Iranian Journal of Applied Linguistics (*IJAL*)

Vol.19, No.2, September 2016, 155-193

Towards a Task-Based Assessment of Professional Competencies

Gholam Reza Kiany*, *Tarbiat Modares Univesity, Tehran, Iran*

Monireh Norouzi, *Tarbiat Modares Univesity, Tehran, Iran*

Abstract

Performance assessment is exceedingly considered a key concept in teacher education programs worldwide. Accordingly, in Iran, a national assessment system was proposed by Farhangian University to assess the professional competencies of its ELT graduates. The concerns regarding the validity and authenticity of traditional measures of teachers' competencies have motivated us to devise a localized performance assessment scheme**.** Therefore, the present study aimed to develop a performance assessment scheme to be used as a benchmark for assessing the professional competencies of ELT graduates of this university. To this end, three assessment tasks and rating scales were developed, piloted, and administered. Next, Haertel's participatory approach was employed to set passing standards for the assessment tasks as well as the whole assessment scheme. Analysis of the data revealed inter-rater and intra-rater reliability coefficients of 0.85 and 0.89. The validity of the assessment scheme was also confirmed by experts' judgments made, to a large extent, on the correspondence between the target domain and test domain skills. Based on the results, the proposed assessment scheme is rendered more efficient and reliable in comparison to traditional tests with regard to the following dimensions: a) higher degrees of reliability and validity of the assessment scheme aimed at the improvement of licensure and program development; b) stronger evidence for inter-/intra- rater reliability and consistency of scoring; and c) an optimized and systematic procedure for setting passing standards based on the consensus of experts' judgments. It is believed that further development of the proposed assessment scheme unlocks its potential to be used as a large-scale teacher assessment model for Farhangian University.

Key words: performance assessment, teacher evaluation, Farhangian University

Article Information:

Received: 20 February 2016 **Revised: 25 July 2016** **Accepted: 25 August 2016**

Corresponding author: Department of foreign languages, Tarbiat Modares University, Tehran, Iran Email address: rezakiany@yahoo.com

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

1. Introduction

The profound effect of teachers on quality education and learners' academic achievement (İşlek & Hürsen, 2014; Sanders, Wright, & Horn, 1997) brings to the forefront the need to evaluate teachers' professional competencies and, consequently, the need to use effective teacher evaluation methods (Marshall, 2009; Medley, 1982). Teacher evaluation, as defined by Shinkfield & Stufflebeam (1995), refers to "the systematic assessment of a teacher's performance and/ or qualifications in relation to the teacher's defined professional role and the school district mission" (p. 86).

The need for teacher evaluation is acknowledged by a number of researchers (e.g., Aseltine, Faryniarz, & Rigazio-DiGilio, 2006; Marshall, 2009; Sergiovanni & Starrat, 2002). Sergiovanni and Starrat (2002), for instance, argued that teacher evaluation contributes to the academic achievement of students by developing teachers' instructional capacity. In a similar vein, Marshall (2009) perceives the theory of action behind supervision and evaluation as follows:

> The engine that drives high student achievement is teacher teams working collaboratively toward common curriculum expectations and using interim assessments to continuously improve teaching and attend to students who are not successful. (p. 731).

In light of the potential impact of teacher evaluation on teachers, learners, and quality education, many state departments of education all around the world have established policies aimed at incorporating teacher evaluation in teacher education (Arends, 2006). Similarly, in Iran, Farhangian University, established in 2011 to educate competent teachers for the Ministry of Education, has set up a national project named *ASLAH* (Evaluation of Professional Competencies) to assess the professional competencies of its ELT student-teachers. For this purpose, performance assessment, written assessment, portfolio, and GPA are used as the criteria for evaluation. Since

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

the above-mentioned project has significant consequences, considerable effort and expertise are required to be invested in the development and implementation of each of its components so that it can fulfil its true potential. Therefore, the present study aims to provide a pilot scheme for the performance assessment and to investigate the validity and reliability of the developed assessment scheme.

2. Review of the Related Literature

In the current era of accountability and testing, high-stakes assessments are considered virtually indispensable for any educational reform movement (Koirala, Davis, & Johnson, 2008). They provide benchmarks for accountability purposes and teacher credentialing and introduce changes in educational practices (Hamilton 2003). High-stakes assessments, and in particular, performance-based assessments, are becoming increasingly popular among teacher education programs for their benefits for teacher learning, teaching quality, and student achievement (Danielson & Marquez, 1998; Delandshere & Arens, 2003).

Research evidence documenting the role of performance-based assessment in quality education has highlighted the role of teacher evaluation in teachers' professional development, learners' academic achievement, and quality assurance (Sanders et al., 1997). As Darling-Hammond (2010) puts it, performance-based assessment, due to its potential to provide contextualized evidence of student learning, uncovers the essential role of teachers in predicting learners' academic achievement. In addition, what adds to the significance of performance-based assessment concerns its positive influence on teachers (Sandholtz, 2012). Darling-Hammond (2010), for instance, refers to the role of performance assessment in increasing teachers' subject matter knowledge, improving classroom management, and designing instruction.

This recognition of the potential advantages of performance-based assessment has led to the application of some teacher evaluation methods, such as the California Teaching Performance Assessment (CalTPA), the

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

Performance Assessment for California Teachers (PACT), and the Fresno Assessment of Student Teachers (FAST) in evaluating teachers' professional competencies prior to awarding them teaching credentials. What follows is a brief explanation of the aforementioned teacher evaluation methods.

In response to Senate Bill 2042, the California Teaching Performance Assessment (CalTPA) was developed by the California Commission on Teacher Credentialing (CTC) and the Educational Testing Service (ETS) along with experienced California educators to evaluate the mastery of Teacher Performance Expectations (TPEs) (Torgerson, Macy, Beare, & Tanner, 2009). TPEs are "the foundational sets of knowledge, skills, and abilities on which all of California's multiple routes to earning a credential are based" (California Commission on Teacher Credentialing, 2016, p.2). The CalTPA incorporates four tasks, each of which measures the TPEs in multiple ways. The tasks include: *Subject-Specific Pedagogy, Designing Instruction, Assessing Learning, and Culminating Teaching Experience.* Using a multilevel task-specific rubric, trained and qualified professionals assess each candidate's performance on these tasks.

