ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 6, No. 5, pp. 1111-1120, May 2016
DOI: http://dx.doi.org/10.17507/tpls.0605.27

Ph. D. Instructors' and Students' Insights into the Validity of the New Iranian TEFL Ph. D. Program Entrance Exam

Reza Rezvani
English Language Department, Yasouj University, Yasouj, Iran

Ali Sayyadi
English Language Department, Yasouj University, Yasouj, Iran

***Abstract*—Owing to their scope, and decisiveness, Ph. D. program entrance exams (PPEE) ought to demonstrate acceptable reliability and validity. The current study aims to examine the reliability and validity of the new Teaching English as a Foreign Language (TEFL) PPEE from the perspective of both university professors and Ph. D. students. To this end, in-depth unstructured interviews were conducted with ten experienced TEFL university professors from four different Iranian state universities along with ten Ph. D. students who sat both the new and old PPEEs. A detailed content analysis of the data suggested that the new exam was assumed to establish acceptable reliability through standardization and consistency in administration and scoring procedures. Conversely, the new exam was perceived to demonstrate defective face, content, predictive, and construct validities. This study further discusses the significance and implications of the findings in the context of Iranian TEFL higher education.**

***Index Terms*—reliability, validity, TEFL Ph. D. programs, University entrance exams, instructors, Ph. D. students**

I. Introduction

Social and educational accomplishments have been firmly tied up with obtaining Ph. D. degree in Iranian context due to the fact that it paves the way for procuring the highest educational degree and consequently reputable jobs. Records of more than 216000 and 240000 applicants sitting the Ph. D. program entrance exam (PPEE) for state universities in 2013 and 2014 respectively (Sanjesh, 2014) are indicative of substantially increasing number of applicants aspiring to qualify for such decisive programs. Despite the annually increasing number of Iranian universities offering Ph. D. programs, administrative limitations are still prevalent in Iranian context. The imbalance between the number of the PPEE applicants and matriculated Ph.D. students, accordingly, has highlighted the sensitivity and significance of such a high-stakes exam and has aroused mounting concerns about it among the applicants and other stakeholders.

Planning and administering Iranian PPEE involved critical modifications in 2012 when the Iranian Ministry of Science, Research, and Technology (IMSRT) resolved to launch a new PPEE in an attempt to pursue educational fairness, to reduce extravagant costs for setting the PPEE, and to admit highly qualified Ph. D. students from all around the country (ISNA, 2012). To this end, universities were deprived from their monopoly on the development and administration of PPEEs, and Sanjesh Organization, a subsidiary of IMSRT, was instead charged with planning, developing, and administering the new PPEEs in national scope. As far as the exam is concerned, two dramatic changes effected were replacing essay type items in the old PPEEs with most radically objective multiple choice ones along with including a number of questions measuring the logical and mathematical intelligences of the new PPEE's applicants. The incentives behind such amendments were to resort to uniform administrative and scoring procedures to achieve the 'educational fairness' through the provision of equal chance of admission.

Such a sensitive and high-stakes exam inarguably needs to demonstrate considerable levels of reliability and validity. From the classical true score (CTS) perspective, reliability is defined as the correlation between the observed scores on two distinct sets of measurement and is measured through internal consistency, stability, and equivalence approaches (Bachman, 1990). This stipulation indicates the empirical nature of reliability. A fundamental concern in investigating reliability, however, is to identify potential sources of error in a given measure and subsequently minimize the effects of such factors on that measure (Bachman, 1990). In CTS all factors other than the ability being measured are considered to be random sources of error, that is, it treats all measurement errors to be unsystematic and unpredictable factors to be minimized. As delineated in CTS, testees' performance on a test, varies as a function of the ability being measured and error including individual attributes of test takers and systematic test method facets. Test method facets are categorized into five groups by Bachman (1990) including the testing environment, the test rubric, the nature of the input the test taker receives, the nature of the expected response to that input, and the relationship between input and response. Despite the statistical nature of reliability, it "may best be addressed by considering a number of factors that may

contribute to the unreliability of a test" (Brown, 2004, p.20). In other words, identification and minimization of the potential impacts of systematic and unsystematic errors can boost the reliability of a given test (Bachman, 1990; Henning, 1987; Neiman, 2011).

Validity has been conventionally classified into content, criterion, and construct validity. Measurement experts, nevertheless, have come to view these as complementary types of evidence to be accumulated in the process of construct validation (see Bachman, 1990; Bachman & Palmer, 2010; Henning, 1987; Johnson, 2001; Messick, 1988). Instead of positing a sliced view of validity, it is defined as a unitary concept which concerns interpretation and use of the information gathered through the testing procedure (Bachman & Palmer, 2010; Messick 1998; Messick, 1992). Messick (1988) argues that viewing different approaches to validation as separate lines of evidence to support a given score interpretation is inadequate. The main processes of validation involve theoretical and operational definition of the constructs of concern, formulating hypothesis regarding the relationship between constructs and other factors interacting with them such as test method facets, and empirically verifying or falsifying the hypothesis through the accumulated correlational or experimental evidences (Bachman, 1990). As products of test scores, such quantitative approaches to construct validation, however, serve critical limitations in view of the fact that they provide no means to scrutinize the underlying processes of test taking (Cohen, 1984; Rezvani, 2010). Hence, language testing researchers have recently begun to take qualitative research into service in order to have more insightful understanding of what test takers actually do when they take tests and what actually tests measure, which has a great deal of potential for providing evidence for construct validation (Weir, 2005). In some cases it may be more appropriate to investigate the appropriacy and adequacy of a test content relevance and coverage in relation to intended course and performance through qualitative examination of experts' insights (Brown, 2004; Fulcher, 2010; Purpura, 1998; weir, 2005). Similar procedures are worthwhile to accumulate evidence to examine potential threats to the validity of test score interpretation and use (Messick, 1992).

