The Journal of Teaching Language Skills (JTLS)
7 (2), Summer 2015, Ser. 79/4, ISSN: 2008-8191
pp. 1-37

INVESTIGATING THE VALIDITY OF PHD ENTRANCE EXAM OF ELT IN IRAN IN LIGHT OF ARGUMENT-BASED VALIDITY AND THEORY OF ACTION

Alireza Ahmadi
Associate Professor
Shiraz University
arahmadi@shirazu.ac.ir

Ali Darabi Bazvand *
PhD Candidate
Shiraz University
alidarabi1350@gmail.com

Rahman Sahragard
Associate Professor
Shiraz University
rsahragard@rose.shirazu.ac.ir

Seyed Ayatolah Razmjoo
Associate Professor
Shiraz University
arazmjoo@rose.shirazu.ac.ir

Abstract

Although some piecemeal efforts have been made to investigate the validity and use of the Iranian PhD exam, no systematic project has been specifically carried out in this regard. The current study, hence, tried to attend to this void. As such, to ensure a balanced focus on test interpretation and test consequence, and to track evidence derived from a mixed–method study on the validity of Iranian PhD entrance exam of TEFL (IPEET), this study drew on a hybrid of two argument-based structures: Kane's (1992) argument model and Bennett's (2010) theory of action. Resting on the network of inferences and assumptions borrowed from the hybridized framework, the study investigated the extent to which the proposed assumptions would be supported by empirical evidence. It also examined the unintended consequences that may possibly be revealed through this validity investigation. Three sources of data informed the present study: (a) Test score data from about 1000 PhD applicants' taking IPEET test administered in 2014, (b) questionnaires completed by university professors and PhD students of TEFL, and finally, (c) telephone and focus-group interviews with university professors and PhD students of TEFL, respectively. The results from the analysis of mixed-method data indicated that all the inferences proposed for this study were rebutted, suggesting that some unintended consequences have happened to the technical as well as the decision quality of this test, hence its invalidity. Findings also

Received: 09/10/2015 Accepted: 29/06/2016
* Corresponding author

provided valuable insights and suggestions for the betterment of the present content and current policy of IPEET in Iran.

Keywords: mixed method study, argument-based validity, theory of action, unintended consequences

1. Introduction

A decentralized assessment system was previously practiced to screen PhD applicants in Iran. In the decentralized PhD exam no central bodies from top-tier decision makers such as Ministry of Science, Research and Technology (hereafter, MSRT) and the National Organization for Educational Testing (NOET) were in control of this admission system. Different universities administered their own examination differently in different formats and at different times. The screening was based on a written performance-based assessment (or sometimes an MC test) in which applicants were required to respond to essay-type knowledge questions (or MC questions) based on which those who passed the cut-score (determined and decided by each specific university) were allowed to attend an interview. The overall evaluation was based on the local written performance assessment, the MC test, and the oral interview. However, this traditional system was claimed in the oral literature not to be scientific and fair enough; that is, most of the PhD students were selected from the MA students of the same university. Furthermore, most of the PhD applicants were relatively dissatisfied with the entrance criteria of the higher education in Iran (Kiany, Shayestefar, Ghafar Samar, & Akbari, 2013). Therefore, these problems casted some doubts on the reliability and validity issues of this type of evaluation.

Currently, following the criticism leveled against the decentralized admission system, a semi-centralized assessment system is practiced for screening PhD applicants. Every year, a resounding number of MA graduates (NOET news, 2013) from different majors, in 30 capital cities in Iran take part in IPEET. As released on the official sites of NOET and MSRT, the apparent intentions behind introducing this test were both to solve some of the deleterious effects of the decentralized local examinations and to take more control and power on the acceptance and non-acceptance of candidates for doctoral programs. Annually, this test is administered in March and the primary results are released on NOET site at the end of May. The IPEET test subsists of a test of academic talent, a general English proficiency test and a specialized knowledge test, all appearing in MC format. The knowledge test which is aimed at measuring the candidates' expertise in the field of Teaching English as a Foreign Language (TEFL) is supposedly related to the courses students have passed in the MA or even BA program. In fact, it assesses the students' specialized knowledge in areas which are assumed to be the prerequisite for entering the PhD program since the PhD program is built on such areas of knowledge. As such, the knowledge test of IPEET includes questions on

linguistics (15 items), foreign/second language teaching methods (15 items), research methods (15 items), language assessment (15 items), theories and issues of language learning and teaching (30 items), and finally sociolinguistics and discourse analysis (10 items). Based on a criterion (cut-off score) determined and decided by MSRT and NOET, some applicants three (sometimes more) times the number of capacities each university reports to these two organizations are introduced to each respective university to be interviewed. The interview questions are related to the participants' research backgrounds, academic records, and expertise (technical knowledge). The final admission will be based on the aggregate scores from the PhD entrance exam in written form and the oral interview.

Given that the written exam plays a gate-keeping role and requires PhD applicants to pay many costly prices to be well prepared for it, this test is of paramount importance in screening PhD applicants for admission into PhD programs in Iran; therefore, it was assumed any technical problem with the content of such test and consequently any inappropriate decision made on the information yielded by it may contribute to some potential problems such as introducing some applicants as false negatives and some others as false positives. Further, it was hypothesized if the present PhD exam is problematic in terms of predictive validity, little success in PhD courses can be demonstrated on the part of PhD candidates being screened through this test, hence creating some potential problems for both post graduate university professors in dealing with these unsuccessful candidates and PhD candidates themselves in fulfilling the course requirements. Although other factors such as the applicants' performance in the interview session, their educational and research background and their GPA scores may have their own effects, it can be claimed all these factors may be more or less dependent on the written exam. For instance, it may be the case that some PhD applicants with good academic and research background and with a good ability in oral performance are unable to show their best ability as they fail the written exam just because the instrument is inappropriate.

1.1 Research questions

This study primarily aimed at investigating the content and use of IPEET in light of argument-based validity and theory of action, throwing some light on the betterment of the technical and decision quality of this test in Iran. More specifically, it tried to answer the following research questions:

RQ 1. To what extent did the characteristics of the test items and the conditions of test administration in the context of IPEET introduce minimal construct irrelevant variance (CIV) in observed scores (Evaluation)?
RQ 2. To what extent are IPEET and its individual subtests internally reliable? Is there any source of unreliability creeping into the test (Generalization)?

RQ 3. What did applied linguistic experts think about (a) the relevance of IPEET test tasks to the content of PhD instructional courses and, (b) the relative success of PhD students in fulfilling the requirements of PhD courses (Extrapolation)?
RQ 4. To what extent did test practitioners apply reasonable decision standards with regard to (a) informing affected stakeholders about the type of decisions they will make on the admission of PhD applicants (b) reporting test scores and score descriptors in a clear and understandable way (c) reporting test scores to test takers in a timely and systematic manner (Intermediate Actions)
RQ 5. What did applied linguistic experts think about the relative effects of the use of IPEET on both university professors in terms of promoting good instructional practice and PhD students in terms of their relative success in PhD courses (Ultimate Effects)?
RQ 6.What possible action mechanisms did stakeholders suggest for the betterment of unintended consequences materialized in the present validity study [in terms of content and decision quality] (Ultimate Actions)?

1.2 Articulating the validity framework

As summarized and presented in Figure 2, this framework subsists of two types of arguments. The measurement argument and the theory of action argument. In the measurement argument three specific claims such as evaluation, generalization, and extrapolation adapted from Kane's interpretive argument were articulated. In the theory of action argument, on the other hand, three sequential claims consisting of intermediate actions, intended ultimate effects, and hypothesized ultimate actions adapted and reconceptualized from Bennett (2010) were localized. More description of details is provided in Table 1 below.

