ATU PRESS

ISSUES IN LANGUAGE TEACHING (ILT)
VOL. 9, NO. 2, 347-375, December 2020
https://doi.org/10.22054/ilt.2021.57203.556

Pre-service and In-service Teachers' Knowledge and Practice of Assessment Literacy: A Dweller in an Ivory Tower

Mahmood Dehqan*
Assistant Professor of TEFL, University of Mazandaran, Babolsar, Iran

Seyyedeh Raheleh Asadian Sorkhi
M.A. in TEFL, University of Mazandaran, Babolsar, Iran

Received: September 5, 2020**; Accepted:** December 23, 2020

Abstract

Teacher assessment literacy plays a pivotal role in teacher education programs; however, there seems to be a lack of either assessment literacy or its implementation. Using an online assessment course, including both theoretical and practical issues, this mixed-method study examined 16 teachers' (8 in-service and 8 pre-service) assessment literacy and the extent to which they implement this knowledge. The quantitative part explored participants' assessment literacy, while the qualitative phase examined the validation of the quantitative results as well as the implementation of assessment literacy in the practical realm. Data were collected via valid and reliable questionnaires, one of which was adapted from Mertler (2003), and the two others were developed by the researchers, along with a practical assessment project. The results indicated that though in-service teachers were more assessment literate at the beginning due to their experience, they were at a lower degree of assessment literacy at their eventual performance in comparison with pre-service teachers. The qualitative analysis explored the lack of teachers' preference for the use of assessment literacy in their classroom practice. The study suggests the inclusion of both theoretical and practical dimensions of assessment literacy in teacher education programs and it proposes doing an in-depth investigation into the difficulties that hinder teachers from putting their theoretical assessment knowledge into practice.

***Keywords*:** Assessment literacy, Pre-service/In-service teachers, Assessment knowledge, Assessment literacy

***Corresponding author's email:** m.dehqan@umz.ac.ir

INTRODUCTION

Assessment literacy (AL) refers to instructors' professional knowledge of assessment, their skills in measurement, and their "understanding of assessment principles and practices" (Taylor, 2009, p. 24). Malone (2013) stressed the importance of the practical aspect of AL and delineated the principles and methods of applying assessment knowledge to classroom practices. Assessment can provide language teachers with information about students' performance and achievement. According to Malone (2013), there is a reciprocal relationship between teaching and assessment in a way that assessment can boost teaching and vice versa. Regarding the pivotal role of AL, Huang, Wang, and Wang (2007) included the notion of AL as a part of pedagogical content knowledge, a requiring base for teaching. As a result of this essential need for AL, different language testing textbooks have been published to facilitate the process. However, in order to develop their AL, language teachers need to have practical training consisting of relevant activities (Fulcher, 2012). Moreover, Boyles (2005), Inbar-Lourie (2008), and Taylor (2009) emphasized the importance of assessment knowledge not only for pre-service teachers but for in-service teachers in improving their ongoing practice.

Although there is agreement on the importance of AL training for both pre- and in-service teachers, there have been fewer inquiries to identify the extent to which each group is assessment literate. In addition, less emphasis has been put on the extent to which each group applies their theoretical AL to practice. Nor is it clear whether there are differences between each group of teachers' amount of AL and its practicality. Filling these gaps, this paper, through an online practical training course, aimed to recognize both groups' amount of AL, assessment practice, and the possible differences.

LITERATURE REVIEW

Teacher Assessment Literacy

Teacher professional development is mainly concerned with improving teachers' practices in the classroom, and AL training as a part of professional development projects enables language teachers to utilize sound and authentic assessment. The concept of AL, first proposed by Stiggins (1991), refers to teachers' capabilities to construct sound assessment which means to plan qualified tasks, to interpret the results appropriately, and to motivate students to actively take part in their learning process (Looney, Cumming, Kleij & Harris, 2017). Taylor (2009) identifies AL as the stakeholders' knowledge of measurement and the implementation of this knowledge. Mertler (2004), in his definition of AL, emphasizes the importance of sound assessment, evaluation, and communication practices as well as the use of assessment to increase students' motivation. In addition, The American Federation of Teachers, the National Council on Measurement in Education, and the National Education Association (1990) proposed AL as teachers' ability in:

1. Choosing appropriate assessment methods concerning the instructional decisions;
2. Developing appropriate assessment methods with regard to instructional decisions;
3. Administering, scoring, and interpreting the results of both standardized and teacher-made tests;
4. Making decisions about students, schools' improvement, and curriculum design via assessment results;
5. Grading students with regard to their assessment;
6. Communicating assessment results to stakeholders;
7. Recognizing inappropriate assessment methods.

Specifically concerned with language teaching and testing and with a

detailed review of language testing textbooks, Davis (2008) categorized the components of language testing as skills (the how-to expertise), knowledge (measurement and language expertise), and principles (concepts of testing such as validity and reliability).

Studies on Teacher Assessment Literacy

According to Ogan-Bekiroglu and Suzuk (2014), studies focused on teachers' AL can be divided into two broad categories: a) studies focused on teachers' AL and its incorporation into practice, and b) studies focused on the development of teachers' AL.