As an alternative to CalTPA, the Performance Assessment for California Teachers (PACT) was developed by preservice teacher preparation programs across California in 2002 to measure effective teaching at the preservice level (Chung, 2008). The PACT program contains two main components: a formative evaluation and a summative assessment (Sandholtz, 2012). While formative evaluation is "based on embedded signature assessments that are developed by local teacher education programs, a summative assessment is based on a capstone teaching event" (Sandholtz, 2012, p. 106)**.** Course-embedded signature assessments are used to assess teacher competency, and the teaching events (TEs) are subject-specific portfolios of teaching "designed to measure and promote candidates' abilities to integrate their knowledge of content, students, and instructional context in making instructional decisions and reflecting on practice" (Pecheone & Chung, 2006, p. 5). In other words, they evaluate a candidate's competency in five domains: planning, instruction, assessment, reflection, and academic

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

language through five integrated tasks. The candidates' performances are scored by trained raters, mainly faculty members and supervisors within teacher preparation programs, on the basis of 12 four-level rubrics divided by tasks (Stewart, Scalzo, Merino, & Nilsen, 2015).

While CalTPA and PACT have been approved to be used by teacher education programs across California, FAST is a locally designed teacher performance assessment system that is specific to one California State University (CSU) campus, Fresno State. FAST is characterized by Teacher Work Sample (TWS). TWS is "a TPA that provides evidence of a student-teacher's ability to meet state and national teaching standards while providing feedback in a form that allows for continuous program improvement" (Torgerson et al., 2009, p. 65) and "requires the teacher candidate to systematically connect teaching and learning" (Girod & Girod, 2008, p. 309). FAST allows for both formative and summative evaluations of the pedagogical competencies of candidates with respect to thirteen TPEs through the following four tasks: *Comprehensive Lesson Plan Project*, *Site Visitation Project*, *Holistic Proficiency Project*, and *Teaching Sample Project*. In addition to designing tasks to measure each TPE twice, a task specific rubric was also developed by Beginning Teacher Support and Assessment (BTSA) program coordinators, content area faculty members, and supervising teachers from local districts to define the elements of the TPEs in a qualitative manner (Torgerson et al., 2009).

The above-mentioned performance-based assessments and other preservice performance evaluations required for licensure, set out to measure teacher candidates' competencies, are increasingly used in teacher education programs. The appeal of such assessments to the research community and teacher preparation programs highlights, as stated earlier, their positive consequences for teaching and learning. Performance-based assessment schemes provide insights into improving teacher education programs as well (Darling-Hammond, 2010). Given the potential benefits of performance-based assessment schemes, it is not surprising that the quest for more appropriate schemes is a continuous process.

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

Despite the considerable potential of performance-based assessment for teacher education, the current teacher evaluation methods used in the Iranian educational systems have been criticized for their failure to improve teacher education programs (Navidinia, Kiany, Akbari, & Ghafarsamar, 2015). In response to these problems, the present study aims to provide a framework for teacher evaluation by proposing a performance assessment scheme to be used to measure the competencies to be developed by practicum courses.

3. Methodology

3.1. Participants

A total of 57 participants were involved in this study. However, along with the requirements of developing a performance assessment scheme, the participants were divided into two different groups for designing tasks and rating scales (group one) and setting standards (group two).

The first group included six participants who were in charge of designing tasks and rating scales and 37 other participants who were involved in the administration of the tasks. All the participants were selected using a convenience sampling procedure. Therefore, a panel of experienced teacher educators from Farhangian centers located in Tehran province established the committee of program leaders. This committee-included one member of the practicum committee, four experienced teacher educators teaching practicum courses, and one testing expert. As for the academic degree, five committee members were Ph.D. holders and one was a Ph.D. candidate of TEFL. Their average age was between 32 to 58 years old and their teaching experience varied between 15 to 32 years.

Among the 37 participants taking part in task administration, there were 34 student-teachers and 3 raters. The student-teachers were selected from among students who had passed Practicum 4 at Bahonar Teacher Training Center which is located in Tehran province. They fell in the age range of 22-27. The three raters who are teacher educators of Farhangian teacher training

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

centers located in Tehran province were selected based on their expertise. In terms of educational degree, one of them was a PhD holder of TEFL and two of them were PhD candidates. Their teaching experience ranged from 8 to 20 years. As regards the age of the participants, they fell in the age range of 30-45.

The second group involved in this study included 20 participants who were responsible for setting standards. This group was further divided into three groups. The first subgroup included the same committee members who participated in the previous phase of the study (program leaders). The second subgroup consisted of ten teacher educators that were involved in the study on the basis of their experience and expertise (panel of teacher educators). A convenience sampling procedure was used to select the second group. Therefore, the teacher educators were selected from Farhangian teacher training centers located in Tehran and Alborz provinces. All of the teacher educators held PhD degrees in TEFL. In terms of teaching experience, they had 5 to 20 years of teaching experience and their ages ranged from 33 to 52 years old. The last subgroup that participated in the standard setting phase consisted of mainly heads of departments at Farhangian teacher training centers in Tehran and Khorasan provinces. Therefore, a total of four heads of departments with PhD degrees formed the third group (policy makers).

3.2. Instrument

Given that a detailed account of the competencies to be measured is the initial step in the development of any teacher evaluation method (Taut & Sun, 2014), the researchers of the current study used ~~a~~ *Performance Assessment Scale* (Kiany, Karimi, & Norouzi, 2017) which had been designed to be used as a benchmark for assessing the professional competencies of ELT student-teachers of Farhangian University. The items of the scale were taken from the content analyses of the Curriculum Document of the English Major (practicum part), in-depth interviews with stakeholders, and the review of the literature. The analyses of the above-

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

mentioned sources resulted in seventeen items which were reduced to thirteen items after running Confirmatory Factor Analysis. In a further step, the items were classified into three main domains including seven items in the first domain and three items in the second and the third domains (Appendix A). In addition to defining the 'what' of assessment, as Taut and Sun (2014) have appropriately maintained, it is crucial to determine the 'how' of assessment. To address this issue, three instruments were developed in this study. What follows is a brief explanation of the instruments.