Reliability and validity of a test are two most critical characteristics of any tests which are in direct line with the import of the decisions to be made based on the test results (Bachman, 1990; Chapelle & Brindley, 2002; Hamp-lyons & Lynch, 1998), that is, the higher the stakes of a test, the more the significance of validation. Accordingly, examining the reliability and validity of a critically sensitive and high-stakes exam like PPEE is of profound significance. Thus, the present study was motivated to explore the reliability and validity of the new TEFL PPEE from the viewpoints of Iranian Ph. D. students and university instructors.

II. Review of Literature

The terms reliability and validity have undergone major conceptual changes over the past decades. Given the fact that the goal of the present paper is to investigate the reliability and validity of the new TEFL PPEE from the viewpoints of Ph. D. students and university instructors, a brief review of these conceptual changes could be of noticeable import. Traditionally viewing, reliability refers to the consistency of measurement from one occasion to another. In objectively scored tests such as multiple choice tests, reliability is basically usually estimated by internal consistency which determines how well the items on a test correlate with each other, whereas in subjectively scored tests such as essays or oral proficiency interviews inter-rater and intra-rater reliability estimation methods are employed (Johnson, 2001). On the other hand, more recent views on the concept of reliability represent it as a broad model which is grounded on the basis of various factors affecting performance of individuals on a given test (Bachman, 1990).

In spite of an ongoing debate on how validity should be defined, one can identify two major periods in the literature concerning the concept of validity in language testing marked by the publication of Messick's (1989) seminal work on validity. These two periods can be labeled as the 'pre-Messick' and the 'post-Messick'. The pre-Messick definition of validity is primarily associated with different types of validity, such as content validity, criterion related validity, and construct validity. Messick's (1989) approach to defining validity, nevertheless, set forth an alternative delineation of validity where he asserts that "validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (Messick, 1989, P. 3). The key point of the post-Mesick conceptualization of validity can be captured in a unified but multifaceted concept (Johnson, 2001).

Test validation is immensely significant for all test users because "accepted practices of the validation are critical to decisions about what constitutes a good language test for a particular situation" (Chapelle, 1999, p.254). Accordingly, review of assessment literature is highlighted by countless studies on examining reliability and validity of numerous proficiency, aptitude, Knowledge, and placement tests (see for example, Carlson et al., 1985; Chi, 2011; Compton, 2011; Dandonolli & Henning 1990; Drollinger et al., 2006; Eda et al., 2008; Greenberg, 1986; Johnson, 2001; Magnan, 1987; Patterson & Ewing, 2013; Sabers & Feldt, 1968; Stansfield & Kenyon, 1992; Thompson 1995; Zhao, 2013).The sensitivity and significance of university entrance exams (UEE), especially in countries where UEEs are perceived as the sole gateways to qualify for university programs, have remarkably necessitated undertaking numerous in-depth inquiries on their reliability and validity around the world (see for example, Frain, 2009; Hissbach et al., 2011; Ito, 2012; Kirkpatrick & Hlaing, 2013). Kirkpatrick and Hlaing (2013), for instance, sought to examine the reliability and validity of the English section of the Myanmar UEE and came to the point that the exam suffered from poor construct and content validity leading to negative washback with regard to learning and teaching. Similarly, Ito (2012) conducted a

validation study on the English language test in the Japanese Nationwide University Entrance Examination and concluded that unlike other tests which enjoyed satisfactory validity, paper-pencil pronunciation subtest suffered from low validity with almost no significant contribution to the total test score. Frain (2009) also examined Korean first year university students before and after sitting the university entrance exams and came to the conclusion that the exam did not properly screen and predict the students communicative competence.

Examining UEEs has also been subject to discussion and research in Iranian higher education context (see e.g. Mahmoudi & Abu Bakr, 2013, Razavipur, 2014; Rezvani & Sayyadi, 2014, Salehi, 2012, Kazemi & Sayyadi, 2014). For instance, the washback effect of UUEs on applicants' motivation (e.g. Kazemi and Sayyadi, 2014), applicants' study plans and strategies (e.g. Rezvani & Sayyadi, 2014), applicants' quality of English learning (e.g. Salehi, 2012; Mahmoudi & Abu Bakr, 2013), teachers' pedagogical strategies (e.g. Salehi & Yunus, 2012; Ramezaney, 2014), and teachers' curricular planning (Ramezaney, 2014), among others, have been pursued recently. The washback effect of the new TEFL Ph. D. program entrance exam on the applicants' study plans and strategies was investigated in a recent comparative inquiry by Rezvani and Sayyadi (2014). That the applicants who sat the old PPEE were required to provide comprehensive and elaborate answers to essentially essay type questions, underscores the requirement to develop and maintain analytic, synthetic and evaluative qualities and capabilities on the part of the applicants. However, the introduction and dominance of multiple choice PPEEs has obviated the need to possess such capacities calling for applicants' lower cognitive abilities of comprehension and recall of information crammed, as argued by Rezvani and Sayyadi (2014).

To conclude, review of the related literature indicates that the reliability and validity of Iranian high-stakes exams have been under-researched. This is more acute when PPEE is a concern. Given its sensitivity, recency, and scope, the present study sought to examine the Iranian TEFL PPEE's reliability and validity from the standpoint of Ph. D. students and university instructors.

III. Objectives of the Study

It seems that no published study has been conducted in order to examine the reliability and validity of the new TEFL PPEE as a critically decisive gate-keeping exam. Thus, the present validation study seeks to examine these two crucial considerations validity from the viewpoints of the instructors and Ph.D. students.

