Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 7, No. 2, October 2017
ISSN 2476-5880

Construction and Validation of a Tool for Measuring English Teacher Candidates' Professional Knowledge: Certification Policy and Practice Evidence from Teacher-Education University in Iran

Gholam-Reza Kiany[1], Parvaneh ShayesteFar[2], Yasser Amoosi[3]

Received: 01 August 2017 Accepted: 22 September 2017

Abstract

Teacher evaluation and certification is a proper approach taken to assess teacher knowledge base and to guarantee that teacher candidates are qualified and have met particular teaching standards. To this end, teacher-education university (i.e., Farhangian University) of Iran recently adopted a teacher evaluation and certification policy, nationally called ASLAH, which mandates all teacher candidates to earn a teaching certification by passing a comprehensive exam. The purpose of the present mixed-method study was to describe the development and validation of 'a written assessment framework', as one of the requirements of ASLAH project, that would measure English teacher candidates' professional knowledge and competency to teach English. The data were collected through a series of interviews with teacher educators and subject-matter experts (N=15), questionnaire surveys that involved 320 English language teachers, and test performance of 62 English teacher candidates. First, the study dealt with the development of the hypothesized framework that included items assessing teacher Content-Knowledge (CK) competency by virtue of nine domain-specific courses recommended by experts' qualitative and quantitative data. The items were empirically determined for content specification and validity, item difficulty and item discrimination. An initial piloting of the newly developed tool to teacher candidates showed tests as valid and reliable instruments for measuring teachers' CK competency. Perceived 'fairness', 'consequences', and 'quality of the results' of the present certification policy and practice were also explored from the eyes of all participants. The results showed that they were not negative about the possible test consequences and fairness. However they did not appear to be strongly positive about the practicality of teacher evaluation and certification project in the present context of teacher-education university. Policy recommendations and implications of the findings were discussed.

.***Keywords:*** *Teacher Evaluation; Teacher Content Knowledge; Certification Testing; Teacher-Education University; Professional Development*

[1] English Department, Tarbiat Modares University, Iran. *(Corresponding author)* Email: rezakiany@yahoo.com
[2] English Department, Farhangian University, Iran. Email: parishayeste@yahoo.com
[3] English Department, Tarbiat Modares University, Iran. Email: yaser.amoosi@yahoo.com

1. Introduction

'Teacher evaluation' is receiving attention worldwide as governments observe the need to delve into educational sectors and investigate them critically to ensure that they are accountable and appropriate to the needs of the youth (Monyatsi, Steyn & Kamper, 2006). Teacher evaluation has mainly focused on several factors, such as teacher knowledge and skills, teacher management, teacher personality and behaviors, and teacher efficacy and effectiveness, which are acknowledged to shape 'teacher quality' (Darling-Hammond, 1997, 2000; Goldhaber & Anthony, 2007; Hanushek & Rivkin, 2004; Mitchell et al. 2001). The focus on teacher quality is much warranted because teacher quality is evidenced as one of the main determinants of student learning and outcomes (Goldhaber & Brewer, 2000; Mangiante, 2011; Marsh & Hattie, 2002), hence, good information about teacher quality can be leveraged to improve instruction and its outcomes (Carey, 2004). As such, when evaluating teacher quality, teacher evaluation needs to pursue information on 'teacher professional development', in particular, on 'teacher professional knowledge and competency' (Stronge, 2006; Stronge & Tucker, 2003).

To evaluate teachers, various endeavors have been made so far, from observing teachers' performance and attributes to measuring the quality of their knowledge and readiness, and credentialing. 'Licensure testing', i.e., certification by virtue of exams, is one of the most widely approaches taken to evaluate teachers' knowledge base and teaching skills, and to grant credentials to them (Boyd, Goldhaber, Lankford & Wyckoff, 2007). Given the importance of providing students with qualified, knowledgeable and ready teachers, it becomes highly crucial that educational programs develop systems of teacher evaluation that can accurately measure teachers' knowledge, effectiveness and teaching readiness. Teacher-quality evaluation and certification as such has been the key concern of many teacher education programs, including Iran's teacher training programs.

When evaluating and certifying teachers, defining 'quality' is fundamental to understanding the role of licensure tests (Mitchell et al., 2001). However, this is not a simple task. Previously, definitions of teacher quality were concerned with teachers' behaviors or technical proficiency, while current definitions of teacher quality are mostly standards-based and are concerned with the 'knowledge, skills, and dispositions' that teacher should demonstrate. Examining standard settings by three internationally known organizations such as NBPTS (National Board for Professional Teaching Standards), INTASC (Interstate New Teacher Assessment and Support Consortium), and NCATE (National Council for Accreditation of Teacher Education), Mitchell et al. (2001) summarized the common standards set by these organizations. According to these standards, teachers: 1) have deep subject matter knowledge, 2) are reflective about their learning, 3) are committed to their students and students' learning, 4) manage and monitor student learning, and 5) are members of a broader community. As these themes suggest, one of the main premises of teacher quality is 'teachers' subject matter knowledge (SMK)'. The standard related to teachers' SMK is described by NCATE as "teacher candidates have in-depth knowledge of the subject matter that they plan to teach as described in professional, state, and institutional standards. They demonstrate their knowledge through inquiry, critical analysis, and synthesis of the subject" (Mitchell et al., 2001, p. 27). Likewise, based on the NBPTS standards teacher candidates who accomplish their training program should

have a rich understanding of the subject they wish to teach and recognize the ways in which knowledge in their subject is created, organized, and applied to real-world context.

To enhance teacher quality and to ensure that all schools are equipped with qualified and competent teacher, teacher licensing policy was adopted in a number of countries and states (Baumert et al., 2010; Boyd et al., 2007; Sadler, Sonnert, Coyle, Cook-Smith & Miller, 2013). This policy mandates teacher candidates to earn a 'teaching certification' by passing different kinds of licensure tests. Clear examples are Praxis Series Tests, Licensure Testing Systems, Pre-Service Teacher Assessment, National Evaluation Series, and a number of other national and state licensing tests employed around the world. In fact, licensing tests are proper means "to provide the public with a dependable mechanism for identifying practitioners who have met particular standards" (AER, APA & NCME, 1999).

Having felt such a need for teacher licensing, and more importantly, having aimed at ensuring and enhancing educational development, in general, and teacher quality, in particular, Iran's Ministry of Education (ME) initiated a reform agenda in its teacher education system, in recent years. This was initially inspired by a general reform wave that was generated and diffused through the overall system of general education of the country in the 2000s. In fact, by 2012, the country's higher-rank policymaking bodies such as the Supreme Council of Cultural Revolution, Higher Education Council, and Ministry of Education reached a consensus to bring about fundamental reforms in the country. The result of this collaborative initiative was 'the Document of Fundamental Reforms in Education (DFRE, 2012)'. The document raises a general reform movement in both general and higher education of the country, from policy to practice, from curriculum to assessment. The document noticeably targets 23 macro policies two of which are specifically concerned with teachers and teacher education program. For instance, macro policy 10 concerns (a) Upgrading teacher profession, (b) Setting an assessment regime for beginning teachers' professional competency, and (c) Setting teacher ranking system for professional development.

In Iran, the Ministry of education (ME) is mainly involved in providing the conditions for individuals who wish to enter the teaching profession. However, with their increasing emphasis on teacher professional competencies, the ME, the MSRT (Ministry of Science, Research and Technology) and the higher-rank Supreme councils endorsed one of the main teacher-training premises of the FRED and therefore, in a collaborative fashion, established an officially national organization for teacher education, called Farhangian University (FU). The establishment of this particular teacher training organization is an essential prerequisite for bringing about a fundamental change in the teacher training program of the country. At a state level, this particular higher education system is organized by a leading institution (locally called Central Organization) to which around 70 provincial centers are linked. Through this new infrastructure, the need for fundamental reform in teacher training program as well as the need for qualified teachers can be met.

With the establishment of Farhangian University as the main organization for training the prospective teachers needed by the ME, subsequent fundamental changes appeared too. For instance, in 2012-2013, attempts were made to develop a FU-specific curriculum for teacher education. Soon after this, another radical change was pronounced by FU's officials. This later change articulated a 'teacher assessment regime' that obviously concerned with 'Professional Assessment for Beginning Teachers' (PABT; or ASLAH, as its Persian Equivalent short-form).

National educators have now come to the view that 'the link between teacher quality and quality education is so strong which mandates all students be taught by highly qualified teachers'. Informed by this view and the international evidence on teacher licensing through established certificate systems (such as PRAXIS I and II), also inspired by the local higher-order educational documents of the country (such as the FRED) as well as the national reports on teacher assessment and evaluation (cf. Kiany et al., 2016; Navidinia et al, 2015), FU authorities have, very recently, set up the PABT project in order to evaluate FU student-teachers' readiness for prospective teaching and certify them. The main reasons behind such a project are to a) assure the attainment of professional competencies by FU's student-teachers during their four-year training program, b) promote their performance levels, and c) increase their motivation for acquiring the contents and skills of the training program.

Therefore, initiatives were taken by Farhangian University to establish a national criterion for teacher assessment and certification (i.e., PABT). The criterion requires that teacher candidates be fully certified by FU evaluation program using a comprehensive measure that includes four major components: 'Written Assessment, Performance Assessment, Portfolio Assessment, and GPA'. Since the PABT (or ASLAH) is a new assessment project and has highly significant consequences, developing reliable and comprehensive measures, as an essential part of the project, is a prerequisite for the success of the project. As one of the PABT four components, 'Performance assessment' has been the focus of a recent study by Kiany, et al. (2016) who developed measures for assessing performance of English-major teacher candidates. The present study focused on the first component of the project, i.e., 'Written Assessment'. It should be noted that the four major assessment measures of the PABT program are targeted to be employed at the end of the four-year training program to assess student-teachers' teaching competencies. Through establishing the assessment and certification program, 'FU's teacher evaluation system' is intended to become more comprehensive and accountable.

The study was an attempt to propose a hypothetical framework for a 'written test' supposed to assess 'English-major teacher candidates' competencies' in the form of 'knowledge base prerequisites' needed for certifying an English-major student-teacher as competent. Additionally, the validity of this initial assessment framework as the outcome of the first phase of the study was also examined. In other words, this study is part of a larger national project (i.e.,Farhangian University national PABT project) which was conducted in the FU context for the purpose of a) developing a written assessment scheme for certifying English-major student-teachers, and b) validating it with a sample of English-major student-teachers of Farhangian University.

2. Literature Review

A review of the available literature on 'teacher development' reveals that one important factor contributing to teacher quality is 'evaluation' of teachers' knowledge and their teaching abilities and skills (cf. Kirkpatrick & Kirkpatrick, 2006; McCaffrey, et al., 2003; Neild & Farley-Ripple, 2008). In this line of growing literature, the importance of teacher evaluation and teacher quality has been extensively discussed in relevance to the importance of teaching recipients, i.e., students, and their outcomes (Goldhaber & Anthony, 2007; Hanushek & Rivkin, 2004; Mangiante, 2011; Marsh & Hattie, 2002; Nye, et al., 2004). While acknowledging many

different factors affecting student outcomes, such as curriculum, funding, class size, and parental participation, Stronge and Hindman (2003) argued for teacher quality as the most influential school-based factors affecting student learning and achievement. Relevantly, Rivkin, et al. (2005) believed that teachers are one of the most important substantial components of a quality education system. Likewise, Freeman and Johnson (1998) highlighted the importance of language teachers when arguing that "lagging behind by almost a decayed, language teacher education has begun to recognize that teachers, apart from the method or materials they use, are central to understanding and improving English language teaching" (p.402). In other words, putting the right people in positions of classroom leadership is an important first step to improve student achievement (Reeves, 2007).