Table 1. Validity framework for IPEET

Inference in the Interpretive Argument	Warrant Supporting the Inference	Assumptions Underlying Warrant	Backing Sought to Support Assumption
Evaluation	Observations of PhD students 'performance on IPEET tasks as well as the characteristics of tasks themselves are evaluated to provide observed scores informative of target academic domain.	1. (Statistical) characteristics of IPEET test items introduce minimal CIV in observed scores and are appropriate for norm-referenced decisions. 2. Test administration conditions introduce minimal CIV and are appropriate for providing evidence of academic target language abilities.	1. Logistic Regression (LR) is applied to investigate gender DIF. 2. Questionnaire and interview data are analyzed to investigate test characteristics. 3. Questionnaire and interview data are analyzed to investigate test administration conditions.

Inference in the Interpretive Argument	Warrant Supporting the Inference	Assumptions Underlying Warrant	Backing Sought to Support Assumption
Generalization	Results from the total IPEET and its subtests are consistent and are considered as estimates of expected scores over multiple tasks and occasions	1. IPEET test and its subtests have an acceptable level of reliability. 2. There is no source of unreliability creeping into the test.	1. Cronbach alpha coefficient is applied to investigate the internal consistency of the test. 2. Test score data and stakeholders' opinion are analyzed to investigate the content and the administration conditions of the test.
Extrapolation	The construct of academic English language teaching abilities as assessed by the specialized section of the IPEET can account for the quality of language performance on relevant tasks in academic PhD courses.	1. The content of IPEET test is fully related to the criteria of PhD courses taught in PhD programs. 2. Performance on IPEET test predicts relative success of PhD students in PhD courses.	1. Experts' judgment is used to investigate the relative relatedness of IPEET test tasks with PhD courses. 2. Experts' judgment is used to investigate the relative success of PhD candidates.
Intermediate Actions	Reasonable decision standards are made for the admission of PhD applicants	1. Test practitioners at NOET inform university professors and PhD applicants about the type of decisions they will make on the admission of PhD applicants. 2. Test practitioners at NOET report test scores in ways that are clear and understandable to PhD applicants. 3. Test practitioners at NOET do not change their admission decisions from one year to another.	1. Stakeholders' opinion is analyzed to investigate the decisions are based on a collective judgment. 2. Stakeholders' opinion is analyzed to investigate the presentation of test scores is understandable to PhD applicants. 3. Stakeholders' opinion is analyzed to investigate the admission decisions are systematic.
Ultimate Effects	The quality of decisions made by policy makers leads to beneficial consequences for most affected stakeholders and influence instructional practice	1. PhD students benefit from the decisions made on PhD admissions through the use of scores from multiple –choice PhD exam of TEFL. 2. The use of the test helps promote good instructional practice and effective learning in ELT instructional settings	1. Experts' judgment is used to investigate the relative success of PhD candidates 2. Stakeholders' opinion is used to investigate the use of IPEET helps promote good instructional practice
Proposed Ultimate Actions	The admission system of IPEET is satisfactory and may not need substantial revision and improvement.	1. University professors suggest the technical and the decision quality of the IPEET is appropriate and may not need substantial revision. 2. PhD students suggest the	1. Logistic Regression (LR) analysis are used to investigate gender DIF, and if so to take some actions 2. Questionnaires and interviews data from

Inference in the Interpretive Argument	Warrant Supporting the Inference	Assumptions Underlying Warrant	Backing Sought to Support Assumption
		technical and the decision quality of the IPEET is appropriate and may not need substantial improvement.	university professors are analyzed to investigate the possible problems with the technical and decision quality of IPEET. 3. Questionnaires and interviews data from PhD applicants are analyzed to investigate the possible problems with the technical and decision quality of IPEET.

3. Method

3.1 Design

This study applied a concurrent triangulation mixed-methods design (QUAL+ QUAN) in which two methods are used in a separate and parallel manner and the results are integrated in the interpretation phase (Dörnyei, 2007). This design is specifically employed for validation purposes. In the present study, the interview data from the qualitative phase was used to provide an in-depth picture of the perceptions of participants on the technical and decision quality of the use of IPEET in Iran. The quantitative data, however, helped to recognize the general factors associated with the views of PhD students, and university professors on the content of IPEET in Iran.

3.2 Participants

The participants in the current study comprised three groups of stake holders. The first group consisted of all the PhD applicants who had taken IPEET in January, 2013, regardless of whether they were subsequently admitted to PhD programs (n= 999).

The second group of participants consisted of 103 PhD students of TEFEL (57 males and 46 females) studying at different PhD programs in Iran, ranging between 25 and 40 years. In fact, this group of participants was selected from the first group. They were selected based on snowball sampling and received a questionnaire through email. Thirty five of them were also interviewed through focus groups.

The third group of stakeholders invited to participate in the present study was a restricted sample of 20 university professors (19 males and 1 female) who virtually had some experience teaching some PhD courses of TEFL in PhD programs in Iran. Their age ranged from 36 to 57 years. Further to their busy schedule and mere unwillingness to cooperate, due to the fact that the number of university professors with the required characteristics was very limited, such a small sample took part in the study. Unlike PhD students, they received the

questionnaires in person. Among them, 12 were recruited for the telephone interview.

3.3 Instruments and data collection

Instruments used to collect data for this study varied. They included IPEET test score data, closed-ended questionnaires, a structured telephone interview and a focus group. Descriptions of the instruments as well as the procedure for their use are as follows:

IPEET test score data. IPEET is a centralized test administered annually to PhD applicants of TEFL in Iran (for more information on this test see the section on local context).This test subsists of a test of academic talent, a general English proficiency test and a knowledge test. Total test score data (the administration of 2013) for all the PhD applicants of TEFL (n=999) were provided by EAO at the request of Shiraz University, Iran. However, for the present study, only the knowledge subtest was investigated. The total test scores were analyzed by Cronbach alpha and logistic regression (LR) to estimate their reliability coefficients and differential item functioning (DIF) respectively. The results informed the generalization and evaluation claims of the instrument.

Focus group interview. Among 103 PhD students who were selected via snowball sampling and subsequently completed the questionnaires, 35 respondents participated in a 30-minute semi-structured focus-group interview. It was made clear to every participant that the purpose of the survey was to investigate the validity of IPEET, and that strict anonymity and confidentiality would be guaranteed. They were divided into groups of five or six based on the respective universities. The visits took place in the course of October 2014 and early May 2015. The language of the interviews was English. The interviews were audio recorded and transcribed. The items intended to gain insight from exam candidates with regard to the characteristics of IPEET test and its subtests (in terms of difficulty), conditions of test administration, and opinions with regard to possible ways of improvement for IPEET. The data informed evaluation, intermediate actions and possible suggestions for improvement leading to ultimate actions.

Telephone interview. The next instrument used during the study was a structured telephone interview conducted with 12 university professors, selected from among 20 participants. Due to budgetary and time constraints they were all interviewed through telephone. For the betterment of the quality of responses, a copy of IPEET test (administered in 1393) was provided to the participants. It was also made clear to every participant that the purpose of the survey was to collect research information, and that strict anonymity and confidentiality would be guaranteed. The interview took place in the course of

November and early January (2014). Each long-guided interview lasted 30 to 45 minutes. The language of the interviews was English. The interview items were aimed at soliciting university professors' opinion regarding: the relevance of the IPEET test tasks to the PhD courses, PhD students' success in PhD programs, the quality of decisions made by testing agencies and possible suggestions for the betterment of IPEET, informing the claims on evaluation and extrapolation, intermediate actions, ultimate effects and ultimate actions. Each interview was recorded on an audio-cassette and subsequently transcribed.