IV. Method

The current study is a qualitative examination of the new Iranian TEFL PPEE's reliability and validity from the standpoint of Ph. D. students and university teachers.

A. Participants and Sampling Method

As argued by Guba and Lincoln (1981), "sampling [in qualitative research] is almost never representative or random but purposive, intended to exploit competing views and fresh perspectives as fully as possible" (p. 276). Accordingly, the current study employed a snowball sampling procedure, a variation on purposive sampling, where the initially selected participants suggested some further informants who could be appropriate for the intended sample.

The selected participants were ten experienced Iranian instructors and ten Iranian TEFL Ph. D. students. The first group of subjects included ten university instructors currently teaching at four different Iranian universities, that is, Shiraz, Esfahan, Sheikh-Bahaie, and Shahre-kord Universities. From among the instructors taking part in the study, six instructors had the experience of teaching both TEFL M. A. and Ph. D. courses and four instructors had taught only M. A. courses. The instructors had at least four years of teaching experience at universities and aged between 43 and 56. Of the instructors, three were females and seven were males. Table 1 summarizes the demographic information of the interviewed instructors.

TABLE 1:
DEMOGRAPHIC INFORMATION OF THE INTERVIEWED INSTRUCTORS

No	Name*	Age	Gender	M. A. teaching experience	Ph. D. teaching experience	Current university
1	Maryam	48	Female	6 years	4 years	Shiraz
2	Meysm	54	Male	13 years	10 years	Shiraz
3	Hamid	46	Male	7 years	5 years	Shiraz
4	Mansur	45	Male	7 years	4 years	Shiraz
5	Nader	44	Male	4 years	_	Shahre-Kord
6	Ahmad	43	Male	4 years	_	Shahre-Kord
7	Javid	50	Male	9 years	6 years	Esfahan
8	Samira	48	Female	7 years	5 years	Esfahan
9	Adel	44	Male	5 years	_	Sheikh-Bahaei
10	Simin	45	Female	5 years	_	Sheikh-Bahaei

Note: The names are fictitious.

A total of ten applicants, 6 females and 4 males, pointed to by the interviewed professors were accessed and interviewed. They were all Ph. D. students who sat both the new and old TEFL PPEE themselves. Of the students

interviewed, four were in their twenties and the rest were in their thirties. Table 2 illustrates the demographic information of the TEFL Ph. D. students taking part in the study.

TABLE 2:
DEMOGRAPHIC INFORMATION OF THE INTERVIEWED STUDENTS

No	Name*	Age	Gender	Ph. D. university
1	Mahmud	28	Male	Tabriz
2	Saeed	29	Male	Shiraz
3	Parvin	31	Female	Shiraz
4	Zahra	34	Female	Shiraz
5	Amin	32	Male	Shiraz
6	Narges	29	Female	Esfahan
7	Nahid	31	Female	Esfahan
8	Reza	28	Male	Esfahan
9	Razieh	35	Female	Tehran
10	Elham	30	Female	Chamran University of Ahvaz

Note: The names are fictitious.

B. Instrumentation

In line with the objectives of the study, semi-structured interviews were utilized to elicit the interviewees' views and reflections about the new TEFL PPEE. To assure the comprehensibility and quality of the interview questions, they were piloted with two instructors and students with comparable characteristics. They were not included in the main study.

C. Data Accumulation Procedure

The interview questions were developed and asked in English. The participants, however, had a choice of responding in Persian or English. Interviews were conducted individually by one of the researchers and took ten to twenty minutes. They were all recorded using an mp3 player with the permission of the interviewees. Once the data were accumulated, they were transcribed into written texts and then analyzed. To ensure the trustworthiness of the findings, intensive care was taken to avoid bias through employing a prolonged and persistent field-work and accounting for participants' language verbatim accounts meticulously documented as recommended by McMillan and Schumacher (2006). When the responses were in Persian, the statements were carefully rendered into English. In addition, the researchers frequently used member checking to check the data informally with the participants for accuracy during the interviews, and were sensitive to discrepant data that did not conform to the emerging patterns.

D. Data Analysis Procedure

Researchers conducting qualitative scrutiny on the data accumulated through interviews have widely advocated interpretation of the collected data thorough content analysis (Elo & Kyngas, 2008). Accordingly, constant comparative content analysis, as suggested by Glaser and Strauss (1967), was employed in the process of data analysis to code the transcribed interviews involving an inductive reasoning process of frequent sifting through the data to identify similarities and patterns of reference in the interview transcripts. Detailed analyses of the similarities and patterns subsequently gave rise to the emergence of an evolving coding system for the categories. The units of analysis and coding schemes, more specifically, were defined and developed during the process of the content analysis; then, the codes were transformed into categorical labels or themes that were repeated or appeared as patterns in the interviews. This iterative procedure, according to Patton (2002), is intended to help the researchers in "developing some manageable classification or coding scheme" as "the first step of analysis" (p. 463). Data analysis proceeded incrementally and once the coherence and saturation of the data were accomplished, conclusions were drawn based on the analyzed data.

V. RESULT AND DISCUSSION

In-depth analysis of the students' and instructors' insights on the reliability and validity of the new TEFL PPEE led to the emergence of the coding schemes and thematic categorizations illustrated in Table 3.

