and Campbell (2005) who developed an instrument called Classroom Assessment Literacy Inventory. The instrument has five components each one including seven questions which are parallel to the seven Standards for Teacher Competence in the Educational Assessment of Students. They are standards "connecting assessments to clear purposes; clarifying achievement expectations; applying proper assessment methods; developing quality assessment exercises and scoring criteria and sampling appropriately; avoiding bias in assessment; communicating effectively about student achievement; using assessment as an instructional intervention" (p. 7). These standards were in fact the basis of developing such an inventory the final version of which has 35 items.

Talking about the other investigations regarding the same domain, which is assessment literacy, it has in fact been the focus of a fair amount of studies focusing on different aspects of evaluation from among whom Checa-García and Guiberson (2019) investigated test validity in morphosyntactic measures. The main aim of this research was to show that differences among groups concerning a morphosyntactic measure which is used to identify specific language impairment (SLI) that cannot assure validity for diagnosis and tracking. The examination can be considered in the same domain as the current one in that it investigated test validity which is an aspect of the present research.

Additionally, Longabach and Peyton (2018) carried out an investigation on the reliability aspect of assessment in which they compared the reliability and precision of subscore reporting methods in the case of an English language proficiency test. The research can be considered in line with the current one in that it also tried to make the teachers aware of the importance of the reliability of the tests as well as introducing them some ways of calculating the point. By the way, at the end of their inquiry, they came to know that the reliability and precision of the two methods of CTT and UIRT, used to evaluate the reliability of the test, were almost similar.

With a very similar goal with that of this research, Kim et al. (2020) conducted a study to survey the assessment literacy essential for interpreting score reports. The results of their attempt provided some suggestions for improving the excellence of score reports. Some of such suggestions were those of clarifying technical terms, involving some information on student progress to enable the teachers monitoring students' language development, and reducing the time interval between the time of test administration and delivery of the scores.

However, regarding teachers' assessment, Atai, Babaii, and Taghipour Bazargani (2017) went through a series of steps to develop a questionnaire aiming at assessing Iranian EFL teachers' critical cultural awareness (CCA) lacking which was a clear gap in assessing such ability in teachers. As in the case of the present study, the researchers went through the very first phase of extensive reading of the literature as well as conducting some interviews with ELT experts to come up with some first draft kind of items. In the second step, they pilot the 37 items with a number of teachers who were available and willing to take part in the study. Component analysis of the data collected showed the presence of the three components of "CCA in ELT Programs" including 20 items, "CCA in ELT Textbooks and Materials" including 13 items, and "CCA in General Terms" including four items.

Instructional practices were the focusing matter of an investigation carried out by Razavipour and Rezagah (2018) to investigate the probable effect of a reform, introduced to the language assessment system in Iran, on English language teachers' assessment practices. The results showed that barriers such as managers, institutions, and individual features of the teachers are the main cause of the slow progress of the reform and the betterment of the classroom practices.

As it was mentioned earlier, although there is a good number of investigations conducted with a focus on assessment and assessment literacy, there was no specific work carried out to develop a questionnaire using which EFL teachers" assessment literacy can be evaluated. Therefore, inquiries such as the present will be very useful to enable those who are responsible for teachers' development to make sure of the amount of the teachers' knowledge considering the matter.

Conclusion

The current research was an attempt in order to come up with a new device (i.e., a questionnaire) using which it is possible to get to know the extent to which teachers are literate in main assessment matters of reliability, validity, interpretability of the results, and efficiency. That is to say, the final goal of this research was to develop and validate a questionnaire for assessing ELTs' assessment literacy. Going through a set of fixed steps, explained in detail in the procedure section, the researcher came up with a questionnaire containing four main components of assessment literacy named reliability with 14 items, validity with seven items, interpretability of the results with one items, and efficiency with three items.

The present questionnaire can be used as a useful tool in evaluating English language teachers' amount of knowledge in the area of assessment which is a crucial point in both deciding about their judgment accuracy and validity as well as thinking of a way to eliminate their weaknesses or even highlight their strengths. Therefore, the conclusion is that this questionnaire is a useful tool for the authorities who are working in the domain of English language teaching and testing. That is, they can make better decisions about what to include in teaching and testing programs or even materials. The other beneficiary group is the English language teachers who can find out their main problematic points when deciding about the scores they assign. Having an enhanced amount of assessment literacy can be considered a great help for them to tailor more variety into their classroom instruction and to make sure of their students' learning as well. As an example, teachers can think of peer observation as a way of increasing the validity of the assessments they go through.