Questionnaires. In order to develop questionnaire items, we relied heavily on two sources. The first was the qualitative, exploratory data gathered from the participants through recorded semi-structured and focus-group interviews in such a way that some major themes emerged and questionnaires were explored and developed based on those themes. The second source was the questions adapted from the relevant literature. As such, two types of questionnaires were developed. The first was completed by PhD students of TEFL (n=103) through email. The same insights and evidence were collected as were sought in the focus-group interviews.

The second questionnaire was responded by the university professors (n=20) having taught some specialized courses in the PhD programs. The questions focused on the same issues as reported in the telephone interviews.

3.4 Data analysis procedure

After the data-gathering process, the next step was to analyze both qualitative and quantitative data. Relying on Glaser and Strauss's (1967) method of constant comparison, the iterative qualitative analysis of data for both focus group and telephone interviews included : 1) reading through transcriptions to obtain an overall flavor of the responses of the interviews and making an exploration of the data; 2) developing a general category scheme of the participants' responses based on specified labels; and 3) aggregating similar codes together to develop themes and identifying categories and sub categories. More details on ensuring the quality criteria of content analysis are presented in the next part.

For the quantitative phase, both test data and questionnaires were analyzed. For the IPEET test score data, Cronbach alpha was applied to estimate the reliability coefficients of the test and its subtests. Further, LR model was used to investigate the possible gender DIF items. As regards the PhD students and university professors' opinions, a series of Binomial tests of significance were used to report the participants' responses to the specified questionnaire items in the form of observed proportions.

3.4.1 Validity of the mixed -method design

Informed by the validity standards of quantitative and qualitative paradigms, early validity efforts for mixed methods studies tended to assess these methods separately (Ary, Jacobs, Sorenson, & Razavieh, 2010). Recently, however, several researchers (e.g, Teddlie & Tashakkori, 2003, 2006) have suggested that the validity criteria for mixed methods research need to be addressed by its own criteria. For example, Teddlie and Tashakkori, (2006) use the term "inference quality" to refer to an overall assessment of validity in mixed methods research. They suggested two ways to examine the inference quality. One approach is the design quality, which deals with the methodological rigor defined as "the extent to which the QUAL and QUAN components of a mixed method study are combined or integrated in a way that the overall design displays complementary strengths and non-overlapping weaknesses of the constituent methods" (Dörnyei, 2007, p. 63) which is affected by within-design consistency (Ary et al., 2010). The second aspect is interpretive rigor, which deals with the accuracy of evaluating the validity of inferences or interpretations (Ary et al., 2010; Kim, 2008). For addressing the design quality, Greene (2007) suggests that researchers adhere to quality criteria, while for attending to interpretive rigor we need consistency of inferences among the findings in terms of "type, intensity and scope" (Dörnyei, 2007).

In line with these recommendations, the present study adhered to the methodological and interpretive standards of qualitative and quantitative approaches. Therefore, special heed was paid to ensure the instruments (test score data, questionnaires and interviews) are appropriate, the procedures used to collect the data through these instruments are systematic and data analysis procedures are based on the standards of mixed method approach. Further, care was exercised to ensure all the processes of data collection, data analysis and data interpretation qualitatively fit the topic, the research questions and the design of the study.

To ensure the quality (validity) of the interview in terms of item development and actual implementation, the present study followed the suggestions made by Cohen, Manion, and Morrison (2007) and Dörnyei (2007). As pointed out by Cohen and his colleagues, "the most practical way of achieving greater validity is to minimize the amount of bias as much as possible" (2007, p.150). As such, the present study attended to four potential sources of bias such as the content of interview items, sample size, specific behavior of the interviewer and the characteristics associated with the respondent. With regard to the quality of content the present study tried to avoid using leading questions (It was astonishing, wasn't it….?) and loaded or ambiguous words. Another equally important issue related to the quality criteria of interview data was sample size. According to Dörnyei (2007), an interview study with an initial sample size of 6-10 might work well. In line with this suggestion, 12 post graduate university professors and five groups of PhD

candidates with each group consisting of 5 or 6 participants were solicited about their perceptions with regard to the validity of IPEET. As regards the behavior of the interviewer as a source of bias, the researcher himself as the interviewer tried to minimize the amount of bias to the least amount possible. For example, he tried to be neutral and attempted not to ask sensitive questions like those that require private answers. Moreover, attempts were made to minimize the issue of "social desirability bias" which unavoidably influences the truthfulness of the interviewees' responses (Dörnyei, 2007). According to Cohen et al. (2007) such practical measures may enhance the reliability of interviews. To satisfy the credibility of the findings, *member checks* strategy was applied. For example, after collecting and analyzing the data, participants were called to check the data they produced during the interview. Soliciting feedback from participants in this way is, according to Maxwell, the "single most important way of ruling out the possibility of misinterpretation of the meaning of what they say and the perspective they have on what is going on" (1996 p. 94). Moreover, by providing sufficient details of the data (in the form of direction quotations) to take the reader into the context being described, the present study addressed the important issue of "thick" description.

As regards the questionnaires, the study paid special attention to maximizing the quality of both their development and their final implementation. To this end, issues such as type face, wording, instructions, coverage, statistical piloting of the items, and authenticity were considered. In order to improve the wording and instructions, for example, the questionnaire items were given to two applied linguistics experts to be checked for their ambiguities, readability levels, type of scaling, redundancies, and clarity. Based on their feedback, some items were discarded and some were added. As for the reliability estimates, PhD students' questionnaire was tested with a sample of participants being as similar to the target population as possible. The overall reliability value for the PhD students' questionnaire estimated through Cronbach's Alpha was turned out to be .86, indicating that the instrument was highly reliable. To establish the authenticity of the data, the researcher explained the purpose of the study, ensured the ethicality of research by promising to protect the identity of the respondents and by ensuring to maintain the confidentiality of the data.

Other important measures were taken to attend to the quality criteria of the present mixed method study. For instance, the study focused on the explicit and systematic use of mixed methods design as guided by Creswell and Clark's (2011) suggestions: First, qualitative and quantitative samples were drawn from the same population to make the data comparable. For example, form among the 103 PhD students who were subsequently given the questionnaire, 35 were interviewed. The same was true for post graduate university instructors; from among the 20 professors who responded to the questionnaires, 12 were interviewed. Second, although, separate data collection procedures for QUAL

and QUAN were used, the results informed the same research questions. Third, the analysis of the data was carried out separately but they were merged in the interpretation phase.

Finally, to enhance the quality of the interpretive rigor of the study, the researcher followed mixed method thinking in sequencing the instruments (for example, first QUAL then QUAN specified as exploratory design), collecting the mixed method data, analyzing the data, interpreting the findings, and guiding the implications or conclusions of the study. All in all, it can be argued that the present mixed-method study followed a systematic procedure in data collection, data analysis, and data interpretation. In this way the present study tried, though not idealistically, to fulfill the requirements of inference quality as suggested by researchers.

4. Results

The present mixed method study applied a hybridized framework to investigate the IPEET in Iran. For the technical quality of the test, evaluation, generalization, and extrapolation inferences adapted from Kane's framework were examined. For the test use and consequences, however, intermediate actions, ultimate effects, and ultimate actions were reconceptualized from Bennett's (2010) theory of action. It is in this order that descriptions of the results for each individual inference are reported in the following sections.