TABLE 3.
MAJOR CATEGORIES, THEMES, AND CODING SCHEMES

The new TEFL PPEE	Theme	Code
Reliability	Administration consistency	T1
	Scoring consistency	T2
	Standardization	T3
	No bias	T4
Validity	Defective face validity	T5
	Defective content validity	T6
	Defective predictive validity	T7
	Defective construct validity	T8

A. The Reliability of New PPEE

1. Instructors' insights

Close examination of the instructors' perceptions was indicative of their sanguine attitudes towards the reliability of the new TEFL PPEE. In other words, the new PPEE, in the instructors' view points, has demonstrated a high degree of reliability on account of controlling factors which lead to its unreliability. Nader, for instance, was of the opinion that the new exam has demonstrated an improved level of reliability in comparison to the old exam and argued that:

• "One considerable advantage of the new exam over the old one is its improved degree of reliability due to employing consistent administration and scoring procedures. The old Ph. D. exams were scored by university instructors who could make unfair evaluations of the exams which were planned in essay type questions and called for subjective evaluations. Standardization and nationalization of the new exam, however, have necessitated taking the advantage of a more systematic and fair scoring procedure which could significantly boost the reliability of an exam", (translated by the researchers).

Samira, Maryam, Hamid, and Javid also reflected on 'consistent administration' (T1) and 'scoring procedures' (T2) as the major grounds for the satisfactory degree of reliability in the new PPEE. Maryam more specifically viewed T1 and T2 as "the main reasons to regard the new PPEE fair" and asserted that "I favor the new exam format [including more objective items] because … [it] is more consistent in terms of the criteria". Moreover, Javid cited that:

• "The positive points regarding the new exam are the uniform scoring procedure utilized and also [the] attempts [made] to develop consistent administration environments. As you know, testing time and environment are two important test method facets which significantly impact the reliability of a test. Fortunately, the new exam, unlike the old one, has been administrated in uniform[day] times with constant time allocations, and [the] applicants are not required to sit the new exam in environments with varying degrees of familiarity. In other words, unlike the past when some students had to go to other cities [other universities] to sit the Ph. D. exam, everyone can sit the new exam in his [own] city", (interviewee's wording).

Furthermore, Adel drew on T1 as well as standardization of the new exam (T3) as the reasons why the new exam demonstrates no bias (T4). More elaborately put, he asserted that:

• "To me, the new exam is fair because it is planned in a standardized way, that is, it provides methods of obtaining samples of behaviour under uniform procedures. Systematized scoring procedures also enhance the new exam's capacity in demonstrating no bias", (interviewees wording).

2. Students' insights

Apart from benefiting from the remarks of the instructors on the reliability of the new exam, the viewpoints of the Ph. D. students were sought and subjected to scrutiny. An analysis of the students' responses to the interview questions which concerned the reliability of the new exam indicated that their perceptions were remarkably congruent with the instructors', and that they deemed the new TEFL PPEE a highly reliable exam. It is also worth mentioning that its consistency in administration and scoring procedures were the main grounds the students evidenced, not unlike the instructors, in their comments and assertions. More simply put, from among the 10 students, 7 students highlighted administration and scoring consistency as the priming features of the new TEFL PPEE enhancing its reliability. Amin for instance, argued that:

• "Administration of the new exam by a dependable organization in charge along with avoidance of subjective evaluations through utilization of a uniform scoring procedure without human interference have changed the pessimistic looks towards the Ph. D. entrance exams. The new exam is more reliable than the old one, I suppose", (translated by the researchers).

Nahid was also of the opinion that the new exam has demonstrated an improved degree of reliability. More specifically put, she stated that "to me, the new exam is perceived to demonstrate a high level of reliability because it is set under uniform administration procedures". Concurring with Nahid, Razieh reflected further on T1 as her main justification for viewing the new PPEE as a noticeably reliable exam and argued that:

• "One advantage of the new exam is that it gives all applicants the same chance to sit an exam with uniform format, testing time, test rubrics, time allocations, criteria for correctness and expected response. As a matter of the fact, controlling such facets under a uniform condition immensely improves reliability of a test", (respondent's wording).

In addition, Mahmud pointed at T2 and T4 as the foundations on which the fairness, and consequently the high reliability of the new exam are built and asserted that:

• "Fortunately, raters' bias and misevaluations do not violate the reliability of Ph. D. exams any more. I sat the old exam three times and despite my great performance in each exam, I could not qualify for the Ph. D. programs due to unfair scoring procedures employed, I suppose",(respondent's words).

In addition, Reza assumed that the new exam, unlike the old one, has controlled some of the factors which make the applicants perform differently under differing conditions including testing times and consistent test rubrics. He also pointed at T4 and argued that:

• "Familiarity of all Ph. D. applicants with the new exam's format creates an enormous opportunity to avoid test bias. To put it more simply, applicants' prior experience of sitting B. A. and M. A. program entrance exams which were developed in similar multiple choice formats awards them with similar degrees of familiarity with the new Ph. D. exam.

However, applicants sitting the old exam used to demonstrate varying levels of familiarity with responding essay type questions in such critical moments", (translated by the researchers).

B. The Validity of New PPEE

Although the respondents expressed quite positive attitudes towards the reliability of the new TEFL PPEE, both instructors and students called the validity of the new exam into critical questions. In what follows their comments will be reviewed.

1. Instructors' insights

As regards the interview questions which concerned the validity of the new TEFL PPEE, a detailed examination of the instructors' responses revealed that the new exam suffers from defective face validity (T5), content validity (T6), predictive validity (T7), and construct validity (T8).

Face validity of an exam concerns the degree to which a test appears as if it measures the knowledge or abilities it claims to measure (Bachman, 1990; Johnson, 2001; Hughes, 1989) and is uphold on the basis of the subjective judgments of observers (Ary et al., 2006; Richards & Schmidt, 2002). In regard to the new TEFL PPEE, three of the instructors viewed the new exam's format as a factor violating the sound evaluation and interpretation of the applicants' potential capabilities. Amin, for instance, concisely stated that "filtering Ph. D. applicants through multiple choice exams is a disaster". His critical comment is also similarly represented in Simin's evaluative view when she commented:

• "Although planning the Ph. D. exam in multiple choice formats improves administration and scoring procedures of the exam, it indeed hinders efficient filtering of applicants. Students' capacities to evaluate and analyze content matters are of fundamental considerations in Ph. D. programs and unfortunately it is impossible to evaluate such capacities through multiple choice exams", (translated by the researchers).