Moreover, English students and learners are the other group whose benefits are even more for the results of the tests they take which may have a critical effect on their lives, and a fair assessment would be the exact thing they always wish for. Those who work on the course of teacher training are also the subjects of this investigation as it may help them think more deeply about the necessity of having such knowledge for language teachers which can lead to a major revision in the courses they provide both preservice as well as the in-service teachers of English.

A limitation of the study was that there was not a lot of people having enough knowledge of assessment especially the four principles of validity, reliability, interpretability, and efficiency of the results so that the researcher could have a semi-structured kind of interview with and have the opportunity to come up with more ideas working as the basis of the item development.

Lastly, the present study's output is merely the first form of a questionnaire developed to judge EFL teachers; assessment literacy which can be used as the starting point to be gone through by other researchers who desire to improve it. Or the questionnaire developed through this research can be used as an instrumentation to do a lot of other investigations in the domain of testing and assessment the upshots of which can yield a fair amount of data to be analyzed both qualitatively and quantitatively.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s40468-020-00106-1.

Additional file 1:. ELTs' Assessment Literacy Questionnaire

Abbreviations

TEFL: Teaching English as a foreign language; ELT: English language teacher; EFL: English as foreign language; CBT: Collocational behavior test; CCT: Classical test theory; ICC: Intra-class correlation coefficient; BA: Bachelor of Arts; MA: Master of Arts; PhD: Doctor of Philosophy; KMO: Kaiser-Meyer-Olkin; CLA: Critical language awareness; SHL: Spanish heritage language; SLI: Specific language impairment; CCA: Critical cultural awareness

Acknowledgements

The authors acknowledge all the faculty members and the Ph.D. candidates of English language teaching and translation of the researchers'institution.

Code availability

SPSS entry was used.

Informed consent

This research involves human participants, and informed consent was recognized and acknowledged in this research.

Authors' contributions

To achieve the purpose of the study which is designing and validating a potential inventory for assessing teacher assessment literacy, ZM conceived of the study and participated in its design and coordination and performed the statistical analysis and help the final draft of the manuscript. FN conceived of the study and participated in its design and data collection and help final draft of the study. Both authors read and approved the final manuscript.

Authors' information

ZM is an associate professor at English translation department of Islamic Azad University, Karaj branch, Karaj, Iran. She has published in the areas of discourse, interaction, and conversation analysis, teaching English as a foreign language and computer-assisted language learning. Currently, she is working on teacher education and development.

FM is a Ph.D. candidate and a lecturer at English translation department of Islamic Azad University, Karaj branch, Karaj, Iran. She has published in the area of assessment and educational evaluation in prestigious journals and attended national conferences on English language teaching

Funding

There is no funding for this research

Availability of data and materials

Data is available for submission if necessary.

Ethics approval and consent to participate

Research ethical issues were established and acknowledged in this research.

Consent for publication

Authors adhere to the Journals Copy Right and publication policies.

Competing interests

The authors declare that they have no competing interests.

Received: 20 March 2020 Accepted: 9 July 2020
Published online: 18 July 2020

References

Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching*, *35*(2), 79–113.

Atai, M. R., Babaii, E., & Taghipour Bazargani, D. (2017). Developing a questionnaire for assessing Iranian EFL teachers' critical cultural awareness (CCA). *Journal of Teaching Language Skills*, *36*(2), 1–38.

Bachman, L. F., Lyle, F., & Palmer, A. S. (1996). Language testing in practice: Designing and developing useful language tests (Vol I). Oxford University Press.

Beaudrie, S., Amezcua, A., & Loza, S. (2019). Critical language awareness for the heritage context: Development and validation of a measurement questionnaire. *Language Testing*, *36*(4), 573–594.

Best, J., & Khan, J. (2006). Research in education. United State: Pearson Education Press.

Brown, H. D., & Abeywickrama, P. (2018). Language assessment: Principles and classroom practices (3rd ed.). USA: Pearson Education.

Checa-García, I., & Guiberson, M. (2019). Test validity in morphosyntactic measures for typical and SLI incipient Spanish–English bilinguals. *Language Testing*, *36*(1), 77–100.

Coombe, C., Vafadar, H., & Mohebbi, H. (2020). Language assessment literacy: What do we need to learn, unlearn, and relearn? *Language Testing in Asia*, *10*(3), 1–16.