From these arguments and the views that take "better teaching is the key to higher student achievement" (Kaplan & Owings, 2002, p. 7), it is understood that enhancing teacher quality seems to be one of the essential missions ahead of teacher education and development programs. 'Teacher evaluation' is one of the most important tools that states and districts can employ for undertaking this mission and for improving the quality of education for all students (Ribas, 2005). That is, teachers are important object of evaluation to make sure they are qualified enough for teaching. Only after such an evaluation one can say whether an education program has worked appropriately (Kirkpatrick & Kirkpatrick, 2006).

Yet, how to evaluate or what to evaluate is the basic concern in teacher evaluation and quality assurance. Most evaluation affairs provide evidence on teacher observable attributes, preparation, and credentials (Goldhaber, 2002; Neild & Farley-Ripple, 2008). The most researched areas are teacher experience and education levels because the two are easily observable and obtainable (Goldhaber, 2002). There are, however, some studies examining teachers' professional qualifications indicating that teachers' quality of preparation, and their credential affect student achievement (Darling-Hammond, 2000; Darling-Hammond & Ball, 2004; Hanushek, 1997; Heck, 2007; Wenglinsky, 2002). More specifically, teacher pedagogical knowledge, content knowledge, and educational attainment are among the most widely evaluated areas.

Various methods and tools, from examining teachers' observable attributes to assessing quality of their preparation, their knowledge base, and their credentials have been used in order to guarantee evaluation effectiveness. One of the worldwide used methods applied for such purposes is 'licensure testing'. Some countries have intensified the educational and academic requirements for teacher candidates by setting licensure tests policy. For instance, Britain licensure testing policy mandates licensing examinations in addition to the final examinations in the colleges (Ross & Hutchings, 2003). These licensing requirements have been formulated into three categories of standards: (1) educational behaviors, attitudes, and values; (2) knowledge and understanding in education and pedagogy; and (3) practical teaching skills. In other words, for being awarded 'Qualified Teacher Status (QTS)' in Britain, teachers must meet Standards for Teaching (e.g., good subject and curriculum knowledge, inspiring, motivating and challenging students, fulfilling wide professional responsibilities, etc.), and Standards for Personal and Professional Conduct (e.g., showing tolerance and respect for the rights of others, building mutual relationships with students, valuing national values such as democracy, liberty and mutual respects, etc.). British teacher candidates must demonstrate that they have met these requirements as prescribed in Britain's teacher education policy.

In France, individuals need to meet three conditions in order to be eligible to enter the teaching profession, such as: 1) earning an academic degree; 2) passing a series of examinations in the disciplines they wish to teach; and 3) passing a year of internship. The teaching license is given only after all these three conditions have been fulfilled (Libman, 2009). Similarly, in the United states of America forty-two states mandate teacher candidates to earn a teaching certification by passing different kinds of licensure tests such as Praxis Series Tests (Praxis I, II, and III), Illinois Licensure Testing System (ILTS), Missouri Pre-Service Teacher Assessment (MoPTA), the National Evaluation Series (NES) (ETS, 2010; Mitchell, et al., 2001). Certification, by virtue of such exams, includes assessing general knowledge and teaching skills, and in some cases coursework and teaching practice (Boyd, et al., 2007). In the U.S., therefore, "most states administer teacher licensing examinations as a kind of guarantee that teachers know enough about their subjects" (Mitchell & Barth, 1999). For instance, PRAXIS II includes S*ubject assessment tests* that measure general and subject-specific knowledge and basic skills, *the Principles of Learning and Teaching test assessments* that measure general pedagogical knowledge, and *the Teaching Foundation Tests* that measure five areas, namely, English, language arts, mathematics, science, and social science.

The teacher evaluation literature indicates that, to date, different tests have been particularly developed for measuring teachers' general, content and pedagogical content knowledge (Baumert et al., 2010; Hill et al., 2004; Krauss et al., 2008; Sadler, et al., 2013; Schmidt et al., 2007). For instance, in an attempt to measure teachers' content knowledge of language and reading, Moats and Foorman (2003) studied teacher knowledge of reading-related concepts and, in a 3-phase process, constructed a measure of teacher content knowledge in language and reading. They aimed to develop a measure which discriminates more competent from less competent teachers regarding language and reading content knowledge. Phase 1 of their study included a measurement of k-2 teachers' content knowledge (n=50). In the second phase, an administration of 'teacher knowledge of language and reading test' to 41 second- and third-grade teachers was done as a pilot of the measure. Phase 3 was the administration of the measure, after the refinement and expansion of the two tests, to 103 third- and fourth- grade teachers. The results of their study showed that the misconceptions about sounds, words, phonemes, and sentences were pinpointed so that these issues could be addressed in future professional developmental teacher training programs.

Given the obvious importance of licensing, most of the English language teaching professions use a licensing system for certifying those teacher candidates who have attained the minimal degree of English language competency necessary to ensure the quality of their prospective instruction. Aligned with this line, the local policymakers, planners and officials of teacher education in Iran have set a specific 'teacher licensure testing program' as a means for certifying teachers' professional knowledge. Given that these means are only one side of the overall quality aimed by English teaching profession, the local reform policies of the profession has recently focused on applying three strategies to ensure the quality of English language teaching context, such as: *gatekeeping assessment* (through entrance examinations), *education programs*, and *certification granting*. These strategies, at an aggregated level, play a key role in the prospective English teachers' preparation. The third strategy, called PABT or ASLAH project in its local words, introduces means to provide the public with a mechanism for identifying those practitioners who have met particular standards of FU's teacher education

policies (FU Archives, 2014). As mentioned before, the project targets four criteria for the purpose of teacher credentials and certifications: Performance Assessment, Written Assessment, Portfolio, and GPA. This study aimed at working on the second criteria, i.e., Written Assessment component as part of the bigger PABT/ASLAH project.

Thus, being grounded in current issues on licensure test development as well as taking insights from research on teacher evaluation and teacher quality assurance, the present study took initiatives in developing an evaluation scheme for FU's English language student-teachers'. For such an important purpose, the Policy Committee from Evaluation and Quality Assurance Department of FU, an advisory panel of FU's teacher educators, subject-matter professors, test development specialists and practicing teachers were involved in the multistage development and validation processes of 'written tests' as measures for evaluating teacher candidates' knowledge base and competencies. The tests assess basic skills of knowledge and subject-matter knowledge, thus, their content varies from the assessment of basic reading, vocabulary and idiomatic expressions to deep subject-matter knowledge in teaching areas (e.g., teaching methodology, second language acquisition, linguistics).

3. Method

3.1 Context of the Study

The study took place at Farhangian University (in Iran) which is an institution mainly responsible for educating teacher candidates for the prospective teaching. Farhangian University's department of English language teaching offers a full four-year under-graduate educational program to individuals willing to become English teachers. In order to obtain an under-graduate degree, they are required to take 4272 hours and 149 related credit courses. The department of English language teaching consisted of both tenured faculty members and instructors/lecturers.

Having admitted at FU through 'State University Entrance Examinations' which are administered each July, student-teachers go through a four-year undergraduate teacher preparation program. The program targets four competencies introduced by the National Curriculum Document of Farhangian University, including Content-knowledge competency, Pedagogical-knowledge competency, Pedagogical content-knowledge competency and General-knowledge competency. Each semester, students are required to pass the final achievement exams (all four competencies included). Based on FU's recent policy, at the end of the four-year undergraduate program all student-teachers are expected to pass the PABT/ASLAH requirements

As previously mentioned, this study was part of the larger project of PABT/ASLAH, with the aim of developing a written assessment framework for evaluation of English language student-teachers' competencies. This written test was used to gauge teacher candidates' readiness for prospective teaching in terms of their knowledge base. The present study specifically focused on the key components of an English written assessment scheme. Overall, the present study was supposed to progress based on a two-phase schedule: development phase and validation phase.

3.2 Participants

The total sample, amounting to 397, included both females and males from three different educational layers: FU's educational policymakers and planners; university instructional layer including teacher educators and experts (N=15), and English teacher and student-teachers. These participants were selected from different centers of Farhangian University.

Teacher Educators were mainly from FU's centers of Tehran (N=10) and Zanjan (N=5). The reasons for selecting teacher educators from these two provinces were availability, accessibility, and time scheduling concerns. In general, all teacher educators were purposefully selected using a snowball method of sampling. Convenience sampling was also a part of the main sampling in that the ease of accessibility and possessing the characteristics required by this study were of the main concerns of the researcher. Classroom experience and subject matter competency were the criteria to be met. Teacher educators and experts with teaching experience above 10 years were selected ($\bar{x} = 21$ years, with the range of 11 to 32). In total, the participant community included 9 males and 6 female, 10 with PhDs and 5 with MA degree.

The study also included **English language teachers** as another layer of participants. Convenience sampling, along with snowball sampling was employed to select the sample that finally comprised of 320 English teachers, 142 females (44.4%) and 178 males (55.6%), 143 with BA. (44.7%), 147 with MA. (45.9%), and 30 with PhD (9.4%) degree. The teaching experience as reported by the participants showed 83 teachers had 0-4 teaching years (25.9 %), 167 teachers with 5-12 teaching years (52.2 %), 57 teachers with 13-20 teaching years (17.8 %), and 13 teachers with more than 20 years of teaching (4.1 %).

The third layer of participants included a total number of 62 **English language student-teachers** who were on the verge of graduation in academic year of 2016-2017. A purposive sampling procedure was adopted, and only those student-teachers who were in their last year of the four-year education program could participate in the study. They were mainly from three FU centers (faculties) in Tehran Province: namely, Shohadaye Makkeh, Bahonar, and Moffateh centers which have educated English language student-teachers for years.

3.3 Instrument and Procedures

The methods used to collect data for this study consisted of document analysis, interviews, surveys, and written tests. Interviews consisted of semi-structured interviews that were used to gather qualitative data. Document analysis comprised of the analysis of higher-order policy documents of Farhangian University and the Ministry of Education. Surveys consisted of 5-point Likert Scale questionnaires developed by the researcher for the purpose of the present study. Finally, Written-tests, comprised of Multiple Choice (MC) items developed as a result of the data collected through the aforementioned instruments, were employed to collect the data related to the second aim of the study.

i. National Policy Documents

The following educational policy documents were used in the first phase of the study: ME's higher-order documents such as the Document of Fundamental Reforms in Education (DFRE), and Farhangian University's higher-order documents such as the FU's National Curriculum Document and its ELT Curriculum Document.

Accessed through the Ministry of Education website, the DFRE was used as a primary source for extracting themes, topics or questions that might be helpful for the construction of the interview instrument. The reason behind employing this document was because it includes an additional informative section on PABT policy and on the role of teacher evaluation in bringing about positive changes in education.

Other documents used as man sources for the construction of both the interview and the tests questions were the National Curriculum Document of Farhangian University and its ELT Curriculum Document. Collected from FU's Quality Assurance Department, the two documents were content analyzed to extract the main themes required for teaching competencies. Four main teaching competencies are highlighted by the documents: a) general knowledge competency (GK), b) general pedagogical knowledge competency (GPK), c) content knowledge competency (CK), and d) pedagogical content knowledge competency (PCK).

General knowledge competency (GK): Understanding issues related to culture, religion, language and politics can be considered as general knowledge competency. This competency provides teachers with the needed knowledge to analyze the situation and provide appropriate decisions.

General pedagogical knowledge competency (GPK): It is comprised of educational knowledge and understanding of the principles and methods of applying it in different situation and the ability to recognize different situations and educational provisions in accordance with the actual situations of education (educational science), and the ability to apply educational research methods in solving simple educational problems (educational research).