A more careful examination of Simin's viewpoint calls attention to the new exam's flaw in predicting efficiently the applicants' expected future performance in Ph. D. programs (T7). Nader correspondingly pointed at the defective predictive validity of the new exam and pointed out that:

• "Ph. D. students are supposed to make future university teachers, and as you know, a university teacher should demonstrate more complex capacities than surface knowledge of technical contents acquired through memorization. In other words, a university teacher and in particular a teacher of M. A. courses should possess analytic capabilities. Unfortunately, the new exam's features in format, content and criteria are indicative of the fact that Ph. D. student admissions through the new exam does not ensure educating and training highly qualified and analytic teachers", (translated by the researchers).

A close analysis of Ph. D. instructors' views on how well the new exam has been able to predict efficiently matriculates' performance as Ph. D. students provides a more accurate view of the new exam's predictive validity. An interesting and common theme, or better concern, emerging from the issues raised by the instructors' of both M.A. and Ph.D. courses concerned its inadequate predictive power. Javid, for instance, complained about its admission of proficiently poor Ph. D. students and favored the old PPEE because, in his view, the old system used to create a more reasonable chance of admitting more qualified Ph. D. students with specific reference to their own policies, expectations, and capacities. Meysam, furthermore, stated that:

• "You know, the Ph. D. students who come and follow the studies are really weak. It shows that the exam is not actually filtering out good students. That is why I believe it has affected the students that we admit", (respondent's wording).

Content validity concerns "demonstrating that a test is relevant…to a given area of content ability" (Bachman, 1990, P. 224). The new exam, however, was perceived by four of the instructors to employ questions partly irrelevant to the expected area of ability. They concurrently questioned the relevance of intelligence items included in the new exam as the main source of defective content validity of the new TEFL PPEE. Meysam, for instance, argued that:

• "The new one is multiple choice, as you know, and they have included other stuffs such as intelligence parts which might not be relevant at all because for Ph. D. students this is not really important", (respondent's words).

Samira also criticized the inclusion of intelligence questions in the new PPEE and stated that:

• "I really do not know why such questions should be included in a high-stakes and sensitive exam like Ph. D. exam. They actually evaluate applicants' mathematical intelligence which has nothing in common with their linguistic intelligence", (respondent's wording).

Construct validity of a test "concerns the extent to which performance on the test is consistent with predications that we make on the basis of a theory of abilities, or construct" (Bachman 1988, p. 51). Two of the instructors were of the opinion that the new TEFL PPEE does not measure what it has to. In other words, the new exam calls for a set of performances which are not consistent with the instructors' expectations of Ph. D. students' required cognitive behaviors. Close examination of Simin's view highlights discrepancy between the abilities which ought to be measured.

• "Students' capacities to evaluate and analyze content matters are of fundamental considerations in Ph. D. programs and unfortunately it is impossible to evaluate such capacities through multiple choice exams. The new exam actually measures applicants' capabilities in recall of memorized knowledge which is of limited significance in Ph. D. programs", (translated by the researchers).

Meysam also assumed defective construct validity for the new exam and argued that "we do not know what really goes in the minds of people who develop Ph. D. questions because they actually measure some trivial traits".

2. Students' insights

Analysis of the students' attitudes towards the new TEFL PPEE revealed that the new exam, as they assumed, is demonstrative of defective face, predictive, and content validities. From among the students, two students believed that the new exam's format does not seem pertinent to effective filtering of students. Saeed, for instance, cited that "multiple choice exams do not serve enough discriminative traits to filter Ph. D. applicants effectively". Nahid also attributed the poor proficiency level of Ph. D. students to the new exam's format. She stated specifically that:

• "Unfortunately, my teachers are not satisfied with the general and technical proficiency levels of the Ph. D. students admitted through the new exam in recent years. They are frequently emphasizing that the students admitted through the old system were of higher levels of proficiency because the teachers themselves used to have direct observations and control on test planning and students admissions. If it is true, it can be because of the new exam's format, I suppose. The utilized multiple choice questions make students limit themselves to memorizing trivial details such as abbreviations. And as you know, rote learning is subject to forgetting. I can remember my first days in Ph. D. programs. I could not remember most of technical content matters because I had just memorized them earlier. It was really embarrassing. I had no justifications to offer when I claimed something in classrooms because I had just easily memorized some sentences", (respondent's wording).

Examination of Nahid's perceptions about the impact of the new exam on the technical proficiency levels of the matriculates along with her teachers' assumptions about the recently admitted Ph. D. students could be demonstrative of not only defective face validity but also defective predictive validity of the new exam.

Frequent references were made to the intelligent questions by four of the students, not unlike the instructors, when they commented about the debatable content validity of the new exam. They concurrently argued that the intelligence questions recently utilized in the new PPEEs are substantially irrelevant to the general and technical contents to be measured in the exam. Razieh, for example, challenging its content validity, asserted that:

• "My second experience of sitting Ph. D exam coincided with the first administration of the Ph. D. exam in its new form. I had no presumption of the intelligence tests. I have unfortunately always been terrible at mathematics, and it was really an embarrassing moment when I encountered such questions. Believe it or not, I thought I was taking an exam other than TEFL exam", (respondent's wording).