Studies on Teachers' Assessment Literacy and its Practicality

Despite the crucial role given to assessment in educational processes (Mertler, 2004; Popham, 2006; Stiggins & Chappuis, 2005), being an assessment literate practitioner is a far-fetched claim among teachers and testing experts as it is convinced by Mertler (1999), Mertler and Campbell (2005) and Popham (2009). For instance, some studies have shown the paucity of teachers' practical AL in spite of their theoretical awareness (Ogan-Bekiroglu & Suzuk, 2014; Seigel & Wissher, 2011). Moreover, Mertler (2004) and Mertler and Campbell (2005) documented teachers' lack of proficiency at the implementation of classroom assessment.

Similar to the conclusions of these studies, in 1991, Berg and Brouwer worked on teachers' knowledge of alternative assessment. At the theoretical level, they found teachers' familiarity with alternative assessment, yet in individual praxis, they recognized teachers' negligence in the implementation of it. Among the studies that focused on the lack of either theoretical or practical AL, the study conducted by Mertler and Campbell (2005) investigated the differences between pre-service and in-service teachers' AL and documented how in-service teachers are more literate than pre-service teachers due to their practical experiences. In this regard, they concluded that practical experiences seem to be more beneficial

than educational programs in the development of teacher AL.

Moreover, Wissehr and Siegel (2008) examined pre-service teachers' AL using their reflection. The teachers asserted that assessment should be at the service of learning, give students enough feedback and motivation and apply multiple methods to facilitate their learning. In their assessment planning project, however, they did not plan their tasks in accordance with what they had asserted in the theoretical realm. In addition, Siegel and Wissehr (2011) worked on pre-service teachers who believed in the alignment of assessment, instructional objectives, and strategies. Nevertheless, in the practical dimension, these teachers' assessment practice in science lessons did not align with what they had claimed as their theoretical beliefs. Ogan-Bekiroglu (2009) also found such discrepancies between theoretical and practical realms in teachers' AL. This study indicated that although teachers confirmed the merits of alternative assessment, they rarely used it in their assessment practice due to lack of time, lack of facilities, crowded classrooms, and what the curriculum obliged them to do.

Ogan-Bekiroglu and Suzuk (2014) also found such limitations in the practical realm of AL. They reported that although in the theoretical domain teachers tend to practice alternative assessment, which pertains more to a constructivist view of assessment rather than the traditional ones, and to give credence to *assessment for learning*, in the practical domain, they tend to practice assessment traditionally and to use *assessment of learning*. In the practical dimension of their study, pre-service teachers confessed that they might not use alternative assessment methods because of some limitations including lack of time and crowded classrooms. In their project assignment, given to pre-service teachers, they found that teachers rather prefer to teach and assess in a traditional way, the way they were taught and assessed. These findings led them to conclude the contradiction between pre-service teachers' knowledge and practice in assessment.

Studies on the Development of Teachers' Assessment Literacy

Boyles's study (2005) examined foreign language teachers' AL in the United States and recommended some training courses teachers should pass to achieve AL. It reported that language teachers should know how to use assessment, how to interpret and analyze the results, and how to use the results for their further teaching processes. It also suggested that teacher development should take place in different forms, happen in different contexts, and occur as part of teachers' educational programs. Mertler (2009), through a two-week classroom assessment workshop, also reported the promotion of in-service teachers' AL and foreshadowed further training needs for the implementation of both theoretical and practical domains of such assessment course.

Tsagari (2017) conducted a mixed-methods study to recognize language teachers' current level of AL as well as their assessment training needs across Europe. The results documented teachers' insufficient level of perceived AL and their inadequate educational programs in language testing and assessment that leads them to "use compensation strategies such as reliance on published assessment materials or the uncontested adaptation of mentors' and colleagues' assessment practice" (p. 54). The paucity of AL was specifically identified in traditional assessment formats teachers employed, the *deficit-oriented feedback procedures* they utilized, and the rare occasions when they used *alternative* forms of assessment.

The influential role of AL training was also emphasized by Deluca and Klinger (2010) who delineated that pre-service teachers' lack of assessment knowledge and confidence is due to the dearth of training courses in the assessment area. They found that pre-service teachers' level of confidence in assessment practice, theory, and philosophy were developed, though in different ranges, due to assessment training in a teacher education program. Their findings revealed that assessment training had more impact on teacher candidates' level of confidence with regard to assessment practice and theory rather than to assessment philosophy. Along

with previous inquiries, Smith, Hill, Cowie, and Glimor (2014) identified the influence of AL training offered to pre-service teachers and found that, as a result of the training program, teachers' beliefs were translated from summative into a formative assessment that could facilitate learning.

More specifically, concerning the paradigm shift in language teaching and learning, i.e. the shift from communicative language teaching towards intercultural tendencies, Scarino (2017) in a three-year study explored intercultural tendencies in language teaching and learning and attempted to develop language teachers' AL according to the sociocultural perspective. The results, which were based on a quadruple cycle including "conceptualizing; eliciting; judging and validating" (p. 24), indicated that in both eliciting and judging cycles teachers need to *(re)conceptualize* their intercultural understanding and their assessment process. The study emphasized that teachers should conceptualize intercultural language learning and learn how to operationalize its principles. They also should make a shift in their assessment process and move from traditional product-oriented assessment towards process-oriented one, so that they could assess students' sociocultural competence and interpret the results in a more inter-subjective mode of judging rather than the objective one.