Content knowledge competence (CK): a relative dominance over the processes and outcomes of domain-specific knowledge; and in some other areas of knowledge (depending on the curriculum and learning requirements of the National Curriculum).

Pedagogical content knowledge (PCK): a competency resulted from a tight interrelatedness and understanding of subject-matter knowledge and pedagogical knowledge which can be used in pedagogical situations and school-based pedagogy in a specific major.

ii. Interviews

The aim of this interview schedule was two-fold: a) to provide information from the instructional layer regarding their views about FU's new assessment project, in general, and the written assessment component and its sub-components and structure, in particular; and b) to help develop a valid instrument for measuring teacher candidates' level of competencies.

The following sources informed designing a valid instrument for collecting the interview data:

-The Document of Fundamental Reforms in Education (DFRE),

-FU's ELT Curriculum Document,

-FU's National Curriculum Document,
-Related literature; and
-The experience and knowledge of the researchers about Farhangian University evaluation policies.

Interview protocol for teacher educators: Interviews with 15 English language teacher educators and experts were conducted from June to July 2016. Scheduled appointments were set to conduct one-to-one recorded interviews with each individual informant teaching different subject matters at different branches of Farhangian University. The interview guide included 8 questions seeking 'how teacher educators felt the new written assessment project measure student-teachers' professional knowledge', about its consequences and influences on teacher quality, its fairness, also about 'their perspectives of whats and hows of the tests' (See Appendix).Before conducting the interviews, the content and face validity of the interview questions were reviewed by one test designer from Farhangian University and one from Tarbiat Modares University. Based on these experts' comments, few modifications in the wording of the questions were made.

All interviews were audio-taped and transcribed soon after recording. Interview data were significantly valuable since they provided the researcher with a thick data stream on 'the whyness and whatness of the written assessment', 'the framework components', 'the number of courses to be included', 'the number of items for assessing each course', 'the type of the questions', as well as 'the precautions to be taken in the development of the written test'.

Later, with the help of this data pool, the researcher was able to construct the instrument required for data collection of the quantitative phase, i.e., 'the Questionnaire on Written Assessment (QWA)'.

iii. Questionnaire Surveys: QWA

The interview data facilitated the construction of the QWA (see Appendix) used to collect quantitative data in the present study. The 26-item QWA was developed for gathering quantitative data and enriching researchers' understanding of the way the written test must be developed. It consisted of three subscales; 'impact and consequences of the proposed written test (18 items taken from a reliable and valid instrument developed by Shayestefar in 2013)'; 'content and structure of the test' (7 items eliciting data on the knowledge/competency types to be included in the written tests, their importance, the relevant courses, and the number of items for each course); and 'personal information' (e.g., gender, degree, years of teaching). Participants were asked to provide their answers on a 5-point Likert-type scale (1=completely disagree, 4=completely agree) for 22 items, and a 3-point Likert-type scale (1=low importance, 2=mid, and 3=high importance) for the rest of the items.

iv. Assessment Scheme: Written Tests

A measure of attainment of English language teaching competencies was aimed to be developed and proposed by this study. These national measures are supposed to be provided by FU as standard tests for all groups of FU's English student-teachers who are supposed to start their

teaching career at schools. To design the initial framework of this MC item measure, detailed content analysis of 9 major courses that were suggested by the interviewees and FU's curriculum were done. The courses are Reading, Vocabulary, Grammar, Linguistics, Phonology, Methodology, Language Testing, Research Methodology, and Second Language Acquisition-SLA. Regarding the content coverage of each course and the required skills, the tests followed the prescribed detailed specifications of the ELT curriculum of Farhangian University. What follows shows an example of a subtest of the overall written framework that finally contained 108 items:

> Reading: It is consisted of three main parts i.e. Pre-reading activities, While-reading activities and Post-reading activities. Out of total 108 items, 12 items are allocated to this subtest. The first two sections (Pre-reading activities and While-reading activities) contain 8 items, 4 items assigned to each respectively. The last section (Post-reading activities) is assessed through 4 related items. The weights of all the items are equal; one score goes to each.

3.4 Analysis

i. Interviews

The audio-recorded data of 15 teacher educators and experts were listened to repeatedly, and then transcribed with care. Thus, the initial step for qualitative data analysis was to review and reread interview notes and transcripts. Participants' confidentiality was maintained through the use of pseudonyms. Teacher educators' interviews played an important source for the creation of '*knowledge-base categories and subcategories'* in this research. Based on the analysis and re-analysis of the transcripts, the researcher designed a preliminary outline to organize the extracted information. Data analysis of this qualitative phase was carried out through using 'inductive analysis procedures' that includes identification and segmentation of the transcripts into separate chunks that are subsequently classified into distinct thematic categories.

The researcher gave each category of information tentative names as he read and reread each transcript. As the transcripts were reviewed, additional themes emerged. Glesne (1999) states, "As the process of naming and locating your data bits proceeds, your categories divide and subdivide…in the early days of data collection, coding can help you to develop a more specific focus" (p.133).

Some of the emerged categories/themes, for instance, were 'quality enhancement', 'professional development', 'effectivity', and 'necessity'. While reading transcripts, some more arguments emerged and these would form new categories/themes. Examples of these were quotes on 'quality of the test' and others related to 'validity', 'reliability', and 'fairness' of the new assessment measures.

ii. Questionnaire Surveys

For the purpose of standardization of the QWA, both exploratory and confirmatory factor analyses were run. Later analyses were done in accordance with the research questions, that is, the data were analyzed in quantitative terms such as reliability estimation of the instruments,

means, frequencies, and normality of distribution by means of SPSS, and confirmatory factor analysis by means of AMOS Software 18 (Arbuckle, 2009).

Having been checked for content validity, the resulting version of the Likert QWA was administered to teacher educators and teachers. An analysis of reliability yielded a Cronbach's alpha value of .78 that is an acceptable index of reliability coefficient.

To examine the internal structure of the QWA Likert-type items, Exploratory Factor Analysis (EFA) was run for its sub-scales. But before this process, the data were checked for factorability through KMO (Kaiser-MeyerOlkin Measure of Sampling Adequacy) test in SPSS. The KMO value was 0.81, exceeding the recommended value of 0.6 (Pallant, 2013), indicating the factorability of the data. The test 'instructional objective scale', 'fairness scale' and 'quality of test results scale', with a total of 18 items, were then factor analyzed with running Principal Component Analysis (PCA). Other items tapping personal information or preferences were not, by nature, included in the analyses. The findings made it discernible that 5 factors with Eigenvalues greater than one could be extracted, explaining 54% of the variance in the pattern of relationships among the items. The percentages explained by each factor were 23.133%, 9.969%, 7.589%, 6.723%, and 6.191%, respectively. The two last factors were excluded from the finale model because of the item loading (having items with smaller loadings than the other factors).

In this study, three factors/sub-scales-**Instructional Objectives** (cognitive & affective outcomes of the Written Assessment Scheme, Items: 1-8), **Fairness** (Items: 9-13), and **Quality/Clarity** of the Written Assessment Scheme (Items: 14-18)-were used to determine the structural pattern in the 18-item scale of the 'Impact and consequences' of the written tests (first section of the QWA). The results of EXA revealed that four items were finally eliminated (items 7, 8, 15 and 17) because they failed to meet a minimum criterion of having a primary factor loading of .30 or above. An examination of the content of the items of this section (section I) of the QWA provided empirical structure for the existence of 3 above-mentioned sub-scales.

It is noteworthy that while these 3 factors/sub-scales emerged in EFA, Confirmatory Factor Analysis (CFA) with AMOS (version 18) was carried out to determine the adequacy of the factor loadings and more information about the structural measurement that could not be provided through EFA-SPSS. Figure 1 presents the measurement model for the variables of 'Impact and consequences' of the proposed written test (section I of the QWA). For this constructed measurement model, factor loadings were allowed to load on only one construct (i.e., no cross loading).

 ISSN 2476-5880

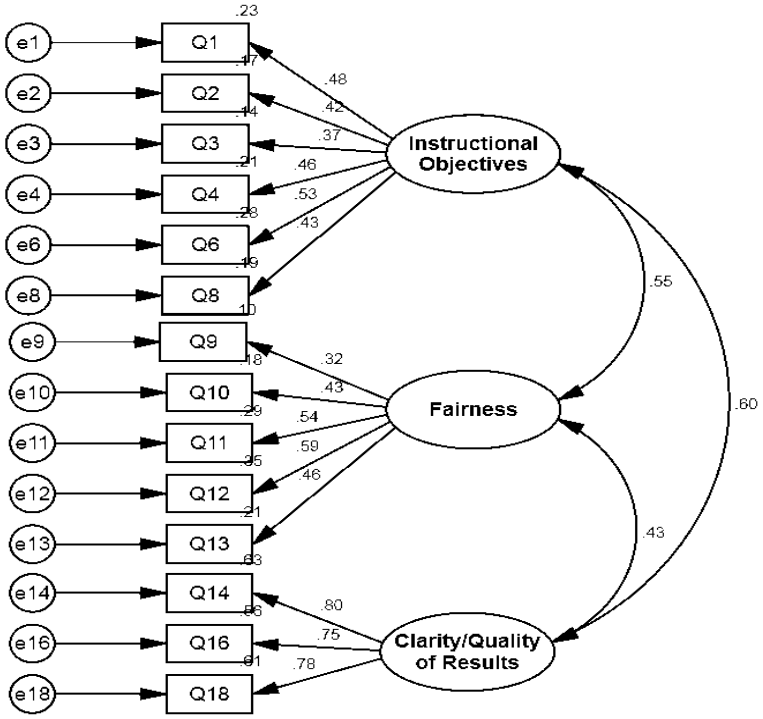

Figure 1. Final confirmatory factor analysis model of impact and consequences of the written assessment measures

The results of CFA showed the obtained Goodness-of-Fit Index (GFI) fell within the acceptable value. The values of .90, very close to it, or beyond for GFI and CFI (Comparative Fit Index) are regarded as indicators of a good fit in CFA-AMOS (Byrne, 1994; Mulaik et al., 1989). In addition to the normed Chi-square, CFI and GFI that are usually reported in CFA-AMOS studies, the RMSEA (Root Mean Square Error of Approximation is another indicator of a model fit. The recommended values of these indices are reported in the literature (Arbuckle, 2013; Ghassemi, 2010). The estimates of the QWA are reported in Table 1. Examining this table, it appears that CFI and GFI are greater than the 0.90 cutoff point. Bearing in mind that the closer the value to 1, the better fitness (Stapleton, 1997), the scale of 'Impact and consequences of the written tests' shows a good fit to the observed data. Table 2 shows the elaborated estimation of the measurement model parameters achieved through CFA-AMOS (including loading estimates, standard errors, squared multiple correlations and critical ratios).