3.2.1. Task Design

Performance assessment is synonymous with task-based assessment (Ross, 2012), hence, designing a systematically valid testing scheme requires a representative set of tasks that "cover the spectrum of knowledge, skills, and strategies needed for the activity or domain being tested" (Frederiksen & Collins, 1989, p. 30). Therefore, the researchers used the expertise of the committee which consisted of one testing expert and a panel of content experts. They held three meetings, each lasting three hours. After detailed and lengthy discussions about the development of tasks inclusive of the competencies that student-teachers were expected to acquire from practicum courses, they reached an agreement on the tasks to be used in the performance assessment scheme so that they could verify that the content was an authentic representation of the competencies. Regarding the length and complexity of the performance tasks, the members followed Kane, Crooks, and Cohen's (1999) suggestions:

- avoid using a small number of lengthy performances to prevent inconsistent scoring, achieve standardization, and enhance generalizability and reliability;
- shorten some steps, for instance, by using simulations;
- select shorter tasks of each kind (in terms of presentation and mode of response);

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

- ask examinees to perform part of a lengthy task.

Taking all these issues into account, three tasks were developed in this study to measure all the competencies offered by practicum courses (Appendix B).

3.2.1.1. Planning Instruction and Assessment for Learning Task

This written task requires student-teachers to demonstrate the ability to plan instruction based on the information they have about students including their goals and proficiency levels. Student-teachers are expected to connect what they know about students to their instructional planning. In other words, their planning must be shaped by and reflect student characteristics. In addition to instructional planning, within this task, student-teachers must design student assessment activities and develop sufficient materials.

3.2.1.2. Instruction Task

It is a written and video-recorded task in which student-teachers must demonstrate the ability to implement the lesson they had designed while making appropriate use of class time, creating an environment of respect and rapport, adapting instruction to students, managing instruction, student behaviors and interactions, implementing learning and assessment activities, as well as assessing student learning.

3.2.1.3. Reflection Task

This written task is based on the reflective observations of student-teachers. They must demonstrate the ability to identify educational and pedagogical problems of the context, recognize student characteristics, and accordingly offer their suggestions to solve the problems on a scientific basis. They must also reflect on the strengths and weaknesses of the implemented lesson and its effectiveness, explain the implications of the reflection process on their professional development, and elaborate on both limitations of their professional development and their plans for further professional growth.

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

3.2.2. Rating Scale

Rating scales refer to "a series of hierarchical levels, with each level providing a proficiency descriptor against which learner performance is measured" (Fulcher, 2012, p. 378). The proficiency descriptors, taken together, pave the way for the operational definition of the construct being assessed (Davies, Brown, Elder, Hill, Lumley, & McNamara, 1996; Fulcher, 1996). The committee members, in addition to developing test tasks, exchanged views on a rating scale and assessors. To be compatible with the Curriculum Document of the English Major, they made an effort to define student-teachers' performances with regard to each of the components in three levels. To this end, they carefully analyzed the Curriculum Document of the English Major and could extract a rating scale from the document for only five factors (Appendix C). Since there were no clear-cut performance levels for the remaining eight factors in the Curriculum Document of the English Major and also in order to have well defined proficiency descriptors, they decided to use Danielson's (2011) performance score levels for the rest of the factors (Appendix D). They also decided to raise the legitimacy of scoring through rescoring some tasks periodically and critically reviewing the procedures for scoring and administrating the assessment (Kane et al., 1999).

3.2.3. Standard Setting

Essential to the consequential validity of any testing system is standard setting (Messick, 1989). Cizek and Bunch (2007) define standard setting as "a procedure that enables participants using a specified method to bring to bear their judgments in such a way as to translate the policy positions of authorizing entities into locations on a score scale" (p. 19). It follows that any attempt aimed at setting standards must adhere to strict guidelines. Therefore, the researchers, in this study, used Cizek's (1996) principles for standard setting. The first principle concerns the participants' awareness of

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

the purpose of assessment and the constructs being measured. Since the researchers and by extension the participants of the present study were well aware of the constructs, that is the professional competencies of ELT student-teachers, and of the purpose of the current assessment scheme which is a licensure and certification testing employed to assess the ELT student-teachers' competencies, it was easy to justify the method used to set standards. The other issue to be considered in selecting an appropriate standard setting methodology deals with supporting the professional acceptability and technical adequacy of the selected method by documenting enough information from the professional literature (Cizek, 1996). To address this concern, the researchers conducted a comprehensive study of the standard setting methods used in different performance assessment schemes, compiled a list of models, and finally based on the context of the present study including purpose, constructs, tasks, and so forth adopted the standard setting model used in PACT. This model is known as the participatory process in the professional literature and is suggested by Haertel (2002).

Therefore, the third instrument used in this study was a passing standard. Following Haertel's (2002) model and Cizek's (1996) guidelines, first, a panel of teacher educators familiar with the scoring process of the assessment scheme were asked to give their initial recommendations for passing standards for each of the tasks and the whole assessment scheme. Their recommendations amounted to five passing standards, with the repeated ones removed, for the first and the second tasks and three passing standards for the third task. A total of three passing standards were also recommended for the whole assessment scheme (see Appendix E, Table 1). Then, a confirmatory group (the members of the committee) reviewed those initial recommendations and selected two passing standards for Task 1, three passing standards for Task 2, two passing standards for Task 3, and two passing standards for the whole assessment scheme (Table 1). Finally, the third group which consisted of the heads of departments of Farhangian teacher training centers reviewed the proposed passing standards, took

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

advice from teacher educators, weighted all the recommendations, took the failure rate into account, and selected a passing standard for each task and one for the whole assessment scheme as the final passing standard:

> *Student-teachers pass the first task if they obtain no level of "1"; they pass the second task if they obtain no more than two levels of "1"; they pass the third task if they obtain no more than one level of "1"; and they pass the performance assessment scheme if they pass all the three tasks.*

3.3. Data Collection

In an attempt to collect more reliable data, the researchers decided to ensure student-teachers' comprehension of the tasks, the purpose of the assessment scheme, and the scoring criteria. They also held a briefing session with the student-teachers lasting for three hours. During the session, the researchers explained the tasks, the competencies measured by each task, the artifacts and documentations to be submitted, and the criteria for assessing performance. An extensive set of data was gathered in this phase of the study. Required by the first task, 34 copies of the designed activities, materials, and reflective commentaries were submitted by the student-teachers. The second task resulted in 34 reflective commentaries and 34 video clips each lasting for about one hour. The third task yielded 34 thick descriptions and required each student-teacher to submit evidence documenting their professional development.