4.1 Evaluation inference

To answer the first research question and to examine this inference, test score data from 999 participants taking IPEET in January 2014 as well as qualitative and quantitative responses from stakeholders were analyzed. This claim was characterized in terms of test characteristic and test conditions.

Assumption 1.Test characteristics. Evidence for this assumption was sought through the statistical method of LR and stakeholders' opinion. Logistic regression has been widely considered as one of the best statistical methods for investigating DIF (Zumbo, 1999); therefore this method was applied to track evidence for gender DIF. The overall results of LR DIF are summarized in Table 2. Worthy of note is that of the total of 100 items analyzed for DIF only twelve items with significant DIF values at 0.05 level of significance were flagged for DIF. As such only the information for these items are presented in table below. As reported in this table, of the 12 items identified as showing DIF four items were detected in the linguistics section, two in the research methods subtest, two in Testing, one in SLA, one in Discourse and finally two items in Sociolinguistics. Further, it is reported that the number of DIF items for males and females was equal, indicating that DIF items might balance out each other in the test level analysis (Drasgow, 1987; Takala & Kaftandjieva, 2000), what Sireci and Rios (2013) call it, DIF "cancellation". With regard to DIF effect

size, the present study followed a "blended decision rule" (Zumbo, 2008) including both the effect size and *p* value. Likewise, it was observed that all obtained R2 values manifested a negligible DIF magnitude (category A); that is, they were smaller than .035 and .05. As such, minimal construct irrelevance variance was introduced in observed scores.

Table 2 LR results for the DIF items identified in IPEET test items.

Item	subtest	Favored	*R*2 **effect size**			⍺2	Category
			UDIF	NUDIF	DIF		
20	L	M	.....	.005	.005	4.488	A
23	L	F	.008	.....	.008	7.953	A
26	L	M	.012	.....	.012	10.231	A
28	L	F	.....	.005	.005	4.19	A
36	R	F	.008	.....	.008	7.312	A
40	R	M	.....	.005	.005	4.163	A
46	T	M	.007	.....	.007	6.31	A
49	T	F	.....	0.019	.019	12.821	A
78	SL	F	.016	.....	.016	9.706	A
95	D	F	.....	.000	.000	22.506	A
97	S	M	.015	.....	.015	11.425	A
99	S	M	.....	.006	.006	4.431	A

Notes. **p* <.05; L= Linguistics; R= Research; T= Testing; SL = SLA; D= Discourse; S= Sociolinguistics; M= Male; F= Female; A = Negligible DIF

With regard to stakeholders' opinion, analysis of transcribed telephone interviews with university professors showed, not quite surprisingly, that the majority were concerned about the difficulty level of the test. They argued that some items were quite easy and some were unduly difficult. As expressed by a university professor: "Sometimes you can see a kind of, I can say, wide differences with most of the items average or above the average in terms of difficulty but you can see a couple of items that are really difficult and a couple of items that are really easy." Another participant added that "well in terms of difficulty level I suppose well, this has not been sufficiently taken well into account for one thing our test is not a standard one, in reality sometimes you find some items unduly difficult and sometimes very easy to deal with and sometimes reasonable you know".

Analysis of transcribed focus group interview also corroborated what university professors perceived of the difficulty level of the items included in IPEET. One of the PhD students eloquently stated:

> I think some questions are really easy. They are written to be answered by M.A students even by B.A students, but most of the questions about 70% are really difficult. These are more important and some questions are so difficult and cannot be even answered by professors. They are really difficult and they are really for memory. We should answer for example; some of the names are really new for us. The dates, some of the

methods. I think they are made at the time of making the questions. So I think the content is not representative of M.A.

PhD students' responses to questionnaires also confirmed the above findings. They, almost, did express the same collective opinion with regard to the level of difficulty of the items. As shown in Table 3, of 103 respondents, about 60 participants (58%, p=.114), answered that the total test is difficult. it is also reported that some sub-tests like teaching(72%) and linguistics (81%) designed based on the BA courses are significantly easy and some others like teaching issues(64%), testing(73%), and discourse (63%) are reported to be significantly difficult.

Table 3. Binomial test for the difficulty of PhD entrance exam of TEFL

IPEET and its subtests		Category	N	Observed Prop.	Test Prop.	Sig. (2-tailed)
Total test	Group 1	easy*	43	.42	.50	.114
	Group 2	difficult +	60	.58		
		Total	103	1.00		
Linguistics	Group 1	easy*	83	.81	.50	.000
	Group 2	difficult+	20	.19		
		Total	103	1.00		
Teaching Methods	Group 1	easy*	74	.72	.50	.000
	Group 2	Difficult+	29	.28		
		Total	103	1.00		
Theories & Teaching issues	Group 1	easy*	37	.36	.50	.006
	Group 2	Difficult+	66	.64		
		Total	103	1.00		
Language assessment	Group 1	easy*	28	.27	.50	.000
	Group 2	difficult+	75	.73		
		Total	103	1.00		
Research methods	Group 1	easy *	69	.67	.50	.001
	Group 2	difficult+	34	.33		
		Total	103	1.00		
Socio linguistics & discourse	Group 1	easy*	38	.37	.50	.010
	Group 2	difficult+	65	.63		
		Total	103	1.00		

* Combined 'Easy' and 'Very easy' responses
+ Combined 'Difficult' and 'Very difficult responses

As such, the knowledge test of IPEET includes questions on linguistics (15 items), foreign/second language teaching methods (15 items), research methods (15 items), language assessment (15 items), theories and issues of language learning and teaching (30 items), and finally sociolinguistics and discourse analysis (10 items).

Inadequacy of the number of MC items is severely prone to the problem of construct underrepresentation which is of major concern to the assessment enterprise. Based on oral literature it seemed that IPEET is problematic in this regard. Conscious of this handicap leveled against IPEET and unable how to

address it, university professors and PhD students expressed concerns in this regard. What could be inferred from collective perceptions of university professors was that the number of items is adequate in its totality but the number of questions allocated to each subtest is disappointing, as complained by one of the participants: "Well when the mode of presentation is MC, then 100 items seems to be an adequate number but that disappoint you when you see that 100 items are to be divided into a number of sub-sections each of which allocated 5, 10 or I don't know 15 questions". Another added: "Well, 5 items for discourse analysis it's very much underrepresented in the questions". This major problem is also the concern of a university professor as a testing specialist. "I think because testing is the part and parcel of TEFL, actually it should include more items and I don't think 15 items actually can show the future performance of candidates' knowledge of language testing, so they are inadequate". Not quite surprisingly, an academically celebrated figure as a participant confirmed that language teaching issues are underrepresented in the IPEET: "actually, the number of items, I mean 15 items on the theoretical issues and 15 on language skills is not enough more items should be included, because this is quite relevant to the nature of the program which is TEFL, more topics, more items or tasks".

PhD students, as interviewed through focus group, did not take up a contrary position. They agreed with university professors with regard to the inadequacy of the number of items. As one of the PhD candidate confided, "There were too many items. I'm speaking about the total test but they were not divided proportionately between sub-tests. Moreover, there were voices of negative view regarding the adequacy of the items: " the number is not ok" or as another said contentedly with regard to discourse sub-test, "it's a kind of discrimination because I'm interested in discourse and 5 for this important topic is not enough".

Thus, the analysis of collective interview and questionnaire responses from stakeholders, then, would seem to reveal that difficulty level as well as inadequate number of items could be introducing construct-irrelevant variance in the observed scores of IPEET. Therefore, the first assumption, stating that the characteristics of IPEET introduce minimal construct-irrelevant variance was rebutted.