Programs for the Development of Teacher Assessment Literacy

Over the past decades, scholars have emphasized the significance of teacher AL and the inclusion of AL courses as essential needs for teacher education programs (Mertler, 2004; Popham, 2006; Stiggins, 1991; Stiggins & Chappuis, 2005). Consequently, teacher assessment training has been developed by the emergence of some AL programs for both pre-service and in-service teachers including:

1. NAC (Lukin, Bandalos, Eckhout & Mickelson, 2004): Nebraska Assessment Cohort program, offered by the University of Nebraska Lincoln, consisted of three six-hour courses among which two six-hour courses were

offered in each of two consecutive summers and a practicum six-hour course was offered during the intervening school year. Teachers participated in all three courses with two opportunities of gaining knowledge about assessment and putting the knowledge and skills they had acquired into practice. The courses were web-based, offering teachers the opportunity to practice assessment literacy online.

2. ALLT (Arter & Busick, 2001; Stiggins, 2001): Assessment Literacy Learning Team was used to enhance AL through a *learning team*. There was a district-wide experience of AL piloted among four groups of teachers and administrators. The participants had the opportunity to discuss and share their successes and failures, thereby understanding the differences between sound and unsound assessment practices and learning how to use classroom assessment to improve instruction and to motivate students.

3. PALS (Lukin et al., 2004): Pre-service Assessment Literacy Study was the program provided for pre-service teachers who were not familiar with actual classroom assessment experience. PALS emerged on the basis of the ALLT model, in which pre-service teachers were paired with practicing teachers. There was a combination of AL training and classroom experience, in which practicing teachers were the leaders of teams.

4. IPALS (Lukin et al., 2004): In-service and Pre-service Assessment Literacy Study was a program similar to PALS but the difference was that both pre-service and in-service teachers were the members of the same AL learning team and the practicing teachers here were the fellow learners instead of the leaders.

5. Triple-A Model of Assessment Literacy (T. H. Wang, K. H. Wang, Wang, Huang & Chen, 2004): *Assembling, Administration,* and *Apprising* are the components of the Triple-A Model of Assessment Literacy which were emerged on the basis of systematic steps of test preparation offered by Miller, Linna, and Gronlund (2009). They stated that the preparation and use of classroom tests and assessment can give teachers valid evidence of students' learning. Hence, considering the learning outcomes, teachers should construct and assemble the tasks and items and

prepare the directions. Teachers then should administer the instruments and score students' responses and finally, they should be able to interpret the scores and apprise students of the results. They believe that only through these systematic steps, beginning with the recognition of instructional objectives and ending with test interpretation, one can measure students' achievement.

Following this systematic sequence, Wang et al. (2004) developed the Triple-A Model as equipment scaffolding the assessment process in their Web-based Assessment and Test Analysis (WATA) system. The content of the Triple-A Model included:

Assembling: teachers construct the questions and assemble the appropriate tests.

Administration: teachers arrange and administer the examination.

Apprising: teachers perform test analysis and item analysis after giving the test.

Following the two categories of research in AL proposed by Ogan-Bekiroglu and Suzuk (2014) and with regard to the NAC program (Lukin et al., 2004) and the Triple-A Model (Wang et al., 2004), this study, through an online treatment and a practical project, attempted to promote AL of both pre- and in-service language teachers, to identify the possible gaps between their theoretical and practical knowledge before and after the treatment and to recognize the degree to which each group utilizes its AL in classroom practice. In this regard, the following research questions are raised:

1. Are pre- and in-service teachers assessment literate before the treatment?
2. Do pre- and in-service teachers' assessment literacy improve as a result of the treatment?
3. In what areas of assessment literacy, do they still need training?
4. Do pre-service and in-service teachers prefer to put their theoretical knowledge of assessment into practice?
5. How do they put their assessment literacy into practice?

METHOD

Design

The present research was both quantitative and qualitative in nature and it was, therefore, designed based on Creswell's (2018) mixed-method convergent approach following a *side-by-side comparison* between quantitative and qualitative data. The main idea behind this design is to collect both quantitative and qualitative data on the basis of the same concepts so that the results inform whether there is convergence or divergence between the two sets of data. While the quantitative phase attempted to identify the participants' AL, the qualitative section tried to increase the validation of quantitative results and identify the participants' implementation of AL in classroom practice.

Participants

The participants of this research included 8 pre-service English teachers (5 females and 3 males) with the age range of 20-23 and 8 in-service English teachers (6 females and 2 males) who were English teachers for almost 2-4 years with the age range of 22-30. Both pre-service and in-service groups were TEFL students of the University of Mazandaran. The participants were referred to anonymously by using codes of "P" and "In" for pre-service and in-service teachers respectively (e.g. P-1 represents pre-service teacher one). Pre-service teachers were about to start their teaching experience, while in-service teachers had already some experience of teaching English either at private institutes or state schools.