3.4. Data Analysis

The data obtained from the implementation of the tasks were submitted to two raters to be scored. Since inconsistency in scoring has an adverse influence on validity and reliability, the researchers made an effort to address the urgent need to train assessors in a uniform manner so that they could enhance reliability and validity of the performance assessment scheme (Fehrmann, Woehr, & Arthur, 1991; Torgerson et al., 2009). With the help

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

of one of the raters who was a faculty member and a member of the program leaders in the present study, the researchers selected two teacher educators with similar teaching experience who had the required pedagogical content knowledge and content knowledge; informed them about the purpose of the assessment scheme; and explained the rating scales to them. In general, the raters were trained to rely on the rubric as the sole criteria for scoring each performance, avoid any bias towards students, and avoid scoring performance on one task under the influence of performance on other tasks.

In a further step to improve raters' understanding of standards, a calibration procedure was used. Calibration is "the process by which an assessor's scores for a specific performance relative to a specific rubric come to match scores determined by experts to be reflective of that same performance using the same rubric" (Torgerson et al., 2009, p. 67). First, each rater rated the performances of five student-teachers separately. Then, their ratings were compared to identify points of disagreement. If the two raters' ratings differed substantially from those of the supervisor and they could not achieve consensus, the supervisor would function as a third rater to resolve the discrepancies. Finally, the whole documentations submitted by the student-teachers were given to the raters to be assessed.

3. Results

As pointed out earlier, this study aimed to develop a performance assessment scheme, and consequently investigate how valid and reliable the developed performance assessment scheme is. To begin to present the results of the study, the evidence documenting the reliability of the assessment scheme is provided. Since uniformity in rating enhances the reliability and validity of assessment, the researchers trained assessors with the help of a faculty member. Then, they examined the ratings of the two assessors to estimate inter-rater reliability. Table 2 provides detailed information about the ratings. As shown by the table, a disagreement of +/- 2 was not observed between the raters in any of the cases. In 92 out of the 102 possible decisions on the first

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

task, the first and the second raters were in absolute agreement. Of the 204 possible decisions on the second task, the raters were in disagreement only on 30 cases and they were in absolute agreement on 120 out of 136 possible decisions on task three. In sum, the resulting reliability coefficient for the whole assessment scheme was .85 and, as stated earlier, points of disagreement were resolved by consensus.

Table 2

Summary of exact matches and disagreement for the tasks

Tasks	**Total possible decisions**	**Exact match**	**+/-1 level**	**+/-2 levels**	**Pass/fail disagreements**
Task 1	102	92	10	0	0
Task 2	204	174	30	0	4
Task 3	136	120	16	0	1
Total	442	386	56	0	5

A further investigation of the reliability of the assessment included the estimation of intra-rater reliability. One of the assessors was asked to rate the performances of ten student-teachers twice. The analysis of the ratings revealed a consistency level of 0.89.

In addition to reliability, the researchers conducted some analyses to estimate the validity of the assessment scheme. To establish the content validity of the assessment scheme, the committee members involved in designing the tasks and rubrics were asked to judge the extent to which the content of the assessment scheme was an authentic representation of the competencies student-teachers were expected to acquire from practicum courses. To put it differently, they were required to examine the correspondence between the skills and knowledge in the target domain and the skills and knowledge needed to perform the assessment tasks. To this

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

end, they analyzed each of the items measured by the tasks, identified the competencies needed to perform the tasks, and determined the corresponding practicum course during which the student-teachers must have acquired the related competencies. Table 3 provides evidence regarding the validity of the assessment scheme. As the results show, the panel of content experts believed that ten of the items were the main foci of practicum courses and only three items were not directly taught. The required competencies for Task 1 were associated with practicum courses two, three, and four. Task 2 measured six items. Three out of six items were the main foci of practicum courses two, three, and four and only three items were not considered to be the main focus of practicum courses. In Task 3, all the four items were covered in practicum courses.

Table 3
The correspondence between target domain and test domain skills

Practicum	**Item**	**Task**	**Competency**
2,3	2	1	Design activities in accordance with content area and student characteristics
3,4	5	1	Design appropriate assessment activities, identify learning outcomes and criteria for assessment
4	3	1	Evaluate, modify, adapt, and develop materials that is suitable for learners and in accordance with content area
2,3,4	10	2	Select and adapt instructional strategies, learning and assessment activities, and materials that are suitable for students and in accordance with instructional purposes
3,4	11	2	Interpret assessment results, monitor students and provide feedback, plan further

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

			instruction
2,3,4	9	2	Allocate instructional time appropriately
-	6	2	Interact with students respectfully and controls and monitors interactions between students
-	7	2	Develop and maintain expectations for behavior
-	8	2	Arrange furniture and use physical resources to enhance learning
1,2	1	3	Obtain information about students' needs, goals, linguistic backgrounds, language proficiency
1	17	3	Describe learning context in terms of educational, emotional, and physical characteristics
1,2,3,4	14	3	Reflect on the instruction, analyze its weaknesses, strengths, and effectiveness, understand implications for further planning and adaptation
4	16	3	Cite reasons for his/her success based on reflections, understand limitations on developing content knowledge and pedagogical knowledge, plan for further professional development

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

5. Discussion

The nature of performance assessment calls for considerable caution to be exercised in its development and interpretation. In fact, "the usefulness of performance assessment for licensure and program improvement depends on the degree to which the scoring is valid and reliable" (Torgerson et al., 2009, p. 73). Thus, basic to performance assessment are the notions of validity and reliability (Bachman & Palmer, 1996).