Assumption 2. Test conditions. With regard to the present study, complaints from PhD students and PhD applicants revealed in the oral literature casted some doubts on the quality of IPEET test administration. As such, during the focus group interviews, PhD students, as a group of stakeholders mostly affected by the administration conditions of the test, were asked their perceptions with regard to IPEET administration conditions. Their responses were classified according to three recurring themes: the exam proctors/inspectors, the testing venues, and the timing of the test administration.

The proctor/inspector factor examined the presence of the inspectors, satisfaction with the behavior of the proctors, calmness of the session, refreshments and finally the possibility of cheating. The testing venue investigated the place and location of testing venues, transportation, finding seats, air-conditioning and lighting of the venues and finally. Suitable time with regard to morning or evening administration, time delay, and the time needed to finish the test were the conditions considered for the timing factor.

The predominant view among the interview respondents was that the proctor/inspector factor was a sort of disappointment. During the focus group interview, one of the PhD students reacted: "Yeah, and about the physical conditions. Aa…the written test…I..aaa…witnessed…I I mean in one end in the classroom I was taking the test. I mean one person…I mean one guy… was just…eeer... making a lot of noise…Yeah, playing with the chair".

Another PhD student opined that "the proctors should not be students of university. They must be trained for exam administration. For example, in our room the announcer forgot to pronounce the time of second exam". Usually, the noises from exam proctors are one major problem introducing construct irrelevant variance into the observed scores. "Proctors need to keep silence in order to avoid distracting the examinees", said a PhD applicant. This unwanted experience happened to many of the participants, when taking the PhD exam. As one of them complained: "Yes, in my view, I think, it is strongly recommended that the proctors be morally silent".

Cheating in test venues had the pride of place on the list of dissatisfaction with the IPEET administration conditions. In his follow-up comments, one of the PhD students lamented: "Although some special measures like designing samples A. B, C have been taken to prevent cheaters from cheating, I believe more rigorous rules should be established in order to prevent them from cheating".

To complete the findings, PhD students were also solicited their responses through a twelve-item questionnaire on test administration condition. The items were divided based on three factors emerged in the analysis of transcribed interviews. The first factor was related to proctors. Three items fell under this category. As illustrated in Table 4 much to our surprise, some were satisfied with both the presence of exam proctors in the site (55%) as well as their appropriate behavior with PhD applicants (78%). However, 75% responded that cheating was possible at the exam site. Collectively, the questionnaire responses did not totally confirm what was revealed in interview transcription that “proctor” factor was problematic in IPEET test administration condition.

Table 4. PhD students' opinion with regard to the proctors/inspectors of the PhD entrance exam

Questionnaire statement	Group	N	Observed Prop.	Test Prop.	Sig. (2-tailed)
1. Were the inspectors/proctors available at the exam session?	Yes	57	.55	.50	.324
	No	46	.45		
	Total	103	1.00		
2. Did the exam proctors behave well?	Yes	80	.78	.50	.000
	No	23	.22		
	Total	103	1.00		
3. Was cheating possible at the exam?	Yes	77	.75	.50	.000
	No	26	.25		
	Total	103	1.00		

Another recurring theme that emerged from the responses of PhD applicants and PhD students in both focus group interview and comments section of the questionnaires was test venue (test location) condition. As mentioned before, the testing venue investigated transportation, finding seats, air-conditioning and lighting, and finally the place and site of testing. The majority of PhD applicants expressed disdain, claiming that the test venue was rife with problems big or small: lack of appropriate facilities, transportation problems, problems with finding seats, and the crowded site. "Regarding location (test venue) of the exam, I think, I mean it was a disaster. “One of the participants confided. Another added that "Transportation was the main concern to many candidates especially those who are living in towns far from the center of the province".

With regard to test administration questionnaire, five items were specified for the test venue. Table 5 summarizes PhD students' responses and presents binomial tests for significance. Results show that PhD students appeared almost unanimous in their perceptions that 'test venue' conditions such as information about the site (87%), transportation (77%), finding seats (78%), air conditioning systems (56%), and lighting (70%) were all appropriate in IPEET condition. Again like “proctors” factor, the results from questionnaires and interview data are not consistent.

Table 5. PhD students' opinion with regard to the testing venues of the PhD entrance exam of ELT

Questionnaire Statement	Group	N	Observed Prop.	Test Prop.	Sig. (2-tailed)
1. Were you appropriately informed about the site (place) of the test?	Yes	90	.87	.50	.000
	No	13	.13		
	Total	103	1.00		

Questionnaire Statement	Group	N	Observed Prop.	Test Prop.	Sig. (2-tailed)
2. Could you commute easily to the testing venue (site)?	Yes	79	.77	.50	.000
	No	24	.23		
	Total	103	1.00		
3. Could you find your seat easily?	Yes	80	.78	.50	.000
	No	23	.22		
	Total	103	1.00		
4. Were the testing venues well-ventilated?	Yes	58	.56	.50	.237
	No	45	.44		
	Total	103	1.00		
5. Was there enough light in the testing venues?	Yes	72	.70	.50	.000
	No	31	.30		
	Total	103	1.00		

Again like proctor factor questionnaire responses for 'test venue' were in opposition with PhD students' opinions in focus group interviews.

The third important factor emerged from comments and focus group responses were timing of the test. Most of the participants contented that one session (just morning) was better than two-sessions (morning and afternoon). They also commented that delay in test administration created some problems for them so that they were not able to show their full potential in the IPEET. "The multiple choice was not proper. Two times exam was boring", remarked a participant. Another commented: "In the two-session exam no facilities were provided for the break between the two sessions. We most found it frustrating". Another problem emerged from transcriptions was “time delay”. By way of illustration, one PhD applicant pointed out: "While the exam was supposed to start at 8, we actually started at 8:30".

Of the twelve items included in the test administration questionnaire, four were related to “timing” of the exam. As Table 6 indicates, most of the participants (75%) reported no problem for time limitation (100 minutes for 100 items) and 65% agreed with the one-session (morning) administration of the test. It is also demonstrated that of the total 103 participants 55% admitted that there was time delay in test administration. Thus, in timing factor quantitative and qualitative findings are virtually convergent, revealing inappropriate test condition. All in all, it can be argued that in terms of amount of time they had to finish the test, there was no problem, but in terms of delay and running the test in 2 sessions, test administration was found problematic by the participants.

Table 6. PhD students' opinion with regard to timing of the PhD entrance exam of ELT

Questionnaire statement	*Group*	*N*	*Observed Prop.*	*Test Prop.*	*Sig.(2-tailed)*
1. Was the time allocation for each sub-test appropriate?	Yes	77	.75	.50	.000
	No	26	.25		
	Total	103	1.00		
2. Do you prefer the exam to be administered in the morning?	Yes	67	.65	.50	.003
	No	36	.35		
	Total	103	1.00		
3. Do you prefer the exam to be administered in the evening?	Yes	36	.35	.50	.003
	No	67	.65		
	Total	103	1.00		
4. Was there any time delay in the administration?	Yes	46	.45	.50	.324
	No	57	.55		
	Total	103	1.00		

Though, the overall results are to some extent confusing and difficult to reconcile, with regard to inappropriate test administration conditions, qualitative responses are more substantially oriented toward *dissatisfaction* than those of questionnaires. However, since these findings reveal there are more than minimal CIV factors polluting test scores, this assumption which purports that test administration conditions introduce minimal CIV cannot be supported.