Instrumentation

Classroom Assessment Language Inventory (CALI)

CALI (Mertler, 2003) was originally adapted from another questionnaire called the Teacher Assessment Literacy Questionnaire (Plake, 1993; Plake,

Impara & Fager, 1993) with a reasonable reliability index consisted of 35 multiple choice items measuring basic concepts of AL. Of the 35 items, 9 which were beyond the scope of this study were removed and only those items which measured the 6 basic AL concepts covered in this study were used. The remaining 26 items were piloted in a similar context, among 20 TEFL students of the University of Mazandaran, and the results have shown a reasonable reliability index (0.70), estimated through KR 20 method. The CALI measured 6 basic concepts of AL including content validity, reliability, dynamic assessment, test function, test kind, and washback. The number of items in each concept is listed in table1.

Table 1: Assessment concepts covered in CALI

Concepts	***Number***
Content validity	7
Reliability	8
Dynamic assessment	4
Test function	4
Test kind	2
Washback	1
Total	26

Open-ended questionnaires

The second data source included two series of questionnaires. The first consisted of 13 open-ended questions, some of which were developed by the authors of this research, and the others were adapted from Berkiroglue and Suzuk (2004). The questions were developed and categorized under the 6 dimensions covered in CALI in order to increase the validation of quantitative data derived from CALI. These questions were particularly used to compensate for the small number of participants and to validate the quantitative results. The second series consisted of 3 open-ended questions which were developed to identify teachers' preference for the feasibility of the theoretical issues as well as their struggles and views about the assessment projects. These 3 questions included:

1. Regarding your test construction projects, could you evaluate the advantages and disadvantages of the implementation of the theoretical issues covered in the tutorials?
2. Did you have any struggles while developing your project? Why?
3. Are you planning to use a different assessment during the course or a final one at the end of the course? Why?

It is worth mentioning that two faculty members of the University of Mazandaran, who were teaching testing and assessment courses for about 10 years, checked the validity of both series of questionnaires.

Data Collection Procedure

Following the NAC program, three tutorials were developed by the researchers. The tutorials, which consisted of texts, pictures, and audio files, were delivered online to both groups of participants separately with reasonable time intervals, and both groups were asked to do their assessment project based on the Triple-A model, mentioned in the literature, throughout the tutorials. Table 2 describes different sections of the tutorials and their brief description.

Table 2: Online tutorials' concepts

Concepts	***Description***
Definition	General information about testing and test scores, and the difference between measurement and evaluation
Test function	Test purposes that divide the tests into two types of prognostic and evaluation of attainment
Test construction	Several steps for making multiple choice items along with some guidelines for test construction
Item analysis	Identifying the characteristics of individual items consisting of item facility, item discrimination and choice distribution
Validation	Identifying the characteristics of items together containing reliability, validity and practicality

The research procedure consisted of the following steps: 1) To evaluate and compare their within- and between-group primary AL, the paper-and-pencil format of CALI was given to both groups of pre- and in-service teachers. 2) With regard to the Triple-A Model, the participants constructed a test, i.e. they practiced assembling. Their tests contained 10 multiple-choice items of vocabulary, grammar, and reading comprehension. 3) According to the NAC program, both groups of participants were provided with the first online tutorial including the primary, general concepts of AL such as the definition, test functions, and test construction. 4) After the first stage of assembling and based on the first tutorial, the participants revised their first version of the test and then practiced the second stage of the Triple-A model, i.e. administrating. Both groups of teachers administered their tests to students who had been taught the content of the tests before. 5) The second online tutorial, containing more specific concepts of AL, was given to both groups of participants separately. The content of the tutorial was mostly concerned with statistical analysis of items such as item facility and item discrimination. 6) After the second stage of the Triple-A Model (administering), and based on the concepts of the second tutorial, participants practiced the third stage, i.e. apprising. 7) Based on their item analysis, participants of both groups revised their tests for the second time and provided the third version of the tests. 8) With respect to the NAC program, the last online tutorial including the concepts of validation, i.e. reliability, validity, and practicality, was delivered to the participants. 9) At the end of the assessment project, participants were asked to calculate their tests' reliability index. 10) The paper-and-pencil test of CALI was administered as a post-test to estimate their possible development in AL within each group and to identify the possible difference of AL between the two groups. In addition, 4 participants from each group were selected randomly and were given the two series of questionnaires.

Data Analysis

The data were collected in a period of 5 months, beginning in April and ending in September 2018. The quantitative data were pre-test and post-test scores of CALI and because of the small number of participants, non-parametric statistical analyses (Mann-Whitney U test and Wilcoxon test) were used. For between-group comparison, the pre-test scores of CALI were used to investigate the difference in entry performance of pre- and in-service teachers, while the post-test scores of CALI were used to test both groups' differences in their eventual performance. In terms of within-group comparisons, the pre-test and post-test of each group were compared together to identify the possible development of each group separately.