The adverse influence of inconsistency in scoring on both validity and reliability (Brown, 2012; Fehrmann et al., 1991; Torgerson et. al., 2009) suggests the need for rater training (Brown, 2012). Similarly, Fehrmann et al. (1991) maintain that "variability across raters may stem from either different conceptions of competence or mastery…and different levels of expertise" (p. 858), and refer to rater training as a means to increase inter-rater agreement.

In this respect, the raters were trained based on Brown's (2012) suggestions to improve both inter-rater reliability and intra-rater reliability. To further improve inter-rater consistency, the researchers followed Erdosy's (2004) suggestion regarding the higher possibility of approaching the task of judging performance based on a shared view of the construct being assessed if the selected raters have similar backgrounds such as similar professional experience. On the whole, the results evidenced the effectiveness of rater training in enhancing the reliability of the assessment scheme since inter-rater and intra-rater reliability coefficients were high.

Another critical aspect of any assessment system is its validity. Since content validity is usually used to support teacher licensure assessment (AERA, APA & NCME, 1999; Wilkerson & Lange, 2003), the researchers first needed to ensure content validity (Ross, 2012), which is in line with Frederiksen and Collins's (1989) proposed criteria for the validity of any

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

performance assessment. Therefore, the researchers conducted content domain analysis (Ross, 2012). The results showed the correspondence between the skills and knowledge in the target domain, that is, practicum courses, and the skills needed to perform tasks in the ten items. Although the remaining three items and the competencies associated with them, as the panel of content experts put, were not of main concern in practicum courses, they are essential teaching skills which were implicitly considered in practicum courses in general. Thus, evidence of the content validity of the current performance assessment scheme was based on the experts' judgment of the committee of content experts (program leaders). In fact, the judgment of the experts' served to confirm that the assessment tasks were meaningful constructs representing significant domains of teaching skills.

Another step towards a valid performance assessment relates to the passing standard. As indicated previously, standard setting is based on experts' judgments; therefore, it follows that essential to an accurate verbal description of performance standards is a panel of qualified judges (Jaeger, 1991; Kane, 1994). A full description of the standard setting panel helps to ensure the match between the characteristics of the group and the judgments that must be made, which, in turn, "provides further evidence for the validity of the resulting decisions (e.g., pass/fail, certify/do not certify, etc.)" (Cizek, 1996, p. 18). A clear understanding of the task of standard setting is also essential. To address this issue, Cizek (1996) suggests that the standard setting panel be informed about the purpose of standard setting and be trained in the application of the selected approach.

In light of what was mentioned above, it can be inferred that the passing standards employed in the present study are valid for several reasons. First, all the participants involved in setting the standards were selected from Farhangian University, and consequently were well aware of the purpose of the assessment scheme. Second, the panel of teacher educators were qualified teacher educators who were introduced by faculty members. Third, the researchers conducted a two-hour briefing session and further online

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

discussions with the participants to ensure their understanding of the methodology. Based on the results of this study, it is believed that further development of the proposed assessment scheme fulfills its potential to be utilized as a large-scale teacher assessment model for Farhangian University.

6. Conclusion

There is a substantial body of research documenting the need for teacher evaluation (Aseltine, Faryniarz, & Rigazio-DiGilio, 2006; Marshall, 2009; Sergiovanni & Starrat, 2002). This study also aims to contribute to this field of research by developing a performance assessment scheme to be used as a benchmark for assessing the competencies of ELT graduates of Farhangian University. The analysis of the results yields conclusive evidence for the validity and reliability of the assessment scheme. Moreover, this study makes major contributions to teacher education programs. To begin with, it must be recognized that the present study is the first empirical research project conducted in Iran's context to investigate the fundamental issue of preservice assessment of professional competencies of teacher candidates, and therefore, paves the way for further research in this regard. Second, consistent with other research studies documenting the strong effect of teacher evaluation on teachers' professional development, learners' academic achievement, and highly qualified education, this study highlights the urgency of incorporating preservice assessment into teacher education programs. Most important is that this study gains insights from international experiences and takes the initiative to use performance-based assessment, though with inherent limitations, rather than knowledge-based tests to measure the competencies of student-teachers, hence, placing performance-based assessment as a top priority for any educational reform movement. Some features of the present study that might be useful to be incorporated into future studies include

- using Haertel's (2002) participatory approach for standards setting which gives voice to all the stakeholders;

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

- providing detailed descriptions of performance levels rather than relying on single words for measuring competencies which are not open to interpretation;
- training raters and applying calibration processes which enhance the reliability of the assessment procedure;
- using simulations and shorter tasks in order for the assessment procedure to be more practical.

Although the results supported the reliability and validity of the assessment scheme, caution must be exercised in interpreting the results. First, as stated by Kane et al. (1999), it is not possible to generalize performance on small samples of tasks to larger domains. For instance, lesson plans are only part of the overall instruction; therefore, evaluating teachers' performance based on the lesson plans they design and implement gives us only a partial picture of teacher's overall classroom performance (Marshall, 2009). Thus, performance assessment cannot be highly generalized to teachers' actual performances in classrooms**.** It should also be stated that the data were mainly collected from student-teachers and teacher educators at different branches of Farhangian University located in Tehran province. This sampling procedure imposes further limitations on the generalizability of the assessment scheme.

Since this study was the first one to provide a model for performance assessment scheme, replication studies are required to verify its findings and as suggested by Brennan and Johnson (1995) to estimate its reliability, identify any areas of bias due to race, gender, ethnicity or cultural-linguistic backgrounds, and investigate test use consequences as part of test validation. Moreover, the assessment scheme developed in this study was restricted to the competencies that student-teachers were expected to acquire from practicum; so, further research is required to investigate the competencies that student-teachers should be equipped with after graduating from

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

Farhangian University and accordingly a more comprehensive performance assessment scheme should be developed.

7. References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1985). *Standards for Educational and Psychological Testing.* Washington, DC: American Psychological Association.