Elaborating on the distinctions between the question types used in the old and new PPEE, Narges cast doubt on the relevance of intelligence items and asserted that:

• "There was another type of tests [in the new exam] called IQ tests. We had several texts in Persian and then several questions posed on each text, and then questions like mathematics which needed calculation. Finding relevance between such questions and what we were supposed to know was a big dilemma for me (respondent's wording)."

Ensuring the reliability and validity as the most fundamental characteristics of a test is the primary concern in test development and use (Chapelle, 1999; Neiman, 2011; Zhang et al., 2013). The expected magnitudes of reliability and validity have direct relation with the significance of the decision to be made based on the test results (Ary et al., 2006; Bachman, 1990; Cohen et al., 2007). Owing to the scope and sensitivity of the decisions to be made based on the applicants' performance on Ph. D. exams, the new TEFL PPEE ought to procure a high degree of reliability and validity. A close examination of the instructors' and students' insights on the reliability and validity of the new TEFL PPEE revealed that both parties of interest, who were practically and intimately in touch with the exam culture, expressed rather disparate views towards the reliability and validity of the new exam.

Given the format of the new PPEE, the participants' perceptions of the new exam suggested an acceptable degree of reliability for its tendency and potential to avoid previously prevalent bias in the old exams through standardization and consistency in administration and scoring procedures. This stance is in line with the general perspectives on more objective multiple choice exams in the literature (e.g. Cronbach, 1980; Dandonolli & Henning, 1990; Haladyna 2004; Johnson, 2001) where objectivity, ease of scoring, and higher consistency are the qualities attributed to multiple choice exams. It is worth pointing the fact that examinees' performances on tests vary as a function of their competencies and characteristics of the test methods. As a matter of the fact, controlling and minimizing the potential impacts of test methods could serve as a booster of test reliability (Bachman, 1990). Bachman's (1990) framework of test method facets presents a set of test characteristics which can potentially influence one's performance on a given test. The research findings in the present study demonstrated that controlling the potential effects of a set of test method facets including scoring procedure, testing time, test format and rubrics, expected response, and time allocations have been a critical consideration in the development of the new Ph. D. exams. These provisions enhanced the reliability of the exam and hence the respondents' technical attitudes.

As regards the validity of the new TEFL PPEE, the present researchers sought to accumulate and interpret complementary sorts of evidence and did not limit their investigations to collecting factual evidence on one type of validity as suggested by Messick (1992). The new exam's defective format and content relevance along with its limited capacity in predicting Ph. D. matriculates' intended behaviors on the basis of well-theorized and well-researched ability characterization were the grounds the participants reflected upon to underline the arguably low validity of the new exam. Simply put, participants' views on different aspects of the new exam were suggestive of the new TEFL PPEE's deficiency in face, content, predictive, and construct validities. In accordance with Haladyna (2004) who set the item development as the primary and most fundamental source of evidence in validating an exam, the results emerging from

the participants' views reflect the substandard item development and consequently the undesirable validity of the new exam.

Moreover, the findings concerning the defective construct validity of the new exam reinforce the findings of Rezvani and Sayyadi (2014) in another study on the new TEFL PPEE where they concluded that the development of the new exam on the basis of inappropriate competency definitions has negatively affected the study plans and strategies the TEFL Ph. D. applicants bring into service in order to prepare for the exam. In other words, it was argued that tapping into their knack of recalling memorized data instead of their ability to make evaluative judgments provokes the applicants to tailor their study plans and strategies towards comprehending and memorizing details to tackle the Ph. D. exam objective questions.

Administration of educational fairness, reduction of extravagant costs for setting the PPEE, and admission of highly qualified Ph. D. students were the IMSRT's main incentives behind the development of the new PPEE (ISNA, 2012). The new exam's tendency to avoid bias through establishing consistency in administration and scoring procedures along with minimizing certain factors which are substantially potential of fluctuating the applicants' performance has appreciably enhanced its level of reliability. It appears that such convincing and advantageous qualities have served the IMSRT to achieve its first goal of developing PPEEs in their new form, that is, consistency in administration and scoring procedures along with the reduced bias in the new exam have apparently administered a nation-wide educational fairness. On the contrary, the participants' insights regarding the predictive validity of the new exam might raise concerns about the IMSRT's policies to truly admitting academically qualified Ph. D. students as the primmest goal. It seems that the admission of competent Ph. D. students through efficient filtering exams is what counts most for universities, and the recently framed policies and plans translated and operationalized in the development and administration of the PPEEs have failed to fulfill such an overriding aim. Defective face, content, and construct validities were the other attributes about which the respondents voiced concerns. More succinctly, it might legitimately be reasoned to presume that the new PPEE developed under IMSRT's supervision is perceived to be fairly reliable but defectively valid. It is argued that the reliability of a test is not a consideration when the test is not valid (Bachman, 1990; Johnson, 2001; Messick, 1989). Therefore, it might be quite justifiable to regard the new exam's fair degree of reliability to be overshadowed by its incapacity to predict effectively the applicants' future performance through taking the advantage of questions developed based on clear constructs characterization, with content relevance and appropriacy.

VI. Conclusion and Implications

This study sought to qualitatively investigate university instructors' and Ph. D. students' insights on the reliability and validity of the new TEFL PPEE. The results generated from a careful content analysis of the accumulated data suggested that the new exam is perceived to produce a fair degree of reliability due to attempts made to minimize the impacts of the potentially adverse factors. Standardization, reduced bias, and consistency in administration and scoring procedures were among the advancements leading to the exam's acceptable level of reliability. The new exam nevertheless was believed to demonstrate defective face, content, predictive, and construct validities. In the respondents' views, the new exam has failed to predict Ph. D. matriculates' intended future performance on the basis of a well-thought-out ability characterization, a properly designed format, and more profound questions with adequate content relevance and coverage.