4.2 Generalization inference

To seek support for this inference and to answer the second research question, we analyzed the reliability of IPEET test score data. As information for individual test items was available, Cronbach alpha was the method of choice. An alpha level of .70 was set for acceptable reliability following a rule of thumb (Kline, 2000). As such, two sources of evidence were presented: Cronbach reliability index of total IPEET test, and insights from test score data and stakeholder's opinion. Each will be presented below.

Assumption 1. Acceptable internal consistency. As indicated in Table 7, Cronbach reliability for the total test is reported to be .873 which is beyond .7 as the rule of thumb criterion. However, when it comes to sub-tests, it is dramatically below .7 due to the low number of items in each section. For example, in discourse and socio, which is considered to be one subtest, this value is very low, considering that each has only 5 items.

Assumption 2. Sources of unreliability. Some factors such as the effect of testees, the structure of the test itself and the administration conditions of the

test may render a test unreliable (Farhady, Jafarpur & Birjandi, 2014). With regard to the present study, these sources of unreliability were influential. Given that a resounding number of PhD applicants with a wide range of abilities take part in IPEET, these differences may contribute to large variances and consequently, the reliability of the test may be overestimated. As regards the test content, findings from evaluation inference showed that the test is problematic both in terms of the difficulty level and with regard to gender DIF items; that is some source of unreliability are inevitable here. Findings from evaluation claim also showed that the administration conditions of the test (in terms of proctors, testing venue, and timing) were not appropriate. This factor can be regarded as another source of unreliability introduced into the context of IPEET.

Overall, it can be argued that, though a high Cronbach reliability is reported for total IPEET which is to some extent natural for every lengthy test of this kind to show this value, low reliability values for individual items together with insights from test score data and stakeholders' opinion, which revealed some sources of unreliability is good evidence to rebut the generalization inference.

Table 7. Reliability statistics for IPEET and its subtests

Type of Test	N of Items	N of participants	Cronbach's Alpha
Total test	100	999	.873
Teaching BA	15	999	.483
Linguistics BA	15	999	.640
Advanced Research	15	999	.640
Advanced Testing	15	999	.676
Teaching issues	30	999	.657
Discourse	5	999	.302
Sociolinguistics	5	999	.382

4.3 Extrapolation inference

To find a reasonable answer for the third research question, we sought evidence from questionnaires and interviews responded by experts (university professors) with regard to: a) the relevancy of content of the IPEET test tasks to the content of PhD credit courses exercised in universities and b) PhD students' success in PhD courses taught at PhD programs, results for each will be dealt with in turn.

Assumption 1. Relevance of IPEET test tasks to PhD courses. To seek support for this part and to answer the related research question, we solicited the

opinions of university teachers through telephone interview. Analysis of transcripts of telephone interview, showed, not surprisingly, that almost all university teachers appeared unanimous in the opinion that the IPEET tasks are partially commensurate to the objectives and requirements of PhD courses of TEFL, claiming that some of items included in the IPEET are beyond the PhD students' level of competence. In like fashion, their overall perception contends that the content of the test has not fully represented the target knowledge use domain of PhD courses. By way of illustration, one of the testing specialists as an associate university professor mentioned" some of the items, as far as I know, are in fact the ones not taught at MA level and some of them are underrepresented in the PhD courses……..so I think they are not perfect". Another professor confirms "actually I checked the questions one by one I found just around fifty percent of the questions are related to the courses in the PhD program". A fellow professor suggested "yeah, actually, I think to my understanding as I had a look at PhD Exam, actually I saw something like 40% reflection of PhD courses".

In order to triangulate the findings from qualitative responses, a self-assessment questionnaire addressing the relevance of IPEET test items to the content of ELT courses was also used. Due to the small number of university teachers taking part in the study, it was not logical to compute statistical analyses for the data. Therefore, only the frequency of responses is described here. The self-assessment questionnaires were comprised of a Likert scale of four choices: 1= *not at all*, 2=*slightly*, 3=*to some extent*, 4= *to a large extent*. It was shown that, on the whole, the clear majority disagreed with the total relevance of IPEET items or tasks with the PhD courses of TEFL. Put it simply, most of the participants selected the choices of not at all or slightly with regard to the relevance of the IPEET items to the target domain, thus corroborating the above perceptions from telephone interview.

To recapitulate, the findings from both interview and questionnaires do not reveal full correspondence between IPEET test tasks and target PhD courses in PhD programs. Likewise, the assumption that performance on IPEET test is fully related to specialized knowledge of PhD courses as target content use domain cannot be strongly supported.

Assumption 2. IPEET's prediction of success in PhD program. To answer the second part of the third research question which seeks to investigate how much success on PhD courses can be predicted based on PhD applicants' performance on the IPEET test, most of the professors took us by unpleasant surprise and confirmed very little chance of success on the part of PhD students. One of them lamented:

> I can count, actually, a number of instances that you see the students have perfectly performed on the test items in entrance examination but you see their performances, actually, are very weak in terms of orientation, in terms of applied linguistics, in terms of, actually, problematizing the

situation, in terms of applying their knowledge in developing scientific phases, I see I don't see, for example, over 50% chance of success on PhD courses.

Others claimed that PhD students, conscious of their language proficiency and content ability handicaps, take pains to make improvements in conducting their research projects. It was also shown that academic writing is a monolithic block ranking at the top among the gaps perceived in the PhD students' repertoire. As one of the participants pointed out, "In the past few years PhD students have been able to publish articles in journals which are not of that much quality and when it comes to their research activities, they are weak in this regard. All in all, I'm not fully satisfied". Another monolithic block perceived by university professors is the drudgery of dealing with students with hotchpotch abilities. As a professor with specialty in discourse remarked, "there are some students with mixed abilities who positively or negatively stand on extremes".

University professors were also asked to complete a questionnaire designed to gain insight on their opinions of the relative success of PhD students. As reported in Table 8, a great majority of participants contented that PhD students' success in PhD courses is a sort of disillusionment. Of the 20 university professors completing the questionnaire, 18(90%) reported that PhD students have problems with language proficiency, 17 (85%) expressed concerns with PhD students' abilities in terms of content courses, 18(90%) felt disappointment with their abilities in academic writing, and finally, 17(85%) contented that PhD students' have problems with basic principles of research. As such, these findings are totally convergent with the results of telephone interview, rebutting the assumption that "Performance on IPEET test predicts relative success of PhD students in PhD courses".

Table 8. Binomial test of university professors' opinion regarding the relative success of PhD students

Questionnaire statement	Group	N	Observed Prop.	Test Prop.	Sig. (2-tailed)
1. Most of the PhD students of TEFL do not have problems with language proficiency.	Disagree+	18	.90	.50	.000
	Agree*	2	.10		
	Total	20	1.00		
2. Most of the PhD students of TEFL do not have problems with the content of specialized courses like SLA, FLA, Discourse etc.	Disagree+	17	.85	.50	.003
	Agree*	3	.15		
	Total	20	1.00		
3. When writing a research paper, most of the PhD students of ELT do not have problems with principles of academic writing.	Disagree+	18	.90	.50	.000
	Agree*	2	.10		
	Total	20	1.00		
4. Most of the PhD students of TEFL do not have problems with basic principles of research.	Disagree+	17	.85	.50	.003
	Agree*	3	.15		
	Total	20	1.00		

* Combined 'Agree' and 'Strongly Agree' responses
+ Combined 'Disagree' and 'Strongly Disagree' responses
□Assumption of minimum 5 participants in each cell not met

4.4 Intermediate actions

To find answers to the fourth research question which tries to examine the inference of intermediate actions, we tracked evidence from questionnaires and interviews (responded by university professors and PhD students) to support the assumptions (collective decisions, full score descriptor, and systematic decisions) proposed for this inference. Each are explained below.