Table 1. Results of Goodness-of-Fit Indices for the QWA

Goodness of fit Indices	**Recommended Value**	**Finale/Modified Estimates**
CMIN/DF	1>, <3	1.902
CFI	>.90	.912

GFI	>.90	.938
RMSEA	>.08	.053

Table 2. Parameter Estimates of the Standardized Factor Loadings, Standard Error (SE), Critical Ratio (CR), and Squared Multiple Correlations (SMC) for the Measurement Model (Impact and consequences of the supposed-written test)

Items		Factors	Estimate	S.E.	C.R.	P	Label
Q8	<---	Instructional Objectives	1.000				
Q6	<---	Instructional Objectives	1.088	.210	5.183	***	par_1
Q4	<---	Instructional Objectives	.967	.205	4.724	***	par_2
Q3	<---	Instructional Objectives	1.438	.343	4.193	***	par_3
Q2	<---	Instructional Objectives	.893	.199	4.481	***	par_4
Q1	<---	Instructional Objectives	1.031	.205	5.035	***	par_5
Q13	<---	Fairness	1.000				
Q12	<---	Fairness	1.212	.218	5.569	***	par_6
Q11	<---	Fairness	1.178	.238	4.949	***	par_7
Q10	<---	Fairness	.960	.212	4.534	***	par_8
Q9	<---	Fairness	.917	.248	3.695	***	par_9
Q18	<---	Clarity/Quality of Results	1.000				
Q16	<---	Clarity/Quality of Results	.860	.069	12.424	***	par_10
Q14	<---	Clarity/Quality of Results	1.010	.079	12.820	***	par_11

All standardized regression weights (equivalent to factor loading) are significant with Critical Ration (CR) CR>1.96, P-value <0.05, and all the error variance ≤1.0, indicating no violation of estimates (Al-Shabatat, Abbas & Ismail, 2010). These values indicate that out of initial 18, 14 measurement variables/items are significantly represented by the 3 latent variables, i.e., **Instructional objectives**, **Fairness**, and **Clarity/ Quality** of the tests results.

iii. The Written Assessment Tests

The present study included development and initial piloting of the assessment measures as the outcome of the PABT project. As such, it was needed that these designed measures (tests) be administered among the participating student-teachers for the piloting and validation purposes. After test performance data together with test-takers' views about the tests were collected from this specific sample, detailed item analyses were conducted along with content and expert validation of the tests. Moreover, the reliability of the overall written test was examined.

4. Results

4.1. Phase I: Determining Content and structure of the assessment framework

The first aim concerned the components of the proposed framework for assessing the professional competence of FU's student-teachers. For such an aim, a triangulation strategy was first adopted to obtain data from multiple sources, including: (1) relevant higher-order national documents; (2) interviews with Farhangian University ELT teacher educators; and (3) teacher surveys.

i. Content Analyses of the FU's Documents: ELT Curriculum and National Curriculum

As mentioned before, two higher-order national documents, that were relevant to teacher education and teacher quality, were content analyzed through a 'deductive content analysis' approach to extract the related themes. The emerging themes were re-examined and reconsidered to obtain convergence of the themes in a consistent picture.

The National Curriculum Document (NCD) defines competency as "*a collection of traits and skills related to all aspects of identity which teachers are in need of for appreciating the situation and improving it (on the basis of Islamic criteria) in order to achieve pure life*" (p.4). According to the content-analyzed NCD, for teachers to be accredited as competent, they must professionally possess the following four major competencies, including Content knowledge competency (CK), General pedagogical knowledge competency (GPK), Pedagogical content knowledge competency (PCK) , and General knowledge competency (GK).

The NCD is a fundamental basis for designing the required courses for each FU major. Of the four main competencies highlighted by the NCD, two, i.e., GPK and GK are common to all FU's majors/disciplines, but not to domain-specific courses of English-Major students. That is why the courses categorized under these two competencies are taught and assessed in Persian. Informed by this and by the present experts' views on nature, content and intended goals of these two competencies, it was decided to exclude them from the 'assessment framework' planning and development. Consequently, the other two competencies, i.e., CK and PCK were targeted for further investigation.

The ELT Curriculum Document of Farhangian University (ELTCD) provides short descriptions for each of the ELT core courses, along with the instructional objectives, curricular planning, and suggested materials for teaching each course. Table 3 indicates the number of course credits required for an English major teacher candidate to pass during the 4-year education program. The result of content analysis revealed five types of courses have been specified by the ELTCD: General courses, Islamic pedagogical courses, General pedagogical courses, Major-specific courses, Elective major-specific courses. At first glance, it is clear that the major-specific course credits have the largest number of credits, in other words major-specific courses account for more than half of the overall credits; 87 out of 148. General course credits have the second largest number of credits; 24 out of 148. Islamic pedagogical course credits (19) and general pedagogical course credits (18) come next, respectively.

Table 3. No. of FU's English Curriculum Credit Courses

Courses		No. of Credits
- Approved by Minister of Science, research and technology - Specific to Farhangian university courses	General courses	21 3
- Islamic pedagogical courses - General pedagogical courses - Major-specific courses - Elective Major-specific courses		19 18 85 2
total		148

Table 4 represents the courses that are identified with CK and Table 5 shows those courses that belong to PCK.

Table 4. Content Knowledge (CK) related courses (taken from ELTCD)

Competency	**Courses**	**Courses**
Content Knowledge (CK)	- Reading Skill (1,2) - Listening Skill (1,2) - Grammar (1,2) - Speaking Skill (1,2) - Writing Skill (1,2) - Media Understanding skill - Creativity through Literature (1,2) - Teaching Material Development	- Phonology - Linguistics (1,2) - Advanced Writing - Translation Skills - Vocabulary - Research Methodology - Language Teaching Methodologies - Teaching Language Skills

Table 5. Pedagogical Content Knowledge (PCK) related courses (taken from ELTCD)

Competency	**Courses**	**Courses**

 ISSN 2476-5880

Pedagogical Content Knowledge (PCK)	- Teaching Philosophy in ELT - Teaching Strategies in ELT - Educational Testing in ELT - Information Technology and Communication Application in ELT - Research and Professional Development in ELT - Project	- Lesson Planning in ELT - Educational Designing in ELT - Teaching Material Content Analysis in ELT - Special Professional Experiences in ELT - Internship

ii. Interview Results

As described before, FU teacher educators and experts who showed their consent to participate in the study were interviewed for two purposes. One was to take advantages of their knowledge, views and experience in teacher training and teacher evaluation to identify the components of teacher knowledge base. Another incentive came from the need to solicit for their appraisal of the overall hypothesized framework, its components and the constituting items (i.e., validity purpose).

As to the units underlying the analysis results, the participants reported 72 themes that were then categorized under 20 units, that were then categorized under 9 main categories, including: '**teacher competency'**, **'quality enhancement'**, **'professional development'**, **'test effectiveness'**, **'test necessity'**, **'test overall quality'**, **'test validity'**, **'test reliability'**, and '**test fairness'**. These categories showed not only the components to be assessed through a written test but also the features of a quality test.

A synthesis of the information obtained from the sources sketched above, i.e., the relevant literature, higher-order documents and interviews resulted into an image of 'teacher knowledge base', its components and number of items' to be included in the hypothesized framework of written assessment. Participants were then solicited for their degree/s of agreement with these proposed constituents, and the significant of each one, also for any further suggestions. All of the 15 interviewees 'agreed' to have CK as one of the essential components of the test. The number of items suggested for this type of knowledge varied. A mean of 102 items was proposed by the participants.

In addition, most of the interviewees (8 out of 15) found pedagogical knowledge (PK) of high importance, however, all of them (14), but one, agreed that this knowledge needs not to be included as one of the constituents of the written test. Therefore, they suggested that it is better not to have items related to this type of knowledge in the test. Moreover, despite the fact that most of the interviewees supposed PCK as very important, they disagreed to have it in the test. Hence, CK was the only knowledge type agreed upon by all of teacher educators. Nevertheless, it was also needed to solicit for more ideas from other layer of the program's participants, i.e., student-teachers, to make sure if they had the same idea. Therefore, the next stage focused on the

inclusion of pedagogical content knowledge (PCK) as one of the variables of the hypothesized assessment scheme.

Specifically, the researcher asked teacher educators' opinion about the possible courses that could be categorized under CK and PCK competencies in order to be assessed through the proposed measures. For such a purpose, the extracted lists of the courses from the ELTCDFU (see above) were presented to the teacher educators, asking them to prioritize the courses on the basis of the importance they perceived each course would have. Table 6 shows the results.

Table 6. Prioritized Courses in each competency category

Competency	**Prioritized Courses**	**Views on the degree of Importance & the need to be assessed**
CK	- Reading - Writing - Grammar - Vocabulary - SLA - Teaching Methodologies - Testing - Research Methods - Phonology - Linguistics	Highly important. To be necessarily assessed through the written tests.
PCK	- Teaching Philosophy in ELT - Teaching Strategies in ELT - Lesson Planning in ELT - Educational Testing in ELT	Important, but No need to be assessed through the written tests.

As Table 6 shows, when CK was concerned, language skill courses (e.g., reading, writing, grammar), Second Language Acquisition, Teaching Methodologies, Testing, Research Methods, Phonology and Linguistics were those courses that received teacher trainers' high consensus. As to the PCK, although they rated four courses as important (i.e., Teaching Philosophy, Teaching Strategies, Lesson Planning and Educational Testing) most of them did not actually feel the necessity to assess these courses. They believed these courses are demonstrated through Performance Assessment component of ASLAH/PABT and no need to double test them via the Written Assessment measures.

iii. Questionnaire Surveys: Teachers and student-teachers' views on the test structure and content

Second part of the WTQ (items 19 to 25) focused on the content and structure of the hypothesized written framework. In order to ask teachers' opinions about the content and structure of the proposed test, the QWA 5-point Likert-scale items ranging from 'strongly disagree' to 'strongly agree' were used. Items 19, 20, and 22 represent participants' agreement on assessing content knowledge, English general proficiency, and pedagogical content knowledge through the test, respectively.

For item 19, '*To what extent do you agree to assess content knowledge (CK) through written tests?*', 54 % of the teachers agreed, 21.6 % disagreed, and 24.4% had "no idea" As to the item 20, *'To what extent do you agree to assess English general knowledge through written tests?'*, 56.8 % of the participants reported agreement rate, 23.8 % disagreed, and 19.4 % selected 'no idea'. As to the item 22, '*To what extent do you agree to assess pedagogical content knowledge through written tests?*', only 7.2 % of the participants reported their agreement, 72.6 % disagreed, and 20.3 % had 'no idea'. Notwithstanding PCK is acknowledged to be one of the important competencies to be acquired for quality teaching, neither the teacher educators nor the teachers were positive about measuring it through the hypothesized test.

When asked to give their ideas about the number of items appropriate for each course (i.e., subtests of the overall written test), the participants reported a mean score of 11 for Reading, Linguistics, Phonology, Teaching Methodology, Language Testing, and Research Methodologies), and 12 SLA, Grammar, and Vocabulary.

After the main components and their constituents (i.e., types of knowledge, courses, number of items assigned to each course, and types/forms of the items) were determined, the study proceeded towards developing a table of specification for each of the above 9 courses.

iv. Test Specifications

Following the assertions that highlight the significance of test specifications for arranging test outlines (cf., Bachman & Palmer, 2010; Raymond & Neuste, 2006), the next stage of the study concerned developing a table of specifications for the proposed framework. In fact, much in line with the way that academic disciplines are organized, test specifications helps outline the topics that examinees are expected to master (Raymond & Neuste, 2006). The following sources were used to develop an initial table of specifications.

The ELT Curriculum Document of Farhangian University (ELTCD): Inasmuch as FU's ELT student-teachers are to be assessed based on the ELT curriculum of the university, the document was further content analyzed in order to propose a detailed table of specification for each of the targeted courses. The ELTCD proposes a syllabus that was employed by this study to develop an initial scheme for table of specifications of each course. Drafts of an initially developed framework for table of specifications were sent to the subject-matter teacher educators to put their comments on.