The qualitative data, which were two sets of open-ended questions, were analyzed based on coding strategies. According to Creswell (2007), the general process for qualitative data analysis is the reduction of the data into themes and categories by coding procedures. This coding procedure can be organized with respect to either *a priori codes* or emergent categories. In order to validate the quantitative results, we used the *a priori* or predetermined codes, as Creswell (2018) named it *qualitative codebook,* which consisted of 6 concepts covered in CALI to have a *side-by-side comparison* between two sources of information. In addition, the coding procedure was "open to additional codes emerging during the analysis" (Creswell, 2007, p. 152). To recognize teachers' preference for the feasibility of AL, the responses of the second questionnaire were analyzed on the basis of the emerging coding system, opening to any new unexpected category. Moreover, from each group, four members' assessment projects were randomly selected to identify the extent to which they put their AL into practice.

RESULTS

Between-group Differences: Pre-test/Post-test of CALI at the Entry and Eventual Performance

To answer the first question of the study stating '*are pre- and in-service teachers assessment literate before the treatment?*', the descriptive statistics of both groups' entry and eventual performance are illustrated in table 3. The Mann-Whitney U test on pre-service and in-service teachers' average pre-test scores of CALI shows that the two groups are significantly different in AL in the entry with the p-value of .04 ($P < .05$) and the in-service teachers are at the higher level of AL with the mean average of 15. Both groups' post-test scores of CALI also indicate the groups' significant differences in AL in their eventual performance though this time, the pre-service teachers are confirmed to be more literate with the p-value of .00 ($P < .05$) and the mean average of 17 (table 4).

Table 3: Descriptive statistics of groups' entry and eventual performance

	Group	**N**	**Mean**	**Standard deviation**
Pretest	Pre-service	8	13	2.00
	In-service	8	15	2.00
Post-test	Pre-service	8	17	1.00
	In-service	8	16	1.00

Table 4: Mann-Whitney U test of groups' pre-test/post-test on CALI

	Pretest	Posttest
Mann-Whitney U	13.00	22.00
Wilcoxon W	49.00	58.00
Z	-2.01	-1.01
Asymp. Sig. (2-tailed)	.04	.00
Exact Sig. [2*(1-tailed Sig.)]	.05	.00
Exact Sig. (2-tailed)	.04	.00
Exact Sig. (1-tailed)	.02	.00
Point Probability	.00	.02

Within-group Differences: Pre-service/In-service Teachers' Performance in Time-line

The Wilcoxon test on both groups' pre- and post-test of CALI was used to answer the second question of the study: *Do pre- and in-service teachers' assessment literacy improve as a result of the treatment?* The result shows their significant progress across time with the p-value of .01 and .00, respectively. Tables 5 shows both groups' performance in the timeline.
As it is obvious, both groups improved in their AL as a result of training. The remarkable result here is that the in-service teachers move less towards progress rather than the pre-service ones.

Table 5: Wilcoxon test of Pre-service/In-service teachers' performance in timeline

Posttest - Pretest	Pre-service teachers	In-service teachers
Z	-2.00	.00
Asymp. Sig. (2-tailed)	.01	.00
Exact Sig. (2-tailed)	.01	.00
Exact Sig. (1-tailed)	.00	.00
Point Probability	.00	.06

Congruency between the Quantitative and Qualitative Results

To answer the third research question, the first set of open-ended questions was used to find how congruent the teachers' answers were with the quantitative results and to find *in what areas of assessment literacy, they still need training.* According to Creswell (2018), a *qualitative codebook* consisting of 6 concepts covered in CALI, i.e. content validity, reliability, dynamic assessment, test function, test kind, and washback, was the criterion for deductive, *a priori* coding system. The findings observed from the qualitative analyses were mostly congruent with the quantitative results:

1. Content validity: Both groups' answers with regard to the concept of

content validity illustrated the congruency between the quantitative and qualitative data:

- I do my best to consider the target lessons (P-3).
- Testing is to assess students according to the materials that were taught earlier (P-6).
- I usually test based on criterion reference (In-3).

2. Reliability: All the participants in both groups showed their literacy in the concept of reliability which again shows the congruency between qualitative and quantitative data:

- Fewer items would not identify students' knowledge properly (P-3).
- I try to design a guideline for scoring. I define a definite criterion (In-5).
- I try to have an ideal number of items, not very low and not very high (In-8).

3. Dynamic assessment: The participants in both groups demonstrated their literacy in dynamic assessment in both CALI and open-ended questionnaires:

- I will give them appropriate corrective feedback (In-3).
- I don't want them to pile up everything for their final exam (P-3).
- I would not base my scores only on one final exam and I would give multiple tests during the course (In-7).

4. Test function: All the participants in both groups mentioned the two main functions of tests (evaluation of attainment and the prognostic):

- Tests are held to evaluate the qualification of teaching and learning

(P-3).

- They are used to measure students' understanding of the course objectives and to make decisions for a person's future goals (p-7).
- It is used for many decisions. It is an evaluation of students' achievements (In-5).