Assumption 1. Decisions based on a collective judgment. During the telephone interview with university professors, the clear majority opinion showed that the type of decisions made by top-tier decision makers is not based on a collective judgment, complaining that they are not informed of any type of decisions made. The respondents' opinions were largely negative with regard to the quality of decisions: "We are not aware of the type of decision. We don't know anything about how they decide…", said one of the participants. Another lamented"…Decisions are not based on a collective judgment, yea it's a matter in Iran that everything is a topsy turvy, they don't have a strict policy. Even if they ask our opinion, they will never act accordingly to what we have told them to do". Still another university professor mentioned: "I say these judgments are based on a collective biases because they try to decide on the content of the items without receiving any judgment from outer circle".

As reported in Table 9, stakeholders' responses to questionnaires also confirmed the qualitative findings. A clear majority of university professors (80%) and about half of the PhD students (58%) contended that policy makers' decisions are not based a collective judgment.

Table 9. Binomial test of stakeholders' opinion regarding collective decisions

	Questionnaire statement	Group	N	Observed proportion	Test proportion	Sig. (2-tailed)
University professors	Policy makers' admission decisions are based on a collective judgment (professors are included)	agree*	4	.20	.50	.012□
		disagree+	16	.80		
		total	20	1.00		
PhD students	Policy makers' admission decisions are based on a collective judgment(PhD students are included)	agree*	45	.44	.50	.237
		disagree+	58	.56		
		total	103	1.00		

* Combined 'Agree' and 'Strongly Agree' responses
+ Combined 'Disagree' and 'Strongly Disagree' responses
□Assumption of minimum 5 participants in each cell not met

Assumption 2. Full representation of score descriptors. Most of the university professors lamented on the lack of presenting a detailed and full report card with regard to score descriptors: "the report card should be based on the multidimensionality here, for the candidates, representative scores based on the performance on different sections of the test should be provided. Their reporting is one-dimensional giving a total score", disdained one of the respondents. All in all the results of the qualitative evidence indicates that the quality of the decisions made by policy makers is inappropriate.

Results from questionnaire data (see Table 10) also confirmed what was concluded in the qualitative section. About 65% of university professors and half of PhD students (51%) disagreed with the way policy makers report test scores.

Table 10. Binomial test of stakeholders' opinion regarding score descriptors

	Questionnaire statement	Group	N	Observed proportion	Test proportion	Sig. (2-tailed)
University professors	Policy makers report and present test scores and score descriptors in ways that are clear and fully representative to the test takers	agree*	7	.35	.50	.263
		disagree+	13	.65		
		total	20	1.00		
PhD students	Policy makers report and present test scores and score descriptors in ways that are clear and fully representative to the test takers	agree*	50	.49	.50	.844
		disagree+	53	.51		
		total	103	1.00		

* Combined ‘Agree’ and ‘Strongly Agree’ responses
+ Combined ‘Disagree’ and ‘Strongly Disagree’ responses
Assumption of minimum 5 participants in each cell not met

Assumption 3. Systematicity of the decisions. As regards the systenmaticity of decisions, university professors expressed that it’s not standard: One of the participants confirmed:

> Yea, I suppose one general problem in the policies made by these two responsible agencies is that the decisions with regard to the percentages assigned to the written form of the exam and the oral form of the exam differ quite unsystematically. For example, for one year 70 percent of the total evaluation is accounted for by the oral test, the next year it’s the other way round. I suppose this actually creates confusion for PhD candidates and universities and it’s not logical.

PhD students also expressed disdain with the systematicity of the decisions made by policy makers. In fact they ignore the ideas and opinions of university

professors and PhD students as major stakeholders. "It is totally a top down decision I think. They do not take into consideration the ideas of the interviewees, participants; sometimes they assign 70% for written exam, sometimes 50% and sometimes 30%. They are not stable…", proposed one of the participants.

As reported in table 11, questionnaire results indicates that university professors (86%) and PhD students (58%) suggested that decisions made by policy makers is not fully based on a collective judgment. As such, the quantitative findings confirm the results from qualitative interview, hence rejecting all the assumptions articulated for the inference of 'intermediate actions' as taken by policy makers.

Table 11. Binomial test of stakeholders'' opinion regarding systematic decisions

	Questionnaire statement	Group	N	Observed proportion	Test proportion	Sig. (2-tailed)
University professors	Decisions made by Policy makers are systematic (do not change from one year to another).	agree*	2	.14	.50	.013□
		disagree+	18	.86		
		total	20	1.00		
PhD students	Decisions made by Policy makers are systematic (do not change from one year to another).	agree*	43	.42	.50	.114
		disagree+	60	.58		
		total	103	1.00		

* Combined 'Agree' and 'Strongly Agree' responses
+ Combined 'Disagree' and 'Strongly Disagree' responses
□Assumption of minimum 5 participants in each cell not met

4.5 Ultimate effects

To track evidence for the fifth research question and inference, we analyzed the results from questionnaire and interview data provided by university professors. The first part of this inference rested on the assumption that the use of the test helps promote good instructional practice and the second part assumed that IPEET predicts success for PhD students in PhD courses, each are dealt with below:

Assumption 1. Effects on instructional practice. During the course of telephone interview, most of the university professors, having some MA courses with MA students (in addition to running some PhD courses of TEFL), took us by unpleasant surprise, acknowledging that the use of the IPEET test did not help promote good instructional practice in those courses, opining that they continued with their own conventional way of instruction. "no as far as I'm concerned, it has no effect on the way I teach. I myself regardless of the type of the exam, we do our own teaching", stated one of the professors. This opinion

was also confirmed by another participant: "this MC exam of PhD cannot have any sort of contributions for promoting my instructional practice [at MA level]".

When responded to the individual questionnaire item investigating the washback effect of using IPEET (see Table 12), a great majority of university professors disagreed with the promotion of a good instructional practice (70%), in ELT courses, specifically the MA ones.

Table 12. Binomial test of university professors' opinion regarding washback

Questionnaire statement	Group	N	Observed Prop.	Test Prop.	Sig. (2-tailed)
The use of the PEEE test helps promote good instructional practice in instructional settings such as MA courses.	Agree*	6	.30	.50	.115
	Disagree+	14	.70		
	Total	20	1.00		

* Combined 'Agree' and 'Strongly Agree' responses
+ Combined 'Disagree' and 'Strongly Disagree' responses

Assumption2. Effects on relative success of PhD students. It was hypothesized that if IPEET was an appropriate instrument, then PhD applicants who are screened through this instrument to enter the PhD courses may have the relative abilities to fulfill the requirements of PhD courses run at PhD programs. This was also hypothesized to be solicited via experts'(university professors) opinion. Since evidence for this part was also sought for the second assumption of extrapolation inference (see the assumption 6.3.2), and the results indicated dissatisfaction with their relative ability, it would be redundant to present the results in this regard. As such the same conclusion holds true for this part of 'ultimate action'. That is, the assumption on the relative effects on PhD students can be rebutted as well. On the whole, residing on the results presented for the two assumptions of 'ultimate effects', this inference is strongly rebutted.

Proposed Ultimate Actions

Qualitative and quantitative analyses of university professors' and PhD students' suggestions for the betterment of content and decision quality of IPEET as 'ultimate actions' were analyzed to support the two assumptions prespecified for the inference of 'ultimate actions' .