Arends, R. I. (2006). Performance assessment in perspective: History, opportunities, and challenges. In S. Castle & B. D. Shaklee (Eds.), *Assessing teacher performance: Performance-based assessment in teacher education* (pp. 3–22). Lanham, MD: Rowman & Littlefield Education.

Aseltine, J. M., Faryniarz, J. O., & Rigazio-DiGilio, A. J. (2006*). A performance-based approach to teacher development and school improvement: Supervision for learning.* Association for Supervision and Curriculum Development, Alexandria, VA.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice, 14*(4), 25-27.

Brown, A. (2012). Interlocutor and rater trianing. In G. Fulcher, & F. Davidson (Eds.), *The Routledge Handbook of Language Testing*. New York, NY: Routledge.

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

California Commission on Teacher Credentialing (2016). *CalTPA Handbook.* Sacramento: California Commission on Teacher Credentialing.

Chung, R. R. (2008). Beyond assessment: Performance assessments in teacher education. *Teacher Education Quarterly, 35* (1), 7-28.

Cizek, G. J. (1996). Standard-setting guidelines. *Educational Measurement: Issues and Practice*, *15*(1), 13-21.

Cizek, G. J., & Bunch, M. B.(2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. UK. Thousand Oaks.

Danielson, C. (2011). *Enhancing professional practice: A framework for teaching*. Association for Supervision and Curriculum Development (ASCD).

Danielson, C., & Marquez, E. (1998). *A collection of performance tasks and rubrics: High school mathematics*. Larchmont, NY: Eye on Education.

Darling-Hammond, L. (2010). *Evaluating teacher effectiveness: How teacher performance assessments can measure and improve teaching*. Center for American Progress.

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and teacher education*, *16*(5), 523-545.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. & McNamara, T. (1999). *Dictionary of language testing.* Cambridge, UK: Cambridge University Press.

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

Delandshere, G., & Arens, S. A. (2003). Examining the quality of the evidence in preservice teacher portfolios. *Journal of Teacher Education, 54*(1), 57-73.

Erdosy, M. U. (2004). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions*. Princeton, NJ: Educational Testing Service.

Fehrmann, M. L., Woehr, D. J., & Arthur, W. (1991). The Angoff cutoff score method: The impact of frame-of-reference rater training. *Educational and psychological measurement, 51*(4), 857-872.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational researcher, 18*(9), 27-32.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208-238.

Fulcher, G. (2012). Scoring performance tests. In G. Fulcher, & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 378-392). New York, NY: Routledge.

Girod, M., & Girod, G. R. (2008). Simulation and the need for practice in teacher preparation. *Journal of Technology and Teacher Education, 16*(3), 307-337.

Haertel, E. H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational measurement issues and practice, 21*(1), 16-22.

Hamilton, L. (2003). Assessment as a policy tool. *Review of research in education, 27*, 25-68.

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

İşlek, D., & Hürsen, Ç. (2014). The evaluation of students' views concerning the teacher qualifications for the total quality implementations. *Procedia - Social and Behavioral Sciences, 116(2),* 4834–4838.

Jaeger, R. M. (1991). Establishing standards for teacher certification tests. *Educational Measurement, Issues and Practices, 9*(4), 15-20.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*(3), 425-461.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5-17.

Kiany, G., Karimi, M., & Norouzi, M. (2017). An assessment scheme for ELT performance: An iranian case of Farhangian University. *Journal of Teaching Language Skills*. In Press.

Koirala, H. P., Davis, M., & Johnson, P. (2008). Development of a performance assessment task and rubric to measure prospective secondary school mathematics teachers' pedagogical content knowledge and skills. *Journal of Mathematics Teacher Education*, *11*(2), 127-138.

Marshall, K. (2009). *Rethinking teacher evaluation and supervision how to work smart, build collaboration, and close the achievement gap*. Jossey-Bass, San Francisco.

Medley, D. M., (1982). *Teacher competency testing and the teacher educators.* Association of Teacher Educators and the Bureau of Educational Research: University of Virginia.

Messick, S. (1989). Validity. In R.L. Linn (Ed.) *Educational measurement* (pp.13-103). Washington, DC: National Council of Measurement in Education and The American Council on Measurement in Education.

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

Navidinia, H., Kiany, G. R., Akbari, R., & Ghafarsamar, R. (2015). EFL teacher performance evaluation in Iranian high schools: Examining the effectiveness of the status quo and setting the groundwork for developing an alternative model. *The International Journal of Humanities*, *21*(4), 27-53.

Pecheone, R., & Chung, R. R. (2006). Evidence in teacher education: The performance assessment for California teachers**.** *Journal of Teacher Education, 57*(1), 22-36.

Ross, S. J. (2012). Claims, evidence, and inference in performance assessment. In G. Fulcher & F. Davidson (Eds.). *The Routledge handbook of language testing*. New York, NY: Routledge

Sanders, W. L., Wright, S. P., & Horn, S. P. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation.*Journal of personnel evaluation in education*, *11*(1), 57-67.

Sandholtz, J. H. (2012). Predictions and performance on the PACT teaching event: Case studies of high and low performers. *Teacher Education Quarterly*,*39*(3), 103-126.

Sergiovanni, T. J., & Starrat, R.J. (2002). Supervision: A redefinition (7th ed.) Boston, MA: McGraw Hill.

Shinkfield, A. J., & Stufflebeam, D. L. (1995). School professionals' guide to improving teacher evaluation systems. In *Teacher Evaluation* (pp. 81-172). Springer Netherlands.

Stewart, A. R., Scalzo, J. N., Merino, N., & Nilsen, K. (2015). Beyond the criteria: Evidence of teacher learning in a performance assessment.*Teacher Education Quarterly*, *42*(3), 33.

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

Taut, S., & Sun, Y. (2014). The development and implementation of a national, standards-based, multi-method teacher performance assessment system in Chile. *Education Policy Analysis Archives, 22* (71).

Torgerson, C. W., Macy, S. R., Beare, P., & Tanner, D. E. (2009). Fresno assessment of student teachers: A teacher performance assessment that informs practice. *Issues in Teacher Education*, *18*(1), 63-82.