Subject-Matter Teacher Educators: the hypothesized framework for table of specifications was then subjected to a review by subject-matter experts. 'Topics', 'weights', 'number of items', 'cognitive level of items' and 'time spent on each item' were the variables to be judged at this stage. This was done in accordance to judgmental weights that require "a group of experts provide direct, holistic judgments regarding the number or percentage of test items per section" (Raymond & Neuste, 2006, p. 213). This study adopted a top-down approach to elicit judgmental weights from the FU's ELT experts. They were asked first to assign 'percentages' to each of the 'main sections' of each course previously decided upon on the basis of the ELTCD. Next, they were asked to assign percentages to the topics specified for each section. Based on the initial weightings offered by the experts, the sections and topics to be included in each of the 9 courses

were identified. Then it was time to assign a number of items to each topic. The next step of test designing stage was to specify the items' cognitive level on the basis of the 'Revised Bloom's Taxonomy' (Anderson & Krathwohl, 2001) in which each of the items is placed on one cognitive level (e.g., synthetic, evaluation or knowledge level).

After the data were collected from the experts, the obtained ratings were carefully reviewed, checked against the ELTCD syllabi, and verified by the team of the present research. The outcome was presented in 9 tables of specification (Table 7 presents an example of table of specifications developed for 'Reading').

Table 7. Detailed specifications for 'Reading Comprehension' measure

Learning/Instructional Objectives		Weight		Topic	No. of Items	Level 1	Level 2	Level 3	Time Spent	Score / point
Reading	Section 1 Pre-reading Activities	33 %	2	Applying previewing	4		*	*		4
			1	understand skimming						
			1	Applying predicting				*		
	Section 2 While-reading Activities	33 %	2	Apply scanning	4	*		*		4
			1	Identifying text organization						
			1	Recognizing textual features		*				
	Section 3 Post-reading Activities	34 %	1	Evaluating generalizations	4			*		4
			1	Applying paraphrasing				*		
			1	Checking summarizing				*		
			1	Understanding outlining			*			
		Total 100%			12					Total 12

v. Setting Cut-off Score

It is argued that people do not receive a certificate or license until they have passed a pre-determined specified level (Mehrens & Lehmann, 1991). Since licensure tests for teaching are used to determine individual's level of teaching competence, a 'cut-off score' determining process must be at work to help interpret individuals' performance on licensure measures. To fulfill this requirement, the policy committee of the present project (the FU's Deputy of Evaluation and Quality Assurance, three FU's experts and two ELT Professors) initially reviewed the recommended cut-off scores by FU's assessment policies (>50% of the total score). However, before making the final decisions, an 'Angoff Method' as a widely-used model for setting a passing score (Livingston & Zieky, 1982) was adopted.

Angoff Method is a score setting method used by test developers to determine the passing score (cut-off score) for a test. Since the passing score of a test should be justified with empirical

data but not decided arbitrarily, the Angoff method is used to predict how many minimally-qualified candidates would answer each test item correctly. To this end, the method relies on informed subject-matter experts who examine the content of test items and independently estimate difficulty values for each item (Mellone & Faben, 2014). The sum of the '*predicted difficulty'* values for each item averaged across the raters and items is the recommended Angoff cut score.

After the cut-off scores were reviewed and initially decided upon by the policy committee of the present project, a copy of the developed tests were presented to the subject-matter experts for a review and determination of the passing scores. They answered each question and estimated the percent of qualified student-teachers who would get answer the item correctly. Their responses were averaged and became the cutting score. In this consensus reaching process, the passing score for the overall written test turned out to be 58 of 108 items. Therefore, A cut-off score of 57 or lower features 'minimally competent'/low competence', 58-90 indicates 'competent' and 91-108 represents 'highly competent' student-teachers.

vi. Item format and initial item pool

As to the forms of the items building the exams, attempts were made to solicit for the experts' views and expertise. This was specifically done through two separate items of the QWT (items 24 and 25). As to the item 24 asking for, '*the extent to which the respondents agreed on short-answer questions*', only 25.7 % reported agreement, 34.3 % disagreed, and the rest of them (40 %) selected 'no idea'. Regarding item 25 tapping '*the extent to which the respondents agreed on Multiple-Choice (MC) questions*', half of them (50 %) showed their high rate of agreement, and 25 % disagreed.

Teachers' ideas were also taken into account. When asked about using short-answer format as one of the alternative formats of the test items, 128 teachers (40%) expressed that they have 'no idea', and only one-fourth of them were positive (n=82; 25.7%), yet those who disagreed formed the second largest proportion of the participants (n=110; 34.3%). With regard to item 25 that assessed if this layer of participants agreed with inclusion of MC questions, 158 (50.4%) of them agreed to have MC items. Still, 80 (25%) of the participants fell on the 'disagreement' side of the scale, while 82 (25.6%) had 'no idea'. Results of Qi-square tests of difference among the observed frequencies showed that the groups of respondents were significantly different from each other in terms of their choices (χ^2 (4) =180.37; $p<.05$).

Apart from the participants' views and perspectives on the most appropriate item format for the present tests, the available literature on test design and item development as well as existing measures for assessing competencies (such as PRAXIS I & II) were also reviewed. Such a review revealed selected-response items as the most appropriate item formats when large-scale assessment of higher order cognitive abilities, achievement, and large domains of knowledge are concerned (Downing, 2006; Haladyna, 2004; Kane, 2006). Compared to this, constructed-response items, however, are much less efficient, typically produce less reliable scores, and may inadequately sample the content of the target domain, therefore reducing content-related validity evidence for the test.

Much in line with the relevant literature as well as informed by the subject-matter experts' expertise and teacher's views, the study then prioritized and proposed MC format for the tests.

Accordingly, all of the experts were contacted and asked to provide the present researcher (and the FU Evaluation and Quality Assurance Department) with the MC items that they have developed or collected for assessing their ELT student-teachers' knowledge of the above-mentioned 9 courses. Subsequently, an initial pool of 600 MC items, covering all the subject-courses in question, was obtained from them. This initial pool of items was carefully examined to check for consistency with the underlying aims of the written assessment plan, its sections, topics and levels, also for item quality. Meanwhile, an ELT professor with years of expertise in test development and validation assisted the researcher with reviewing, examining and selecting the corresponding items. Therefore, the most corresponding and plausible items for each of the courses were selected from the collected pool of items. Each of the designed tests was then given to one FU's subject-matter expert to comment on the test content, form and format of the test and its items. Based on the received comments, few items were modified and finally each of the 9 tests included 12 items for assessing teacher competencies in respective courses.

4.2. Phase II: Test Validation

The second aim of the study concerned the validity of the proposed framework developed for assessing the professional competencies of FU's student-teachers.
The outcomes of phase I were examined in phase II or the phase of the test validity. In other word, the 'quality' of the tests developed in phase I, and their probable 'impacts' on the ELT context of FU were examined through the eyes of the present ELT community stakeholders. In addition, 'content validity' and test 'Item Characteristics' were also explored to make sure of the validity of the developed tests. The analysis of interview data and the relevant results are presented below.

4.2.1. Teacher Educators and Subject-matter Experts: Interviews

The purpose of the qualitative data was to capture experts' reflections on and perceptions of the quality and relevance of the newly developed tests. Teacher educator and subject-matter experts were asked to participate in this interview stage of the study. Specifically, they were asked to express their opinions about the potential contributions and relevance of the written tests to the general quality of FU's four-year teacher education program and the quality of the program's graduates, in particular. Two related open-ended questions were asked to gouge into the participant's opinions on the issue.

- How do you evaluate the contribution of these 'written tests' to the issue of 'quality' of FU's English student-teachers?, in other words,

- Do you think these 'written tests' might have effects on the quality of FU's teacher education program (in terms of professional development and competency development)?

In a similar approach taken before for the interview analyses, informants' responses to these two questions were also analyzed through an inductive data analysis approach. Each interview transcription was analyzed separately, and eventually the responses emerged to be categorized into four major categories, each with its corresponding themes. The four categories

identified were: *(a) quality improvement*; *(b) knowledge development*; (c) *teaching preparedness*; and (d) *test development technical issues*.

Some of the extracts are:

> …*"As one of the criteria, I see it as a good tool as it helps student-teachers to update the knowledge they have learned at the end of their four-year teacher education program". (Participant #7)*
>
> … *"It definitely will have positive effects on the students. Even if student-teachers will not be able to answer the questions correctly, it would help them learn by realizing their problems and deficits." (Participant #13)*

However not every response was positive as evidenced by one of the teacher educators.

> "*When it comes to quality assurance in Iran, in my opinion, there must be a test at the end of teacher education program. As you know, student-teachers are becoming too happy-go-lucky with today's teacher evaluation system. As a matter of fact, they neither are nor even being assessed." (Participant #3)*

Overall, the teacher educators' responses were positive regarding the issue of assessing teachers' teaching knowledge through the present written tests. Based on the participants' responses, one can claim that a written test criterion, as one of the PABT's four criteria, is deemed necessary in order to come up with a comprehensive evaluation system through which quality assurance can be attained.

Furthermore, in respect to the potential effects of the tests, the teacher-educators' were positive and reported that the written tests can positively influence the overall FU system of teacher education and student-teacher, in particular.

4.2.2. Teachers' and teacher educators' views: QWA Surveys

In order to obtain a broader picture of the perceived fairness and quality of the proposed written assessment policy as well as its perceived consequences on 'learning objectives', inclusion of other representatives of the program seemed essential. Investigation of EFL teachers and student-teachers' perceptions of the tests consequences (N=320) was made possible through the QWA which was validated before (see Method Section). When asked about 'instructional objectives', 'quality of the test results' and 'test fairness', participants' answers indicated that they were not negative about the written tests' effects and consequences nor about the program fairness. However, they did not appear to be 'strongly positive', either ($X < 3.5$). The obtained means for 'Instructional objectives', 'Quality of Results', and 'Fairness' were 3.229, 3,271, and 3,254, respectively.

In addition to accumulation of validity evidence in support of the tests consequences (Section I of QWA), more evidence was needed to evaluate the practicality of the written assessment plan through the eyes of the teachers and student-teachers. Likewise, evidence

indicating how they perceived the importance of each course was deemed to be essential. This was done by means of the second section of the questionnaire instruments.

First, to see whether the written assessment policy is turned into practice in the real context of Farhangian University, the participants were asked to give their opinion about the 'practicality' concerns (item 25). Approximately, more than half of them (190; 59.4%) thought it practical and feasible to administer in the FU's context; 86 (26.9%) perceived very little practicality for the tests; and the remaining 44 participants (13.8%) appeared doubtful. Further, a Qi-square test of difference was used to explore any significance difference between the participants' views of the test practicality. The result showed that the difference in response proportions was significant (χ^2 (4) =87.19; p<.05).

4.2.3. Subject-matter Experts' Views: Content Validity

According to the latest edition of the Standards for Educational and Psychological Testing (AERA, 1999), tests that are used for credentialing intentions (licensure and certification) concentrate on a candidate's current skill, knowledge, or competency in a particular domain. Relevantly, according to Educational Testing Service "the process of licensure serves as a gateway into a profession, as a license is often required for entry into an occupation" (ETS, 2014, p. 2). Such evidence shows how important are credentialing measures in making inferences and decisions about the candidates' performance levels. Therefore, what is integral to the quality of these measures is the evaluation of their validity. According to American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (AERA), "Validation of credentialing tests depends mainly on content-related evidence, often in the form of judgments that the test adequately represents the content domain of the occupation or specialty being considered. Such evidence may be supplemented with other forms of evidence external to the test" (AERA, 1999, p. 157).