5. Test kind: The qualitative analysis of the CALI revealed teachers' lack of knowledge in test kind that refers to the difference between large-scale, norm-reference tests (NRTs) and teacher-made, criterion-reference tests (CRTs). Likewise, the quantitative results indicated this paucity. Only P-3 who answered a question of the CALI in test kind area correctly answered the same in the qualitative questionnaire:

If the instruction was provided by the teacher, a teacher-made test would be suitable.

However, some incompatibilities were found between the quantitative and qualitative data. For instance, two participants, who did not answer the questions of the CALI in this area, revealed some clue about the test kind in the open-ended questions, though they did not know the exact concept:

- I use teacher-made tests because standardized tests are used for large-scale testing. I rarely use standardized tests for my class (In-8).

This means that most of the teachers in both groups were still in need of training in the concept of the difference between NRTs and CRTs.

6. Washback: The in-service teachers' post-test of CALI indicated their progress as a result of training. This result was also congruent with their qualitative answers:

- I usually teach based on the material, not the test (In-3).
- We should not forget that tests are at the service of learning and

teaching (In-7).

- A teacher should not teach for testing (In-8).

However, the pre-service teachers did not show any awareness in washback. In addition, only one of those two pre-service teachers, who answered the question in the CALI correctly, revealed congruent result in her qualitative answer:

- I would put a bit more emphasis on the contents of the final exam, but I will never reveal the exact questions (P-3).

The result of this part documented that both pre-service and in-service teachers were still in need of detailed instruction in the washback area. According to the participants' responses to the first qualitative questionnaire, it seems that almost all the findings of the quantitative phase are congruent with their qualitative counterparts. Therefore, it can be said that both groups of teachers are assessment literate in the concepts of the test function, content validity, reliability, and dynamic assessment. In addition, both groups of teachers, especially pre-service ones, lack the knowledge of the concept of washback as well as the knowledge of differences between test kinds, i.e. NRTs and CRTs.

Assessment Literate Teachers' Preference for the Feasibility of the Theoretical Issues

The second set of open-ended questions was designed to answer the fourth question and to find *if pre-service and in-service teachers prefer to put their theoretical knowledge of assessment into practice*. The inductive, emerging coding system identifies two main themes, each of which is elaborated below:

1. Time is money

Almost all of the in-service teachers mentioned that they prefer not to use their theoretical knowledge of assessment in their class because of some reasons:

- Calculating the reliability of the test is not easy and most of the methods are time-consuming (In-8).
- Many methods are not practical and are not even feasible in the given situation (In- 7).
- Finding the reliability index is very helpful but time-consuming and because of the lack of payment and time teachers don't consider them (In-3).
- On the contrary, pre-service teachers seem to agree with the feasibility of the theoretical issues:
- This really helps an inexperienced to be a teacher. It covers interesting methods that help teachers (P-7).
- They opened my eyes to testing (P-7).

Overall, it seems that pre-service teachers are rather optimistic about the feasibility of the theoretical knowledge in assessment since they are not still in the picture! They are not still involved in the teaching and testing process and the possible difficulties that in-service teachers mention such as the lack of time and payment.

2. Teacher is marginalized

While In-service teachers emphasized the lack of time, pre-service teachers focused on their own authority in assessment practices, as one of them mentioned her worry over the value of the test she had constructed:

- I was worried about the result of the test because I was sure that students wouldn't take the test seriously. They were sure that I could do nothing with their scores since it was not required by the institute (P-3).

  As P-3 mentioned, it is obvious that teachers prefer not to use their

own teacher-made tests, constructed according to the theoretical issues, because of the institutional policies. It seems that teachers' AL is not applicable to the educational context and their identities as teachers are marginalized.

Moreover, all participants preferred to give different tests across the course and to apply dynamic assessment as an important facet in the language learning process, yet it seemed that in some cases the testing procedure including the content, the time, and scoring was predetermined by the school principal or the educational policy. For instance, one of the participants mentioned the limitations derived from the school's policy:

- I want to use different things, but I cannot since teachers are restricted and the procedure for doing this is pre-determined in our schools (In-7).

Implementation of Teacher' Assessment Literacy into Practice

At the endpoint, the participants' assessment projects were analyzed to answer the last question which is to find *how they incorporate their AL into practice*. Out of 8 participants, only one (In-3) revised her test based on the guidelines in the tutorials and item analysis. In her revised version, she described all aspects of item analysis, the index of item discrimination, item facility, and the choice distribution of each alternative, in detail.

Others' revised test versions were still problematic in some areas. For instance, In-7 identified the problematic items after item analysis, but in his revised version he just omitted those items without making any change.

After writing her test three times, In-5 still had some problematic items in those areas which were mentioned in the first tutorial:

- All of the alternatives should be grammatically correct (tutorial 1).
- In the same item, the alternative should be of similar length,

difficulty, and type of grammatical structure (tutorial 1).

However, in her revised version, she did not implement the guidelines. In one item, one of her alternatives was grammatically wrong, and in another one, she used four verbs as the alternatives of which one of them was a phrasal verb and the others were simple.

In her revised test version, In-8 also had a problematic item. She did not incorporate the guidelines into her practice since in an item one of the alternatives (the answer) was an adjective and the others (distractors) were a noun.