Wilkerson, J.R. & Lange, W.S. (2003). Portfolio, the pied piper of teacher certification assessments: Legal and psychometric issues. *Education Policy Analysis Archives, 11*(45).

Notes on Contributors:

Gholam Reza Kiany is an associate professor of Applied Linguistics at Tarbiat Modares University. He is also the vice-chancellor of supervision, evaluation, and quality assurance at Farhangian University. His research interests include teacher education, assessment, teacher evaluation, and program evaluation.

Monireh Norouzi has an M.A. degree in English Teaching from Tarbiat Modares University. Her research interests include teacher evaluation, performance assessment, and teacher cognition.

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

Appendix A: Performance Assessment Scale

Domain 1: Planning and instruction	**Domain 2: The classroom environment**	**Domain 3: Professional development and responsibilities**
1) Demonstrating knowledge of students;	1) Creating an environment of respect and rapport;	1) Reflecting on teaching;
2) Designing learning activities;	2) managing student behavior;	2) Growing and developing professionally;
3) Developing instructional materials and resources;	3) Organizing physical space	3) Reflective observation
4) Designing student assessments;		
5) Time management;		
6) Engaging students in learning;		
7) Using assessment in instruction		

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

Appendix B: Assessment Tasks

Task 1: The planning instruction and assessment for learning task

You are going to teach a group of 30 teenagers at beginners' level studying at grade 8 (key stage 8) at school. Some of them have background knowledge of English from Private institutes.

The lesson you are going to teach is about "My favorite food". Design learning and assessment activities. Decide on the parts of the textbook that you would use without change and the parts that you would need to adapt in your design.

In this task, you will demonstrate your ability to plan instruction and assessment that is shaped by student characteristics and is appropriate to the learning goals of the unit.

This task measures:

- Designing learning activities
- Designing student assessment
- Developing instructional materials and resources

You should submit:

- A copy of assessment artifacts including the assessment activities, scoring scales, and rubrics answer key, and directions.
- A copy of the learning activities
- A copy of the materials you used or designed (if any), e.g. worksheets, games, etc.

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

- A reflective commentary on the experience containing answers to specified questions

Task 2: The Instruction Task

Implement your instructional planning and assessment activities. If necessary, adapt your instructional planning based on the assessment results. Make sure all the learners are involved and can benefit from the session.

In this task, you will demonstrate your ability to manage class time, make use of instructional resources, manage student interaction, create positive learning environment, and assess student learning.

This Task measures:

- Creating an environment of respect and rapport
- Manage student behavior
- Time management
- Use assessment in instruction
- Engaging students in learning
- Organizing physical space

You should submit:

- A video clip portraying the required features of your teaching
- selected student assessment responses
- A detailed reflective commentary on the experience containing answers to specified questions

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

Task 3: The Reflection Task

Watch the film of your own instruction in task two.
Describe what you have observed. What focal point/problem do you come up with after watching the film? What evidence do you have for the focal point/problem you have found?

In this task, you will demonstrate your ability to learn important details about a classroom of students, show how you would apply this information to your future planning for these students, reflect on the whole process and its implications for improving your teaching effectiveness and professional development.

This task measures:

- Reflective observation
- Growing and developing professionally
- Demonstrating knowledge of students
- Reflecting on teaching

You should submit:

- A full description of your reflective observation
- Responses to the questions
- Evidence on the professional development activities in which you engaged

Appendix C: Performance levels extracted from the Curriculum Document of the English major

Item 2: Designing Learning Activities

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

1) The learning activities are designed based on unit goals but are not well organized and fail to address any individual or collective problems.

2) The learning activities are well organized and are designed based on unit goals. They address individual or collective problems.

3) The learning activities are well organized and are designed based on unit goals. They effectively address individual or collective problems, are differentiated, and leave room for further adaptation.

Item 3: Developing Instructional Materials and Resources

1) The student-teacher can adapt materials to student characteristics but cannot develop new materials.

2) The student-teacher can adapt materials to student characteristics and develop new materials that are not suitable to students or instructional outcomes.

3) The student-teacher can adapt materials to student characteristics, develop appropriate instructional materials creatively by drawing on teaching and learning experiences.

Item 5: Design Student Assessment

1) The student-teacher can design assessment but cannot identify learning outcomes and criteria accurately.

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

2) The student-teacher can design assessment, identify learning outcomes and criteria accurately, and adapt assessment to groups of learners.

3) The student-teacher can design assessment, identify learning outcomes and criteria accurately, and adapt assessment to individual learners.

Item 16: Growing and Developing Professionally

1) The reflective commentary includes the activities undertaken by the student-teacher but does not provide evidence on the professional skills, limitations, and suggestions for further professional development.

2) The reflective commentary includes the activities undertaken by the student-teacher and provides evidence on the professional skills and limitations.
3) The reflective commentary includes the activities undertaken by the student-teacher and provides evidence on the professional skills, limitations, and practical suggestions for further professional development.

Item 17: Reflective Observation

1) The student-teacher can describe the learning context but cannot identify educational and pedagogical problems.

2) The student-teacher can describe the leaning context in an organized manner, identify educational and pedagogical problems, and offer solutions to solve the problems based on the context.

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

3) The student-teacher can describe the learning context in an organized manner, identify educational and pedagogical problems, offer solutions to solve the problems, and support his/her solutions on a scientific basis.

Appendix D: Performance levels extracted from Danielson's (2011) framework

Item 1: Demonstrating Knowledge of Students

1) Teacher displays little or no knowledge of students' skills, knowledge, and language proficiency and does not indicate that such knowledge is valuable; Teacher displays little or no knowledge of students' interests or cultural heritage and does not indicate that such knowledge is valuable; Teacher displays little or no understanding of students' special learning or medical needs or why such knowledge is important.

2) Teacher recognizes the value of understanding students' skills, knowledge, and language proficiency but displays this knowledge only for the class as a whole; Teacher recognizes the value of understanding students' interests and cultural heritage but displays this knowledge only for the class as a whole; Teacher displays awareness of the importance of knowing students' special learning or medical needs, but such knowledge may be incomplete or inaccurate.