In a similar vein, Lissitz and Samuelsen (2007b) call on measurement professionals to reconsider their focus on construct validation and instead concentrate attention on content validity. They postulate: "we are attempting to move away from a unitary theory focused on construct validity and to reorient educators to the importance of content validity and the general problem of test development" (Lissitz and Samuelsen, 2007b, p. 482). To this aim, "the content domain to be covered by a credentialing test should be defined clearly, and it should be justified in terms of the importance of the content for credential-worthy performance in an occupation or profession ... Some form of job or practice analysis provides the primary basis for defining the content domain … [T]he emphasis for licensure is limited appropriately to the knowledge and skills necessary for effective practice" (AERA, 1999, p. 161).

Many studies have used 'content specialists' to evaluate the instructional or content domain representation of a test or assessment (Dolmans, Gijselaers, & Schmidt, 1992; Sireci & Geisinger, 1995). 'Content specialists' are persons with in-depth knowledge of the subject-matter who are willing to review items to ensure that each item represents the content and level of cognitive behavior desired (Haladyna, 2004). However, Polit and Hungler (1999) believe that a complete objective method of establishing content validity does not exist. Therefore, content validity is highly dependent on judgments (Polit & Hungler, 1999), and one practical way of

assuring content validity is subject-matter experts' content review (Beanland & Schneider, 1999; Haladyna, 2004; Polit & Hungler, 1999).

Taking insights from these evidential claims for the necessity of examining the content of credentialing tests/measures into the present developed schemes, the study aimed to take advantages of the content specialists' judgments to determine the content of the newly developed tests or the level of representativeness of their items in relation to the previously specified contents. That is, in addition to content relevance that was checked through specification of the behavioral domain (i.e., content relevance), the extent to which the test tasks and items represent the behavioral domain was also carefully examined (i.e., content coverage). Five subject-matter experts agreed to participate in the whole content validity stage. These content specialists were presented with the finalized and approved items and the tables of specification for each of the test subject/course. They were asked to rate the items based on a 5-point Likert-scale item, ranging from 'highly representative=6' to 'not representative at all=1', while matching with the previously-verified tables of specification. Item ratings were analyzed using a descriptive statistics procedure. A mean number was calculated for raters' responses to any of the items for each of the courses.

An overall mean was calculated for the raters' ratings for item representativeness of each test. 'Reading' appeared to have the lowest mean among all the courses (x=3.89). 'Grammar' and 'Vocabulary' were the next subjects with the lowest means (3.90 and 3.91, for Grammar and Vocabulary, respectively). Although these general proficiency courses/subjects, as a whole category, received the lowest ratings, still the means are high enough (x>3.5) to be regarded as content-representative. One possible explanation for these lower means might be 'a less detailed specification' of the content to be taught and evaluated determined by the ELTCD.

SLA was rated as the most highly content-representative item, showing the largest mean among all subjects (x=4.35). 'Language Testing', with an overall x of 4.33, appeared as the second highly representative subject. Similar ratings were found for 'Research Methodology' (x=4.29), 'Teaching methodology' (x=4.27), 'Linguistics' (x=4.27), and 'Phonology' (x=4.26), all indicating the content-representativeness of these decided upon test measures.

The results, in general, suggest that the present content specialists (i.e., 5 subject-matter experts) have found the courses sufficiently representative of the content presented by their respective tables of specification. In other words, from the perspectives of the present experts the test subject are perceived to measure what they are purported to measure.

4.2.4. Item Characteristics and Key Check as Evidence for Construct validity

As another distinct sources of validity (AERA, 1999), 'construct validity' is the extent to which a test measures a theoretical attribute (Beanland & Schneider, 1999; Polit & Hungler, 1999). In Bachman and Palmer's (1999) terms, test constructs can be viewed as definitions of abilities that help us state specific hypothesis about the relationship between these abilities and observed behaviors. Thus, in conducting construct validation, testing the hypothesized relationships between test scores and abilities is what is needed. In other words, test scores can be viewed as behavioral manifestations of test constructs. The approach that has been used most extensively in construct validation studies includes a number of statistical procedures (e.g., Correlational coefficients, discriminant analysis, etc.) to examine the hypothesized relationships between tests

scores and the constructs (Bachman & Palmer, 1999, 2010). One application of these statistical procedures is to investigate the relationships between item characteristics and item performance to provide sources of evidence in construct validity (Ibid). In this regards, Haladyna, 2004, Masters et al., 2001 and Violato, 1991 suggested that the construct validity of MC tests should be established using item response analysis such as item difficulty and item discrimination and distractor evaluation, as well as item key check.

Taken this perspective, the present study focused on item characteristics and item keys and the extent to which the items measure the domain of knowledge being examined.

4.2.4.1. Item Response Analysis (Item Difficulty and Item Discrimination)

Two types of item response analyses were run to examine item properties and detect flawed items. 'Item Discrimination' (ID) is used to diagnose item construction flaws which may have resulted in poor discrimination between high and low ability test takers. Item difficulty (shown as P) procedure is useful for identifying 'difficult' items. Although item analyses are not synonymous with test validity and the results are used for evaluating 'individual' items' quality, they present whether the items are tapping into the underlying construct, or being interpreted in a way other than the test construct intends to measure.

All items of the newly developed written tests were thus analyzed using 'Item Response Analysis' procedure. The item discrimination index measures the differences between the percentages of students in the upper group with that of the lower group who obtained the correct responses (Sim & Rasiah, 2006). At first, the total number of student-teachers in the upper 25% who obtained the correct response (US) and the lower 25% who obtained the correct response (LS) was counted. The higher the discrimination index, the test item can discriminate better between students with higher test scores and those with lower test scores. Items with discrimination index between 0.25-0.39 are considered 'acceptable', and items with discrimination index larger than 0.4 are considered 'excellent'.

In the MC written test developed in this study, 14 items (13%) had values smaller than .24, and most of the items (87%) had ID ranging between acceptable values of .25-1.00. That is, the overall written can be considered as an appropriate measure when discrimination power is concerned.

Only phonology test included 3 items (out of 12) of small ID values, indicating 25% of the test did not satisfactorily differentiate between good and weak students. Other subjects included one or two items to be revised due to their low ID index. Taken Brown (1983) and Crocker and Algina's (2008) recommended values of >.2, it appeared that only 10 (9%) of the test items showed low ID index. Thus, it can be claimed that most of the MC items used in these study were good or satisfactory items which would not need any modification or editing. 59 out of 108 items showed discrimination index equal or higher than 0.4, indicating that these MC items were excellent test items for discriminating between poor and good test-takers.

Regarding another type of Item Response analysis, i.e., item difficulty, Sim and Rasiah (2006) explain an item is considered 'difficult' when the difficulty index value is less than .37 and an item is considered 'easy' when its difficulty index value is greater than .80. Overall, only 6 items (5.5%) had values below the recommend value for very difficult item (Grammar: items

5, 10; SLA: items 10, 11; Teaching Methodology: item 4; and Research: item 4) and no item was found as 'the easiest' one.

To explore if the level of item difficulty in the present data correlate with the level of item discrimination, Pearson Correlation was conducted between the two variables (see Table 8).

Table 8. Pearson correlation between difficulty and discrimination levels

		Item discrimination	Item difficulty
Item discrimination	Pearson Correlation	1	-.217*
	Sig. (2-tailed)		.024
	N	108	108
Item difficulty	Pearson Correlation	-.217*	1
	Sig. (2-tailed)	.024	
	N	108	108
*. Correlation is significant at the 0.05 level (2-tailed).			

Pearson correlation value showed that discrimination index correlate poorly with difficulty index (r= - .217). The correlation is significant at 0.05 level (2-tailed). Negative correlation signifies that with increasing difficulty level, there is a decrease in discrimination level. In other words, as the items get easy, the level of discrimination index decreases consistently. Almost 95% of the items were found to have 'optimum difficulty level' (0.50) that, in turn, leads to maximum discrimination between high and low achievers. Since there is no easy item in the whole written test, the test displays significant discrimination powers.

4.2.4.2. Answer Key Check

In addition to ID and P analyses, item Answer keys were also carefully checked. According to Haladyna (2004), 'key check procedure' helps determine whether the correct answer to an MC item is truly correct and ensure that the correct answer is the only correct answer and there is no other correct answer to the item. He proposes that key check must be conducted by a number of subject-matter experts. In the present study, key check analysis was done with the help of the FU' subject-matter experts who made sure that there is only one actual correct answer to each of the MC items. Even though very minor inconsistencies were observed for few items, they were resolved following further discussions and illuminations. Using item key check and the resulting degrees of consistencies yield an informative picture of the construct under measurement. The

overall results of Item Response Analysis and Answer Key Check provided evidence required for judging the quality of the newly developed written assessment measures.

4.3. Student-Teachers' Competency Level

The newly developed tests were initially administered among the presents sample of English student-teachers (N=62) who were studying at FU during the educational year of 2016-2017 and spending their last two semesters. In addition to obtaining evidence for the validity of the tests, as presented above (4.2), such an initial piloting of the tests helped assess the knowledge level of the present student-teachers through their test performance.

Meanwhile, the reliability index was estimated using the KR-21 method and the overall test emerged as a satisfactory reliable measure (.89).

4.3.1. Descriptive Statistics: Overall Performance

Using SPSS-24, the test performance data were checked for descriptive statistics to obtain Mean, Standard Deviation, Mode, Median, and Range of scores. Additionally, Skewness and Kurtosis values for scores distribution were also obtained. Table 9 reports the summary statistics for the present teacher-educators' sample.

Table 9. Descriptive statistics: Student-teachers' performance statistics

Statistics	No. of Items	Minimum	Maximum	Mean	St. Deviation	Mode	Median	Skewness	Kurtosis
Results	108	19	78	48.50	15.34	42	47	.336	-.498

As Table 9 indicates, the total mean score was 9 points lower than the set cut score (x=48.5, <58), showing that there is a comparatively high proportion of minimally-competent testees in the present sample (x<58). The scores ranged from 19 to 78, with the most frequent score of 42. The median was 47, a value close to the obtained mean score, representing slight difference between these two measures of central tendency. The Skewness and Kurtosis values were between the suggested levels of ±½ by Bulmer (1979), ranging from -.498 to .336 for the variables, thus indicating no evidence of overly peaked variables. The obtained median value means that 50% of test-takers scored below 47 which, in turn, indicates that the 'low-to-near competency level (i.e., minimally-competent)' overrides 'complete competency level' in the present sample. In other words, only 25 percent of student-teachers got total scores at or above the cutting score and no one was found to be 'very competent (i.e., full mastery level)' at time of graduation.

5. Discussions and Conclusion

The question of teacher evaluation has been emphasized by scholars who posit teacher quality as an important key to student higher academic attainment and quality learning. Research, in

particular, by Kirkpatrick and Kirkpatrick (2006), Kaplan and Owings, 2002, Goldhaber and Antony (2007) and Mitchell et al. (2001) were among the ones that discussed the quality of learning and instruction in terms of the quality of teachers. They acknowledged 'teacher evaluation' as an important process to obtain information not only about teacher effectiveness and readiness to teach but also about the effectiveness and efficiency of teacher education programs. To guarantee teacher effectiveness, various tools and methods have been devised, adopted and implemented over the past three decades, ranging from observation, portfolios, attributes checklists, students' and peers' rating surveys, etc., to performance assessment, licensure testing and certification. As one of the most widely used measures of teacher evaluation, 'licensure tests' attempt to assess teacher knowledge-base and professional skills to ensure that teachers are qualified to teach (Mitchell, et al., 2001).