P-2 analyzed her test and identified the problematic items; however, in her revised version she did not incorporate the theoretical issues mentioned in tutorials. In addition, in her last revised version, she did not consider the guidelines of the tutorials. For instance, in the first tutorial, it was mentioned that avoid using general statements. However, she used general knowledge in one of her stems. Moreover, in another item, two of her alternatives had the same meaning. Other guidelines of the first tutorial were:

- Avoid using non-of-the-above or all-of-the-above alternatives
- Avoid using negative statements or double-negation structures because they are likely to be overlooked. In unavoidable cases, the negative marks should be bold, underlined, or capitalized.

However, in his revised test version, P-3 used negative words in two stems without underlining or bolding. P-7 also had a grammatically wrong item in her last revised version.

Overall, after three tutorials containing some theoretical issues about assessment, both groups of teachers enhanced their AL, though in the majority of cases they did not put them into practice.

DISCUSSION

This study tried to shed light on the possible differences between pre-service and in-service teachers' AL, their promotion in this area as a result of training, their further needs in this stage of teacher development, and their possible differences in the practical implementation of this knowledge.

The study revealed that in-service teachers are at a higher level of AL regardless of any training course. This higher degree of AL might be due to in-service teachers' degree of experience in teaching and assessment in comparison with that of pre-service teachers. This incipient finding seems to be in line with Mertler's (2004) study emphasizing that teachers learn more about assessment from the practical realm than from the theoretical one.

Previous studies have emphasized the importance of AL in teacher education programs (Mertler, 2004; Popham, 2006; Wang et al., 2007). Moreover, many scholars have argued the importance of the practical dimension apart from its theoretical side (Fulcher, 2012; Malone, 2013), and others such as Boyles (2005), Inbar-Lourie (2008), and Taylor (2009) have focused on the significance of AL as a need for both pre-service and in-service teachers. Following these emphases, this study through practical training for both pre-service and in-service teachers attempted to enhance teachers' AL and to identify their further needs. The results show that an online assessment course containing some theoretical issues along with the Triple-A Model of practice could be beneficial for teachers.

In addition, both quantitative and qualitative phases of the study indicated that all teachers, especially pre-service ones, are still in need of detailed instruction in washback as well as in the concept of test kind, i.e. NRTs versus CRTs. Hence, these areas of knowledge in AL are recognized as further training needs.

The unique consideration of this study could be the findings pertinent to the second research question, pointing to the extent to which each group improves in AL. Although the participants' entry performance

documented in-service teachers' higher level of AL in theoretical dimension due to their greater degree of practice, which was also demonstrated by Mertler (2004), the eventual investigation of both groups' performance revealed that pre-service teachers are more assessment literate as a result of training. The possible reason for this finding could be that pre-service teachers were more eager to achieve a greater deal of AL or they might pay more attention to the tutorial content since they are not aware of the real context of teaching and testing, and therefore they may not be in the picture. They may not be familiar with those problematic issues in the assessment area such as lack of time, lack of wage, crowded classroom, and the obligations offered by the educational system to obey the predetermined rules and administer the predetermined tests, and hence their post-test performance was better than that of in-service teachers. Therefore, regardless of both groups' progress as a result of an assessment course as is also documented by Smith et al. (2014), it seems that the course was more beneficial for pre-service teachers rather than the in-service ones.

Another importance of the current research could be the qualitative findings related to participants' preference and their implementation of AL. The study evinces a fundamental contradiction between the theoretical and practical realms and it shows that practice, though helpful for teachers to enhance their AL, could be a barrier that hinders further developments. This contradiction between theoretical and practical dimensions is in line with the research conducted by Ogan-Bekiroglu and Suzuk (2014). In the theoretical investigation of pre-service teachers' assessment literacy, Ogan-Bekiroglu and Suzuk (2014) identified teachers' tendency towards alternative assessment, yet in practical dimension, they found out teachers' repulsion in applying alternative assessment in their own classroom due to some restrictions such as lack of time, crowded classroom, and their traditional manner of teaching and assessment that pushed them to teach and assess in a way that they were taught and assessed. The possible reasons for the contradictory findings of this study, i.e. the difference between the theoretical and practical realm of AL, could be little time available for

teachers, financial issues, teachers' marginalized identity, and the dominance of the educational system over teachers' decision.

CONCLUSION AND IMPLICATIONS

This study suggests the inclusion of assessment course in teacher education programs that contains not only the theoretical issues but some practical opportunities for test construction, enabling teachers to understand their strengths and weaknesses in the assessment. The study also suggests that being more engaged with assessment does not necessarily transform the theoretical knowledge into practice. It finds practice as a barrier to an in-depth study of assessment especially for in-service teachers who are involved in teaching and its difficulties. Some of the reasons that prevent teachers from improving their AL could be the lack of time, the lack of payment, and the restrictions due to the educational policy.

Regarding the abovementioned problems, the study suggests that the theoretical dimension of AL lives in the ivory tower and that more considerations should be given to the practical realm of assessment. The significant contribution of this study to the field of AL is to help educate policymakers, specifically in the realm of language teaching and assessment, to change or establish principles that give language teachers a greater degree of authority in their assessment practice so that they could bring their AL to practical dimension.