Creswell, J. W. (2009). *Research design: Qualitative and mixed methods approaches*. London: Sage Publications.

Darabi Bazvand, A., Khorram, A., & Mirsalari, S. A. (2018). Establishing an argument-based validity approach for a low-stake test of collocational behavior. *Journal of English Language Teaching and Learning*, *10*(22), 27–48.

De Corte, W. (2000). Estimating the classification efficiency of a test battery. *Educational and Psychological Measurement*, *60*(1), 73–85.

Deygers, B., & Malone, M. E. (2019). Language assessment literacy in university admission policies, or the dialogue that isn't. *Language Testing*, *36*(3), 347–368.

Dornyei, Z. (2007). Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies. Oxford University Press.

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. (1997). An English language placement test: issues in reliability and validity. *Language Testing*, *14*(2), 113–139.

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

Green, B. A. (2014). Program evaluation and language assessment. In A. J. Kunnan (Ed.). The companion to language assessment: Abilities, contexts and learners volume III (pp. 443-456). John Wiley & Sons, Inc.

Guerrero, M. D. (2000). The unified validity of the four skills exam: applying Messick's framework. *Language Testing*, *17*(4), 397–421.

Herppich, S., Praetorius, A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., … Klug, J. (2018). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education*, *76*, 181–193.

Hinton, P. R., McMurray, I., & Brownlow, C. (2014). *SPSS explained*. New York: Routledge.

Hougen, M. C. (2015). *Fundamentals of literacy instruction and assessment, 6-12*. Maryland: Paul H. Brookes Publishing Co., Inc.

Kim, A. A., Chapman, M., Kondo, A., & Wilmes, C. (2020). Examining the assessment literacy required for interpreting score reports: A focus on educators of K-12 English learners. *Language Testing*, *37*(1), 54–75.

Klebanov, B. B., Ramineni, C., Kaufer, D., Yeoh, P., & Ishizaki, S. (2019). Advancing the validity argument for standardized writing tests using quantitative rhetorical analysis. *Language Testing*, *36*(1), 125–144.

Longabach, T., & Peyton, V. (2018). A comparison of reliability and precision of subscore reporting methods for a state English language proficiency assessment. *Language Testing*, *35*(2), 297–317.

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly*, *3*(1), 31–51.

Mertler, C. A., & Campbell, C. (2005). Measuring teachers' knowledge & application of classroom assessment concepts: Development of the" assessment literacy inventory". Online Submission.

Ng, W. S., Xie, H., & Wang, F. L. (2018). Enhancing teacher assessment literacy using a blended deep learning approach. Paper presented at the International Conference on Blended Learning.

Ockey, G. J. (2014). Exploratory factor analysis and structural equation modeling. In A. J. Kunnan (Ed.). The companion to language assessment: Abilities, contexts and learners volume III (pp. 140-160). John Wiley & Sons, Inc.

Razavipour, K. (2013). Assessing assessment literacy: Insights from a high-stakes test. *Research in Applied Linguistics*, *4*(1), 111–131.

Razavipour, K., & Rezagah, K. (2018). Language assessment in the new. *English curriculum in Iran: managerial, institutional, and professional barriers*, *8*(9), 1–18.

Salimi, E. A., & Farsi, M. (2018). An Investigation of assessment literacy among native and nonnative English teachers. *Journal of English Language Teaching and Learning*, *10*(22), 49–62.

Schils, E., van Der Poel, M., & Weltens, B. (1991). The reliability ritual. *Language Testing*, *8*(2), 125–138.

Tommerdahl, J., & Kilpatrick, C. D. (2014). The reliability of morphological analyses in language samples. *Language Testing*, *31*(1), 3–18.

Weideman, A. (2019). Assessment literacy and the good language teacher: four principles and their applications. *Journal for Language Teaching*, *53*(1), 103–121.

Weideman, A. (2019a). Degrees of adequacy: the disclosure of levels of validity in language assessment. *Koers*, *84*(1), 1–15.

Weideman, A. (2019b). Validation and the further disclosures of language test design. *Koers*, *84*(1), 1–10.

Xu, Y. (2018). Assessment in the language classroom: teachers support student learning. *Language Assessment Quarterly*, *15*(4), 423–425.

Youn, S. J. (2020). Managing proposal sequences in role-play assessment: Validity evidence of interactional competence across levels. *Language Testing*, *37*(1), 76–106.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.