The results of the study might raise the policy makers' and test developers' awareness about how reliable and valid the newly designed TEFL PPEE is viewed by two prime parties of interest directly in touch with it. It is suggested to draw upon the views and expertise of TEFL assessment experts and university instructors of Ph.D. courses to pursue a more scientifically profound PPEE development approach involving better construct characterization along with more accountable format and content determination.

VII. Limitations and Research Suggestions

The examination of PPEE reliability additionally calls for undertaking empirical analysis drawing on numerical data systematically accumulated, presently unavailable. Though recommended, it should be acknowledged as one of the limitations of the study that the researchers heuristically pondered upon only the respondents' views on the exams' reliability and validity. The study was also limited in its scope. Examining a larger sample, perhaps through more quantitative approaches would provide more comprehensive and complementary validation evidence of the exam. Meanwhile, considering the likely effects such a sensitive exam may have on different stakeholders, it might be fruitful to examine the washback effects of the new PPEE on the applicants and university instructors from various aspects. To scrutinize the validity of the exam from a different perspective, it may be worthwhile to set the predictive utility as the evidence supporting the validity of the exam through examining correlations between matriculates' performance on the new PPEE and their scores on the future exams in Ph. D. programs.

References

[1] Ary, D., Jacobs, L. C., Sorensen, C. (2006). Introduction to Research in Education. B Belmont: Wadsworth, Cengage Learning.

[2] Bachman, L. F. (1990). Fundamental Considerations in Language Testing. Oxford: Oxford University press.
[3] Bachman, L., & Palmer, A. (2010). Language Assessment in Practice. Oxford: Oxford University Press.
[4] Brown, H. D. (2004). Language Assessment: Principles and Classroom Practices. London: Longman.
[5] Carlson, S. B., Bridgeman, B., Camp, R., & Waanders, J. (1985). Relationship of admission test scores to writing performance of native and nonnative speakers of English. New Jersey: Educational Testing Service.
[6] Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19* (2), 254-272.
[7] Chapelle, C. A., & Brindley, G. (2002). Assessment. In N. Schmitt (Ed.), *An introduction to applied linguistics* (pp. 268-288). London: Arnold.
[8] Chi, Y. (2011). Validation of an academic listening test: Effects of" breakdown" tests and test takers' cognitive awareness of listening process (Unpublished doctoral dissertation). University of Illinois, Urbana-Champaign.
[9] Cohen, A. D. (1984). On taking language tests: What the students report. *Language Testing, 1* (1), 70-81.
[10] Cohen, L., Manion, L., & Morrison, K. (2007). Search methods in Education. New York: Routledge.
[11] Compton, M. T. (2011). Development, item analysis, and initial reliability and validity of a multiple-choice knowledge of mental illnesses test for lay samples. *Psychiatry Research, 189* (1), 141-148.
[12] Cronbach, L. J. (1971). Test validation, In R. L. Thorndike (Ed.), *Educational measurement* (pp. 443-507). Washington, DC: American Council on Education.
[13] Cronbach, L. J. (1980). Essentials of Psychological Testing. New York: Harper and Row.
[14] Dandonoli, P., & Henning, G. (1990). An investigation of the construct validity of the ACTFL Oral Proficiency Guidelines and Oral Interview Procedure. *Foreign Language Annals, 23*(1), 11-22.
[15] Drollinger, T., Comer, L. B., & Warrington, P. T. (2006). Development and Validation of the Active Empathetic Listening Scale. *Psychology & Marketing, 23*(2), 161–180.
[16] Eda, S., Itomitsu, M., & Noda, M. (2008). The Japanese Skills Test as an On-Demand Placement Test: Validity Comparisons and Reliability. *Foreign Language Annals*, 41(2), 218-236.
[17] Elo, S., & Kyngas, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing, 62*(1), 107-115.
[18] Frain, T.J. (2009). A comparative study of Korean university students before and after a criterion referenced test (Unpublished master's thesis). University of Southern Queensland, Australia.
[19] Fulcher, G. (2010). Practical Language Testing. London: Hodder Education.
[20] Glaser, B. G., & Strauss, A. L. (1967). The Discovery of Grounded Theory: Strategies for Qualitative Research. Chicago: Aldine Publishing Company.
[21] Greenberg, K. L. (1986). The Development and Validation of the TOEFL Writing Test: A Discussion of TOEFL Research Report 15 and 19. *TESOL Quarterly, 20*(3), 531-544.
[22] Haladyna, T. M. (2004). Developing and Validating Multiple-choice Test Items (3rd edition). New Jersey: Lawrence Erlbaum Associates.
[23] Hamp-Lyans, L,. & Lynch, B. K. (1998). Perspectives on validity: A historical analysis of language testing conference abstracts. In A. J. Kunnan (Ed.), *Validation in Language Assessment* (pp. 253-276). New Jersey: Lawrence Erlbaum Associates.
[24] Hashmi, M. A. (2000). Standardization of an Intelligence Test for Middle Level Students (Unpublished master's thesis). Bahauddin Zakariya University, Multan.
[25] Henning, G. (1987). A guide to language testing: Development, evaluation and research. Massachusetts: Newbury House.
[26] Hissbach, J. C., Klusmann, D., & Hampe, W. (2011). Dimensionality and predictive validity of the HAM-Nat, a test of natural sciences for medical school admission. *Medical Education, 25*(3), 1-11.
[27] ISNA, (2012, November 21). Why did the Iranian Ph. D. program entrance exams altered to be semi-centralized [Online forum comment]. Retrieved from http://isna.ir/fa/news/91090100296.
[28] Ito, A. (2012). A Validation Study on the English language test in a Japanese Nationwide University Entrance Examination. *Asian EFL Journal, 7* (2), 101-127.
[29] Johnson, M. (2001). The Art of Non-conversation. New Haven & London :Yale University Press.
[30] Kazemi, A., & Sayyadi, A. (2014). Effect of University Entrance Exam on Gifted High School Students' Motivation Scrutinized: An Iranian Perspective. *International Journal of Applied Linguistics & English Literature*, 3 (5), 150-165.
[31] Kirkpatrick, R., & Hlaing, H. L. (2013). The Myanmar University entrance examination. *Testing in Asia, 14* (3), 1-15.
[32] Magnan, S. S. (1987). Rater reliability of the ACTFL Oral Proficiency Interview. *The Canadian Modern Language Review, 43*(1), 267-276.
[33] Mahmoudi, L., & Abu Bakr, K. (2013). Iranian Pre-university English Teachers' Perceptions and Attitudes towards the Iranian National University Entrance Exam: A Washback Study. *International Journal of Education & Literacy, 1*(2), 57-83.
[34] McMillan, J. H., & Schumacher, S. (2006). Research in education: Evidence-Based Inquiry. New York: Pearson Education.
[35] Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer, & H. I. Braun (Eds.), *Hillsdale* (pp. 153- 186). New Jersey: Lawrence Erlbaum Associates.
[36] Messick, S. (1989). Validity. In R.L. Johnson (Ed.), *Educational Measurement* (pp. 3-104). New York: American Council on Education.
[37] Messick, S. (1992). The interplay of evidence and consequences in the validation of performance assessments. New Jersey: Educational Testing Service.
[38] Muchinsky, P.M. (1993). Validation of intelligence and mechanical aptitude tests in selecting employees for manufacturing jobs. *Journal of Business and Psychology, 7* (4), 373-382.
[39] Neiman, D. (2011). Exercise Testing & Prescription. New York: McGraw-Hill.
[40] Patternson, B. F., & Ewing, M. (2013). Validating the Use of AP Exam Scores for College Course Placement. London: College Board Reports.
[41] Purpura, J. E. (1998). The development and construct validation of an instrument designed to investigate selected cognitive background characteristics of test-takers. In A. J. Kunnan (Ed.), *Validation in Language Assessment* (pp. 111-140). New Jersey: Lawrence Erlbaum Associates.