One of the limitations of this study could be the probable unequal levels of AL of pre- and in-service teachers due to the latter group's authentic practice of assessment. Regarding the findings of the present research, further studies are suggested to investigate the possible reasons for the discrepancies between the theoretical and practical realms in assessment, namely some contextual factors. More attention also needs to be paid to ways of overcoming the difficulties in the practical realm so that teachers could apply their assessment literacy to their classrooms. A large-scale, longitudinal research might be needed to have an in-depth investigation of

teachers' incorporation of AL into practice. In addition, further studies need to be done on other aspects of AL (e.g. self/peer assessment, portfolio assessment) and their feasibility.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Mahmoud Dehqan http://orcid.org/0000-0003-4344-7307
Seyyedeh Raheleh Asadian Sorkhi http://orcid.org/0000-0001-5291-4873

References

American Federation of Teachers, National Council on Measurement in Education, and National Education Association (1990). Standard for teacher competence in educational assessment of students. *Educational Measurement: Issues and Practice*, *9*(4), 1990, 30-32. Retrieved from http://www.unl.edu/buros/bimm/html/article3.html

Arter, J. A. & Busick, K. U. (2001). *Practice with student-involved classroom assessment: A workbook and learning team guide.* Portland, OR: Assessment Testing Institute.

Berg, T., & Brouwer, W. (1991). Teacher awareness of students' alternative conceptions about rational motion and gravity. *Journal of Research in Science Teaching*, *28*(1), 3–18.

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

Creswell, J. W. (2007). *Qualitative inquiry and research design* (2nd ed.). Thousand Oaks, CA: Sage.

Creswell, J. W. (2018). *Research design: Qualitative, quantitative, and mixed methods approach* (5thed.). Thousand Oaks, CA: Sage.

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

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

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

Huang, S. C., Wang, K. H., & Wang, T. H. (2007). Designing a web-based assessment environment for improving pre-service teacher assessment literacy. *Computers & Education, 51*, 448-462.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing, 25*, 328-402.

Looney, A., Cumming, J., Kleij, F. V. D. & Harris, K. (2017). Reconceptualising the role of teachers as assessors: Teacher assessment identity. *Assessment in Education: Principles, Policy & Practice*, *25*(5), 442-467.

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

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

Mertler, C. A. (1999). Assessing student performance: A descriptive study of the classroom assessment practices of Ohio teachers. *Education, 120*, 285-296.

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

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

Mertler, C. A., & Campbell, C. (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, Quebec, Canada.

Mertler, C. A. (2009). Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. *Improving Schools, 12*, 101-113.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching (*10th ed.). Upper Saddle River, NJ: Pearson Education.

Ogan-Bekiroglu, F. (2009). Assessing assessment: Examination of pre-service physics teachers' attitudes towards assessment and factors affecting their attitudes. *International Journal of Science Education, 31*(1), 1-39.

Ogan-Bekiroglu, F., & Suzuk, E. (2014). Pre-service teachers' assessment literacy and its implementation into practice. *The Curriculum Journal*, *25*(3), 344-371.

Plake, B. S. (1993). Teacher assessment literacy: Teachers' competencies in the educational assessment of students. *Mid-western Educational Researcher, 6*(1), 21-27.

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

Popham, W. J. (2006). Needed: A dose of assessment literacy. *Educational Leadership, 63*(6), 84-85.

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

Scarino, A. (2017). Developing assessment literacy of teachers of languages: A conceptual and interpretive challenge. *Papers in Language Testing and Assessment, 6*(1), 41-63.

Siegel, M. A., & Wissher, C. (2011). Preparing for the plunge: Preservice teachers' assessment literacy. *Journal of Science Teacher Education, 22*(4), 371-391.

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

Stiggins, R. (2001). *Student-involved classroom assessment (3rd ed.).* Upper Saddle River, NJ: Prentice Hall.

Stiggins, R. J., & Chappuis, J. (2005). Using student-involved classroom assessment to close achievement gaps. *Theory into Practice, 44*(1), 11-18.

Smith, L. F., Hill, M. F., Cowie, B., & Gilmore, A. (2014). Preparing teachers to use the enabling power of assessment. In C. Wyatt-Smith, V. Klenowski, & P. Colbert (Eds.), *Designing assessment for quality learning* (pp. 303–323). Dordrecht: Springer.

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

Tsagari, D. (2017). Assessment literacy of foreign language teachers around Europe: Research, challenges and future prospects. *Papers in Language Testing and Assessment, 6*(1), 41-63.

Wang, T. H., Wang, K. H., Wang, W. L., Huang, S. C., & Chen, S. Y. (2004). Web-based assessment and test analysis (WATA) system: Development and evaluation. *Journal of Computer Assisted Learning, 20*(1), 59-71.

Wissehr, C., & Siegel, M. A. (2008). *Unlocking assessment secrets: What are pre-service teachers' views of assessment?* Paper presented at the annual meeting of the Associations for Science Teacher Education, St. Louis, MO.