Tests for such purposes are grounded in 'common teaching standards', linked to 'local standards set by local policies' and to 'policy proposals for teacher assessment' (Darling-Hammond, 2010), at the same time that they must conform to the 'test designing standards' (AERA, APA & NCME, 1999, 2014; Tannebaum, 2011). For instance, in the United States 'Teacher Performance Assessment Consortium' under the auspices of the American Association of College of Teacher Education (AACTE) and the Council of Chief State School Officers (CCSSO) joined together to create a common initial licensing assessment that can be used nationwide for licensing teacher effectiveness. This also indicates that the development process for such standardized high-stakes tests usually involves a board of representative educators (Zieky et al., 2008) or advisory panel of teacher educators and teachers. These experts help determine and define content areas that should be covered on the tests, create specifications to guide the development effort, make passing-score recommendations, and review, revise and approve all test tasks, questions and items.

Being grounded in such current issues on licensure test development as well as taking insights from research on teacher evaluation, the present study took initiatives in developing an evaluation scheme for FU's EFL student-teachers'. For such an important purpose, the Policy Committee from Evaluation and Quality Assurance Department of FU, an advisory panel of FU's teacher educators, subject-matter professors, three test development specialists and practicing teachers and students were involved in the multistage development process of 'written tests' as part of the bigger PABT/ASLAH project. At different main steps, the study attempted to take advantage of the participants' expertise, experience, views and perspectives. This corroborates Shohamy's (2001) view that a social dialogue is needed for development and validation of tests as social activities. According to her, this can be achieved through involving a variety of stakeholders such as policymakers, test-designers, teachers, students and parents in the whole process. Hence, when talking about the teacher certification and credential issues there is a need to first encompass different perspectives, ideologies or values of the program planners, teacher educators, teachers and test designers. Then, extensive surveys and investigations from test recipients' perspectives should be followed to confirm test validity, fairness and practicality (Darling-Hammond, 2010; McNamara & Roever, 2006).

Driven by a recent teacher certification policy by FU, the present study got advantages of the information obtained through the abovementioned sources to embark on planning, designing, and validating the written assessment framework that helps determine EFL student-teachers' competencies they develop over their four-year training program. The recent certification policy

requires written tests to be taken by all EFL student-teachers seeking a teaching certification. Decisions are made based on the candidates' performance on the written tests that together with three other criteria of comprehensive ASLAH (i.e., performance assessment, portfolio and GPA) forms one of the final hurdles in their student career. The present certification process extends those arguments by Kane (2001) in that setting a licensure standard is, in effect, setting a policy about teacher requirements including the type and amount of knowledge and skills that beginning teachers need to have.

As to the specific components/sections of the tests, a synthesis of the information obtained through the following multiple sources led to the decisions about competency type, test subjects, and test format: content analyses of the national policy documents of teacher assessment and FU's ELT curriculum, interviews with FU teacher educators and subject-matter experts, teachers and student-teachers' surveys, and insights from theoretical and empirical research findings on teacher evaluation, teacher quality and effectiveness as mentioned above. On such basis, the proposed written assessment scheme included nine major content knowledge (CK) subject-matter courses that a newly licensed/certified EFL teacher should know in order to perform his or her teaching job competently. Before the main piloting of the tests, their perceived usefulness and effectiveness were explored in terms of 'consequences on instructional objectives', 'quality of results' and 'fairness' from the perspectives of EFL teachers and student-teachers. The obtained evidence showed 'moderate' degrees of consistency between the program's idealized outcomes and consequences and the ones teachers perceived. They did not report negative perspectives, however. On the other hand, teacher educators and subject-matter experts conceptualized the licensure tests as leading to 'teacher quality improvement, knowledge development, and preparedness' and to 'improving test development technical issues'. Furthermore, when 'test practicality' was concerned, the majority of participants believed in the practicality of the new tests in the FU context.

Afterward, based on the results of determining subject matter courses and qualifying scores, an advisory panel of subject-matter experts defined the specifications of each of the nine courses to guide the subsequent phase of test development efforts which involved test development specialists too. Once developed, initially reviewed, modified, and piloted, the tests were explored for 'test validity and reliability' and analyzed for 'item and test quality'. The results of 'content' and 'expert' validity as well as 'item statistics' showed that the designed tests can be used as appropriate measures to assess EFL teachers' competencies.

In terms of the main subtests, the tests do not exactly correspond with the main subtests of the widely known tests such as PRAXISs. An obvious reason is that the present written assessment framework has been specifically developed for assessing Iranian EFL teachers of the nationwide teacher training faculties. In other words, the present tests, like many other certification tests in the world, are not only context-specific but domain-specific; therefore, do not, by nature, measure mathematics, science or social science knowledge that are assessed through PRAXISs. However, like the PRAXIS exams, they include separate sections/subtests that assess subject-specific teaching skills and knowledge. The present tests also follow Schulman's (1987) suggestions in that they include both general and specialized CK courses suggested by these scholars. The newly developed written tests were therefore used as a main tool to assess the extent to which the student-teacher participants of the study meet the qualification requirements for English language teaching.

Overall, 'teacher certification' signifies the program effectiveness question which is at the core conception of evaluation. To learn about the factors and events which shape effectiveness of a new policy (such as FU's ASLAH policy and certification), evaluation should not focus only on the input and the ultimate output but on the nature of the output and factors which shape it (e.g., teachers' competencies as an output of teacher training programs but as an input to teachers' future teaching career). Such an evaluation lens involves engagement of a range of stakeholders in a wider sphere of social dialogue that helps obtain information and furthers our understanding of the program's missions, objectives, input and the intended output (Shayestefar, 2013). Without making such an evaluation endeavor, there is no way to provide a comprehensive solution that systematically and progressively gauge how knowledgeable and skillful the EFL student-teachers are and whether they are on track for becoming ready for future teaching in social-cultural contexts of schools. Equipped with this information, FU's training programs and trainers can support those non-certified individuals who need extra support and improvement.

References

Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of educational objectives: Complete edition, New York: Longman.

Al-Shabatat, M. A., Abbas, M., & Ismail, H. N. (2010). The Direct and Indirect Effects of the Achievement Motivation on Nurturing Intellectual Giftedness. *International Journal of Human and Social Sciences*, *5*(9), 580–588.

Arbuckle, J.L. (2009). *IBM SPSS Amos 18 User's Guide*. Chicago, IL: IBM.

Arbuckle, J. L. (2013). *IBM SPSS Amos 22 user's guide*. Chicago, IL: IBM.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (2014). *Standards for educational and psychological Testing*. Washington, DC: American Educational Research Association.

Bachman and Palmer's (1999). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F. & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, a., … Tsai, Y.-M. (2010). Teachers' Mathematical Knowledge, Cognitive Activation in the Classroom, and Student Progress. *American Educational Research Journal*, *47*(1), 133–180.

https://doi.org/10.3102/0002831209345157.

Beanland, C., & Schneider, Z. (1999). *Nursing Research: methods, critical appraisal and utilisation*. (Mosby, Ed.). Sydney.

Boyd, D., Goldhaber, D., Lankford, H., & Wyckoff, J. (2007). The Effect of Certification and Preparation on Teacher Quality. *The Future of Children*, *17*(1), 45–68. https://doi.org/10.1353/foc.2007.0000.

Brown, F. (1983). *Principles of educational and psychological testing* (3rd ed.). New York: Holt: Rinehart and Winston.

Bulmer, M. G. (1979). Principles of statistics. New York: Dover.

Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows*. Thousand Oaks, CA: Sage Publications.

Carey, K. (2004). The real value of teachers. Using New Information about Teacher Effectiveness to Close the Achievement Gap. *Thinking K-16*, *8*(1), 1–44.

Crocker, L., & Algina, J. (2008). *Introduction to Classical and Modern Test Theory*. *Cengage Learning*. USA, Mason, Ohio: Cengage Learning.

Darling-Hammond, L. (1997). *Doing What Matters Most: Investing in Quality Teaching* (First Edit). United States of America: Prepared for the National Commission on Teaching and America's Future. https://doi.org/ED419696.

Darling-Hammond, L. (2000). Teacher Quality and Student Achievement : A Review of State Policy Evidence Previous Research. *Education*, *8*(1), 1–44. https://doi. org/10.1038/sj.clp.

Darling-Hammond, L., & Ball, D. L. (2004). Teaching for high standards: what policymakers need to know and be able to do.pdf, (November 1998), 1–33.

Dolmans, D. H. J. M., Gijselaers, W. H., & Schmidt, H. G. (1992). Do Students Learn What Their Teachers Intend They Learn? Guiding Processes in Problem Based Learning. San Francisco, California: Paper presented at the Annual Meeting of the American Educational Research Association.

Downing, S. M. (2006). Selected-Response Item Formats in Test Development. In T. M. H. Steven M. Downing (Ed.), *Handbook of test development* (pp. 287–301). Lawrence Erlbaum Associates.

English Language Teaching Curriculm Document of Farhanian University (2014). Farhangian University Education Department, Iran.

ETS (2010). *The Praxis Series Passing Scores State by State*.

Freeman, D., & Johnson, K. E. (1998). Education, Reconceptualizing the Knowledge-Base of Language Teacher. *Annual Review of Applied Linguistics*, *32*(3), 397–417.

https://doi.org/10.1017/S0267190500200032.

Ghassemi, V. (2010). *Structural Equation Modeling in Social Science Research Using AMOS Graphics*. Tehran: AGHAH Publication.

Glesne, C. (1999). *Becoming qualitative researchers: An introduction* (2nd ed.). Longman.

Goldhaber, D. (2002). The mystery of good teaching: Surveying the evidence on student achievement and teachers' characteristics. *Education Next*, *2*(21), 50–55.

Goldhaber, D., & Anthony, E. (2004). Can Teacher Quality Be Effectively Assessed? *Quality Assurance*, *89*(1), 134–150. https://doi.org/10.1162/rest.89.1.134.

Goldhaber, D., & Anthony, E. (2007). Can Teacher Quality Be Effectively Assessed? National Board Certification as a Signal of Effective Teaching. *Review of Economics and Statistics*, *89*(1), 134–150. https://doi.org/10.1162/rest.89.1.134.

Goldhaber, D. D., & Brewer, D. J. (2000). Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement. *Educational Evaluation and Policy Analysis*, *22*(2), 129–145. https://doi.org/ 10.3102 /01623737022002129.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Hanushek, E. A. (1997). Assesing the Effects of School resources on Student Performance: An Update. *Handbook of the Economics of Education*. https://doi.org/10.1016/B978-0-444-53444-6.00004-3.

Hanushek, E. A., & Rivkin, S. G. (Steven G. (2004). How to Improve the Supply of High-Quality Teachers. *Brookings Papers on Education Policy*, (1), 7–44. https://doi.org/ 10.1353/pep .2004.0001.

Heck, R. H. (2007). Examining the Relationship Between Teacher Quality as an Organizational Property of Schools and Students' Achievement and Growth Rates. *Educational Administration Quarterly*, *43*(4), 399–432. https://doi.org/ 10.1177/0013161X07306452.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematical knowledge for teaching. *Elementary School Journal*, *105*, 11–30. https://doi.org/ 10.1086 /428763.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.

Kane, M. (2006). Content-Related Validity Evidence in Test Development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 131–153). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Kaplan, L. S., & Owings, W. A. (2002). *Enhancing Teaching Quality*. Bloomington, Indiana: Phi Delta Kappa Educational Foundation.

Kiany, G. H., ShayesteFar, P. & Ahmadishokooh, A. (2016). Teacher Educators' Evaluation Model. *American Journal of Educational Research, 4*(2), 210-220.

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating Training Programs*. San Francisco, California: Berrett-Koehler Publishers, Inc.