[42] Ramezaney, M. (2014). The washback effects of University Entrance Exam on Iranian EFL Teachers' Curricular Planning and Instruction Techniques. *Social and Behavioral Sciences*, 98, 1508-1517.
[43] Razavipur, K. (2014). On the Substantive and Predictive Validity of University Entrance Exam for English Majors. *Journal of Research in Applied Linguistics, 5*(1), 77-90.
[44] Rezvani, R. (2010). Modeling Test-taking Strategies and Their Relationship to Translation Test Performance: A Structural Equation Modeling (SEM) Approach (Doctoral dissertation). Shiraz University, Shiraz.
[45] Rezvani, R. & Sayyadi, A. (2014). A qualitative study on the Wash-back effects of the new Iranian TEFL Ph.D. program entrance exam on applicants' study plans and strategies: Instructors' and applicants' insights. *Journal of Research in Applied Linguistics, 5*(2), 113-127.
[46] Richards, J.C., & Schmidt, R. (2002). Longman Dictionary of Language Teaching and Applied Linguistics, 3rd edition. London: Longman.
[47] Sabers, D. L., & Feldt, L. S. (1968). The Predictive Validity of the Iowa Algebra Aptitude Test for Achievement in Modern Mathematics and Algebra. *Educational and Psychological Measurement, 28* (1), 901-907.
[48] Salehi, H., & Yunus, M. M. (2012). The Washback Effect of the Iranian Universities Entrance Exam: Teachers' Insights. *GEMA Online Journal of Language Studies, 12* (2), 609-628.
[49] Stansfield, C. W., & Kenyon, D. M. (1992). Research on the comparability of the Oral Proficiency Interview and the Simulated Oral Proficiency Interview. *System, 20*, 347-364.
[50] Stager, J. L. (1993). The comprehensive Breast Cancer Knowledge Test: validity and reliability. *Journal of Advanced Nursing, 18* (1), 1133-1140.
[51] Thompson, I. (1995). A study of Inter-rater Reliability of the ACTFL Oral Proficiency Interview in five European languages: Data from ESL, French, German, and Spanish. *Foreign Language Annals, 28*, 407-422.
[52] Weir, C. J. (2005). Language Testing and Validation: An Evidence-based Approach. New York: Palgrave Macmillan.
[53] Zamani, G., & Rezvani. R. (2014). A Comparative Study of Iran's TEFL and English Translation UEEs: Do High-stakes Tests Assess Critical Thinking? *Theory and Practice in Language Studies*, 4(2), 379-386.
[54] Zhang, L., Aryadoust, V., & Zhang, L. J. (2013). Development and Validation of the Test Takers' Metacognitive Awareness Reading Questionnaire (TMARQ). *The Asia-Pacific Education Researcher, 38* (3), 50-65.
[55] Zhao, Z. (2013). Diagnosing the English Speaking Ability of College Students in China –Validation of the Diagnostic College English Speaking Test. *RELC Journal, 44*(3) 341 –359.

Reza Rezvani did his undergraduate (English Literature) and master degrees (TEFL) at Shiraz University. He received his Ph.D. degree from the same university in 2010. He has taught Translation Studies and TEFL courses. His areas of interest include Language and translation Assessment, Materials Development & Evaluation, Teacher Education, and Teaching Language Skills.

Ali Sayyadi holds a bachelor's degree in English Literature and a master's degree in English Language Teaching from Yasouj University, Yasouj, Iran. His research areas of interest include language assessment, discourse analysis, conversation analysis, and vocabulary learning.