3) Teacher displays understanding of individual students' skills, knowledge, and language proficiency and has a strategy for maintaining such information; Teacher recognizes the value of

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

understanding students' interests and cultural heritage and displays this knowledge for individual students; Teacher possesses information about each student's learning and medical needs, collecting such information from a variety of sources.

Item 6: Creating an Environment of Respect and Rapport

1) interactions (both between teacher and students) are negative or inappropriate characterized by sarcasm, putdowns, or conflicts.

2) interactions (both between teacher and students) are sometimes appropriate and sometimes inappropriate characterized by generally appropriate but occasionally disrespectful interactions.

3) interactions (both between teacher and students) are appropriate. Students treat each other with respect and exhibit complete respect for teacher and trust him or her.

Item 7: Managing Student Behavior

1) classroom is characterized by no expectations for standards of conduct, no monitoring of student behavior on the part of teacher, and no response to misbehavior on the part of teacher.

2) classroom is characterized by low expectations for standards of conducts, monitoring student behavior on the part of teacher which sometimes misses the activities of some students, and sometimes appropriate and sometimes inappropriate teacher response to misbehavior.

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

3) classroom is characterized high and clear expectations for standards of conduct, respectful monitoring on the part of both teacher and students, effective response to misbehavior on the part of teacher, and appropriate student behavior

Item 8: Organizing Physical Space

1) The classroom is unsafe, or learning is not accessible to some students; The furniture arrangement hinders the learning activities, or the teacher makes poor use of physical resources.

2) The classroom is safe, and at least essential learning is accessible to most students; Teacher uses physical resources adequately. The furniture may be adjusted for a lesson, but with limited effectiveness.

3) The classroom is safe, and learning is equally accessible to all students; Teacher uses physical resources skillfully, and the furniture arrangement is a resource for learning activities.

Item 9: Time Management

1) doesn't allocate instructional time appropriately and consequently can't implement the whole planned lesson (much time is lost between activities).

2) allocates instructional time sometimes appropriately and sometimes inappropriately (only some transitions are efficient which results in some loss of instructional time and in skipping some parts to finish the planned lesson)

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

3) allocates instructional time appropriately by establishing appropriate procedures for routine tasks and managing transitions to maximize instructional time.

Item 10: Engaging Students in Learning

1) Activities and assignments are inappropriate for students' age or background. Students are not mentally engaged in them; Instructional groups are inappropriate to the students or to the instructional outcomes; Instructional materials and resources are unsuitable to the instructional purposes or do not engage students mentally; The lesson has no clearly defined structure, or the pace of the lesson is too slow or rushed, or both.

2) Activities and assignments are appropriate to some students and engage them mentally, but others are not engaged; Instructional groups are only partially appropriate to the students or only moderately successful in advancing the instructional outcomes of the lesson; Instructional materials and resources are only partially suitable to the instructional purposes, or students are only partially mentally engaged with them; The lesson has a recognizable structure, although it is not uniformly maintained throughout the lesson. Pacing of the lesson is inconsistent.

3) Most activities and assignments are appropriate to students, and almost all students are cognitively engaged in exploring content; Instructional groups are productive and fully appropriate to the students or to the instructional purposes of the lesson; Instructional materials and resources are suitable to the instructional purposes and engage students mentally; The lesson has a clearly defined structure

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

around which the activities are organized. Pacing of the lesson is generally appropriate.

Item 11: Use Assessment in Instruction

1) Inadequately uses assessment results to determine student achievement and to plan further instruction (inaccurate or no feedback, inappropriate or no adaptation)

2) Minimally uses assessment results to determine student achievement and plan further instruction (limited or minimal feedback, sometimes appropriate and sometimes inappropriate adaptation,)

3) Appropriately uses assessment results to determine student achievement and to plan further instruction (accurate and detailed feedback, appropriate adaptation)

Item 14: Reflecting on Teaching

1) Teacher does not know whether a lesson was effective or achieved its instructional outcomes, or teacher profoundly misjudges the success of a lesson; Teacher has no suggestions for how a lesson could be improved another time the lesson is taught.

2) Teacher has a generally accurate impression of a lesson's effectiveness and the extent to which instructional outcomes were met; Teacher makes general suggestions about how a lesson could be improved another time the lesson is taught.

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

3) Teacher makes an accurate assessment of a lesson's effectiveness and the extent to which it achieved its instructional outcomes and can cite general references to support the judgment; Teacher makes a few specific suggestions of what could be tried another time the lesson is taught.

Appendix E: Passing Standards

Table 1: The recommendations for passing standards

Tasks	Teacher educators	Program leaders	Policy makers
Task 1	Student-teachers pass if they obtain: 1) no level of "1" 2) 3 levels of "3" 3) at least 1level of "3" 4) two levels of "2" or "3" 5) no more than 1 level of "1"	Student-teachers pass if they obtain: 1) no level of "1" 2) no more than 1 level of "1"	Student-teachers pass if they obtain: no levels of "1";
Task 2	Student-teachers pass if they obtain: 1) no more than two levels of "1" 2) no less than 2 levels of "3" 3) 3 levels of "2" or "3" 4) no level of "1" 5) no more than 3 levels	Student-teachers pass if they obtain: 1) no more than two levels of "1"; 2) 3 levels of "2" or "3" 3) no less than 2 levels of "3"	Student-teachers pass if they obtain: no more than two level of "1";

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]

	of "1"		
Task 3	Student-teachers pass if they obtain: 1) no more than 2 levels of "1" 2) 2 levels of "2" 3) no more than 1 level of "1"	Student-teachers pass if they obtain: 1) no more than one level of "1"; 2) no more than 2 levels of "1";	Student-teachers pass if they obtain: no more than one level of "1";
Assessment scheme	Student-teachers pass if they obtain: 1) no more than 3 levels of "1"; 2) 5 levels of "2" or "3"; 3) pass all the tasks	Student-teachers pass if they obtain: 1) no more than 3 levels of "1"; 2) pass all the tasks	Student-teachers pass if they pass: all the three tasks

[Downloaded from ndea10.khu.ac.ir on 2026-11-06]

[DOI: 10.29252/ijal.19.2.155]