Krauss, S., Brunner, M., Kunter, M., Baumert, J., Blum, W., Neubrand, M., & Jordan, A. (2008). Pedagogical content knowledge and content knowledge of secondary mathematics teachers. *Journal of Educational Psychology*, *100*(3), 716–725. https://doi.org/10.1037/0022-0663.100.3.716.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ.

Lissitz, R. W., & Samuelsen, K. (2007b). A Suggested Change in Terminology and Emphasis Regarding Validity and Education. *Educational Researcher*, *36*(8), 437–448. https://doi.org/10.3102/0013189X07311286.

Mangiante, E. M. S. (2011). Teachers matter: Measures of teacher effectiveness in low-income minority schools. *Educational Assessment, Evaluation and Accountability*, *23*(1), 41–63. https://doi.org/10.1007/s11092-010-9107-x.

Marsh, H. W., & Hattie, J. (2002). The Relation Between Research Productivity and Teaching Effectiveness: Complementary, Antagonistic, or Independent Constructs? *The Journal of Higher Education*, *73*(5), 603–641. https://doi.org /10.1353/jhe.2002.0047.

Masters, J. C., Hulsmeyer, B., Pike, M., Leichty, K., Miller, M., & Verst, A. (2001). Assessment of Multiple-Choice Questions in Selected Test Banks Accompanying Text Books Used in Nursing Education. *J Nurs Educ.*, *40*(1), 25–32.

Mc Caffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating Models for Teacher Accountability*. *Distribution* (Vol. 158). Santa Monica, CA: RAND Corporation.

McNamara, T. & Roever, C. (2006). *Language Testing: The social dimension.* Oxford: Blackwell Publishing.

Mehrens, W. A., & Lehmann, I. J. (1991). *MEASUREMENT AND EVALUATION in Education and Psychology* (4th ed.). Belmont, CA: Wadsworth/Thomson Learning.

Mellone, I., & Fabern, C. (2014). Are They Mission Ready? Using the Modified Angoff Method To Set Cut Scores. *Proceedings of Interservice/Industry Training, Simulation, and Education Conferen*ce (I/ITSEC), P.1-11, N. 14060.

Mitchell, K. J., Robinson, D. Z., Plake, B. S., & Knowles, K. T. (2001). *Testing Teacher Candidates: The Role of Licensure Tests in Improving Teacher Quality*. *Committee on Assessment and*

Teacher Quality, Mitchell, K.J., Robinson, D.Z., Plake, B.S., and Knowles, K.T., editors. Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. WaSh: National Academy Press. https://doi.org/10.17226/10090.

Mitchell, R., & Barth, P. (1999). EXAMINATIONS THINKING K-16. *Thinking K-16*, *3*(1).

Moats, L. C., & Foorman, B. R. (2003). Measuring teachers' content knowledge of language and reading. *Annals of Dyslexia*, *53*(1), 23–45. https://doi.org/ 10.1007/s11881-003-0003-7.

Monyatsi, P., Steyn, T., & Kamper, G. (2006). Teacher perceptions of the effectiveness of teacher appraisal in Botswana. *South African Journal of Education*, *26*(3), 427–441.

Mulaik, S. a., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, *105*(3), 430–445. https://doi.org/10.1037/0033-2909.105.3.430.

Navidinia, H., Kiani, G. R., Akbari, R., & Samar, R. G. (2015). identifying the requirements and components of a model for English language teachers' appraisal in Iranian high schools. *Language Related Research*, *6*(2).

Neild, R. C., & Farley-Ripple, E. (2008). Within-School Variation in Teacher Quality: The Case of Ninth Grade. *American Journal of Education*, *114*(3), 271–305. https://doi.org/10.1086/529503.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How Large Are Teacher Effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237–257. https://doi.org/10.3102 /01623737026003 237.

Pallant, J. (2013). *A step by step guid to data analysis using IBM SPSS* (5th ed.). New York: McGraw Hill.

Polit, D., & Hungler, B. (1999). *Nursing research: Principles and Methods*. Philadelphia: Lippincott Williams & Wilkins.

Raymond, M. R., & Neuste, S. (2006). Determining the Content of Credentialing Examinations. In T. M. H. Steven M. Downing (Ed.), *Handbook of test development* (pp. 181–223). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Reeves, D. (2007). New ways to hire educators. *Educational Leadership*, *64*(8), 83–84.

Ribas, W. B. (2005). *Teacher Evaluation that Works!!: The Educational, Legal, Public Relations [political] & Social-emotional [E.L.P.S.] Standards & Processes of Effective Supervision & Evaluation*. (R. Education, Ed.) (illustrate).

Rivkin, S. G., Hanushek, E. A., & Kian, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, *73*(2), 417–458. https://doi.org/10.1002/polq.12145.

Ross, A., & Hutchings, M. (2003). *Attracting, Developing and Retaining Effective Teachers in the United Kingdom of Great Britain and Northern Ireland*. *Institute for Policy Studies in*

Education. London Metropolitan University.

Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N., & Miller, J. L. (2013). The influence of teachers' knowledge on student learning in middle school physical science classrooms. *American Educational Research Journal*, *50*(5), 1020–1049. https://doi.org/10.3102/0002831213477680.

Schmidt, W. H., Tatto, M. T., Bankov, K., Blömeke, S., Cedillo, T., Cogan, L., Schwille, J. (2007). *The preparation gap: Teacher education for middle school mathematics in six countries*. East Lansing, MI: Michigan State University.

Shayeste-Far, P. (2013). *Evaluation of the changing high-stakes university entrance assessment of english language: Learning from multiple approaches*. Unpublished Ph.D. dissertation. Tarbiat Modares University.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing, 18*(4), 373-393.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*, 1–22.

Sim, S. M., & Rasiah, R. I. (2006). Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. *Annals of the Academy of Medicine Singapore*, *35*(2), 67–71.

Sireci, S. G., & Geisinger, K. F. (1995). Using Subject-Matter Experts to Assess Content Representation: An MDS Analysis. *Applied Psychological Measurement*, *19*(3), 241–255. https://doi.org/10.1177/014662169501900303.

Stapleton, C. D. (1997). Basic concepts in exploratory factor analysis (EFA) as a tool to evaluate score validity: A right-brained approach. *Paper Presented at the Annual Meeting of the Southeast Educational Research Association*, 1–19.

Stronge, J. H. (2006). *Evaluating teaching. A guide to current thinking and best practice*. Thousand Oaks: Crowin Press.

Stronge, J. H., & Hindman, J. L. (2003). Hiring the best teachers. *Educational Leadership*, *60*(8), 48-52.

Stronge, J. H., & Tucker, P. D. (2003). *Handbook on teacher evaluation: Assessing and improving performance*. Larchmont, NY: Eye On Education.

The Document of Fundamental Reforms in Education (2012). The Supreme Council of Cultural Revultion, Iran.

The Document of National Curriculum (2010). Ministry of Education of the Islamic Republic of Iran.

Tannebaum, R. (2011). *Setting standards on the PRAXISS series tests.* R & D Connections, ETS, N. 17.

Violato, C. (1991). Item difficulty and discrimination as a function of stem completeness. *Psychological Reports*, *69* (3 Pt 1), 739–43. https://doi.org/ 10.2466/pr0.1991.69.3.739.

Appendix

Dear Participants,

To fulfill the graduation and certification requirements for its teacher candidates, Iran's Farhangian University (Teacher Training University) adopted 'a comprehensive exam' policy (locally called ASLAH) for a teacher evaluation purpose. According to this policy, four criteria are required to certify teacher candidates' qualifications for teaching: performance assessment (25%), written assessment (25%), portfolio (25%) and GPA (25% of the total score).

The purpose of this survey is to explore a) perceived fairness, consequences, and the quality of the results of the new certification policy and practice, from your eyes (items 1-18 below); and b) priority of the exam subject-matter courses on the basis of the importance and weights you think each course would have, and the item format of each (items 19-25).

Please choose the alternative which seems appropriate.

All information is confidential.

Thank you for your participation.

- **Section A:** Please choose the response alternative that reflects your perspective of the new programme's consequences, fairness, and quality of results.

	Statements	4 Strongly Agree	3 Agree	2 Disagree	1 Strongly Disagree	No Idea
1	The tests will not drive student-teachers to a *rote-learning* form of learning.					
2	The tests will increase student-teachers' *sense of problem solving* that is required for EFL teaching.					
3	The tests will influence student-teachers' *quality of learning*.					
4	The tests will increase student-teachers' *motivation for learning*.					
5	The tests will reduce student-teachers' *learning stress and anxiety* level.					
6	The tests will have an important role in helping student-teachers *develop competencies required for EFL teaching.*					
7	The tests will increase teacher educators' *quality of teaching.*					

8	The tests will lead to transferring more knowledgeable EFL teachers to the Ministry of Education.					
9	The tests will be administered under equal conditions to all student-teachers (time, procedures, scoring).					
10	The tests will assess the *real abilities and competencies.*					
11	The tests provide a *fair tool* for assessing student-teachers' competencies.					
12	The tests can provide student-teachers with *equal opportunities* to access to the test content and test-preparation materials.					
13	*Fairness* with the present tests will be achieved more through provincial- level administration than state-level administration.					
14	The criteria and procedures for measuring and recording student-teachers' competencies and the relevant interpretations and decisions are clear to the student-teacher before the tests' administration.					
15	The test performance interpretations can provide decision makers with *sufficient information about the student-teachers' competency level* to make meaningful and fair decisions about them.					
16	The test scores and the relevant decisions are timely announced.					
17	The decisions that are made about student-teachers' eligibility to enter the teaching profession are useful for each stakeholder.					
18	The test scores and the relevant interpretations and decision are announced confidentially to each teacher candidate.					

- **Section B: Content and Structure of the Tests: Subject-matter course priority**

19. To what extent do you agree with the inclusion of content knowledge (CK) in the written test scheme?
(Strongly agree-agree-disagree-strongly disagree-no idea)

20. To what extent do you agree with the inclusion of English general knowledge in the written test scheme?
(Strongly agree-agree-disagree-strongly disagree-no idea)

21. How important is each of the following subject-matter courses for measuring CK knowledge? How many items do you suggest for each curse?

Suggested Subject-matter Course (100 items)		Highly important 3	Moderately important 2	Low importance 1	Your suggested number of items	Your suggested course/s
General Proficiency Knowledge Courses	Reading Comprehension					
	Writing					
	Grammar					
	Vocabulary & Idioms					
Content	SLA					

Knowledge (CK) Courses	Teaching Methodologies					
	Testing					
	Research Methods					
	Phonology					
	Linguistics					

- 22. How important is each of the following subject-matter courses for measuring the Pedagogical Content Knowledge (PCK)? How many items do you suggest for each course?

Suggested Subject-matter Course (50 items)		Highly important 3	Moderately important 2	Low importance 1	Your suggested number of items	Your suggested course/s
Pedagogical Content Knowledge (PCK) Courses	Teaching Philosophy in ELT					
	Teaching Strategies in ELT					
	Lesson Planning in ELT					
	Educational Testing in ELT					

23. To what extent do you agree with the inclusion of short-answer items in the written test scheme?
(Strongly agree-agree-disagree-strongly disagree-no idea)

24. To what extent do you agree with the inclusion of Multiple-choice items in the written test scheme?
)Strongly agree-agree-disagree-strongly disagree-no idea)

25. Given the present condition of Farhangian University, to what extent do you think the present written assessment test program is practical?
(Very practical-practical-partially practical- impractical)

- **Section C: Personal information**

26. Please choose the appropriate option or write the relevant information as required.
Your Gender*: female*___ *male*____ Total years of teaching experience: _____*years*
Your Academic degree(s): *BA*___ *MA*____ *PhD*___ *other*____
Further suggestions:
_____
_____