Assumption 1. University professors' suggestions for the betterment of IPEET. First, university professors were solicited their opinions via the telephone interview. Analysis of results gave birth to six general themes such as 'application of knowledge', 'relatedness of MA courses', 'specialized interest', 'significant role for IPEET', 'collective development of questions', 'academic writing', and finally 'overall change'. Not surprisingly, some university professors argue that some of the questions included in IPEET are disappointing. "The IPEET taps into the memorized knowledge of candidates (90%). Test items should test candidates in terms of creativity, in terms of

application of knowledge", said a participant. Moreover, among the list of suggestions proposed for the improvement is including essay type questions in IPEET test. By way of example, one of the participants pointed out: "If we have an essay type exam instead of MC, then well… we would have better candidates"- a statement which was corroborated by another participant: "questions should change from recognition to production one that is essay-type would be better to select qualified candidates". That is because essay type questions reflect a better picture of PhD applicants' competence.

Another recurring theme perceived by participants was academic writing. They complained that this important skill has been neglected in the content of IPEET. One of the participants mentioned: "I suppose if ,for example, one or two essay-type questions could also be added to the MC to check students writing ability , esp. academic writing, this would again let the final decision be made more logical".

The third common theme extracted from the interview data was the "significant role of IPEET" in the admission process. Most of the participants claim that IPEET test should be given more weight compared with local interviews. "that exam should have a significant role in admission, for example, 50% of the final admission decision should be based on the results of this test", proposed one of the participants.

As another theme observed in the interview data, "collective development of IPEET items" occupied the pride of place on the list of suggestions. Being disdained with the invisibility of their voices in the content of IPEET, university professors proposed unanimously that different professors from center and periphery universities should have an equal hand in the development of IPEET test items. "It would be better if more actually, universities, let's say, are involved in test development but now it is not the case", suggested one of the participants.

When taking IPEET, PhD applicants should be required to select their specialized field of interest (testing, research, teaching, discourse, and so on). This was a common suggestion observed in the perceptions of the participants. One of them expressed: "We should select the, I mean, candidates based on the interest. We should construct our test according to the capacities and interest of candidates and those who wish to participate in our program should know in advance that our department is discourse oriented".

As a common theme elicited from the respondents, the "overall change" of the IPEET was substantially suggested. University professors contented that this test should be reshuffled.

As such it can be claimed that the current procedure is far from being perfect but it still needs improvement. An associate professor in language testing approved: "so if the content quality of the test goes up, this would bring the decision making to a sort of even-handedness and justice".

University professors were also asked to fill out questionnaire items intended to gather their opinion regarding the suggestions for the improvement of IPEET. Descriptions of Binomial tests are presented for each group of stake holders. Table 13 summarizes university professors" responses and reports binomial tests for significance.

Results indicate that university professors are almost unanimous in their views that the present content and the current policy of IPEET should be changed or improved. Of the 20 university professors completing the questionnaire, 19 (95%) agreed that academic writing should be added to the content of IPEET, 19(95%) contended that the items should be based on MA courses, 16(80%) suggested that items should be task-based, 19(95%) proposed that the items should test students' application of knowledge, 16(80%) acknowledged that the items should not be designed based on the content of BA courses, 18(90%) confirmed that the scores from IPEET should be given more weight, as compared with those of PhD interview, 16(80%) admitted that all professors should have an equal hand in developing the questions, 19(95%) welcomed that the specialized interest of test takers should be taken into account, and finally 18(90%) were satisfied with the overall change and improvement of the content of IPEET. As regards the decision quality (including items 11-15), of 20 participants 17 (85%) agreed with the decisions to be systematic, while 16 (80%) contended for both the collective decisions and qualitative articles. As for the opinions with regard to centralized system, this number was lowered to 7(35%). And finally 15 participants (75%) agreed that the current policy should be changed.

Table 13. Binomial test of university professors' suggestions for ultimate actions taken to change the content of IPEET

Questionnaire statement	Group	N	Observed Prop.	Test Prop.	sig. (2-tailed)
1. As a sub-test, academic writing should be added to the content of PhD Entrance Exam of ELT	agree*	19	.95	.50	.000
	disagree+	1	.15		
	Total	20	1.00		
2. The questions included in PhD Entrance Exam of ELT should be based on the content of MA syllabi or MA courses.	agree*	19	.95	.50	.000
	disagree+	1	.15		
	Total	20	1.00		
3. The items included in the PhD Entrance Exam of ELT should be task- based.	agree	16	.80	.50	.012
	disagree+	4	.20		
	Total	20	1.00		
4. The items included in PhD Entrance Exam of ELT should test PhD applicants' application of knowledge rather than their memorized knowledge	agree*	19	.95	.50	.000
	disagree+	1	.15		
	Total	20	1.00		
5. The questions related to some courses at BA	Agree*	16	.80	.50	.012

level are not logical to be included in PhD Entrance Exam of ELT.	Disagree+	4	.20		
	Total	20	1.00		
6. The PhD Entrance Exam of ELT should have a more significant role in admission process than the oral interview	Agree*	18	.90	.50	.000□
	Disagree+	2	.10		
	Total	20			
7. part of IPEET scores should be allocated to research background	Agree*	17	.85	.50	.003□
	Disagree+	3	.15		
	Total	20	1.00		
8. Academically celebrated professors from both center and periphery universities should have an equal hand in developing questions for PhD Entrance Exam of ELT	Agree*	16	.80	.50	.012□
	Disagree+	4	.20		
	Total	20	1.00		
9. At the top of the answer sheet distributed among PhD applicants should be a short check list requiring PhD applicants to tick their field-specific interest	Agree*	19	.95	.50	.000□
	Disagree+	1	.15		
	Total	20	1.00		
10. Overall, the current PhD Entrance Exam of ELT should be changed or improved.	Agree	18	.90	.50	.000□
	Disagree+	2	.10		
	Total	20	1.00		
11. The decisions made by policy makers should be systematic (should not change from one year to another	Agree*	17	.85	**.50**	.003□
	Disagree+	3	.15		
	Total	20	1.00		
12. Decisions made by policy makers should be based on a collective judgment	Agree*	16	.80	.50	.012□
	Disagree+	4	.20		
	Total	20	1.00		
13. More attention should be paid to quality rather than the quantity of articles as research backgrounds.	Agree*	16	.80	.50	.012□
	Disagree+	4	20		
	Total	20	1.00		
14 The evaluation of PhD applicants' educational background should be based on a centralized system	Agree*	7	.35	.50	.263
	Disagree+	13	.65		
	Total	20	1.00		
15. The current policy of IPEET should be changed or improved.	Agree*	15	.75	.50	.041
	Disagree+	5	.25		
	Total	20	1.00		

* Combined 'Agree' and 'Strongly Agree' responses
\+ Combined 'Disagree' and 'Strongly Disagree' responses
□Assumption of minimum 5 participants in each cell not met

Assumption 2. PhD Students' suggestions for the betterment of IPEET. Analysis of focus group interviews with PhD students gave rise to five important themes such as 'significant role for IPEET', 'application of knowledge', Students' specialized interest', 'questions based on MA courses', and 'the overall change' of IPEET. 'Academic writing' and 'collective development of questions' deemed significant in telephone interviews were not considered important in the perceptions of PhD students.

In the views of PhD students, weight assigned to IPEET is problematic and the scores from this test should be given more weight as compared with those of PhD interview. "The weight… So there should be given a more standardized weight for example, 70 percent to multiple choice, and 30 you mean", remarked one of the participants.

Measuring test takers' "application of knowledge" was another important category receiving substantial value in the eyes of PhD students. They don't think this test tap into their target content abilities as practiced in PhD courses. One of them purported: "Yes. I don't believe in multiple-choice question, I believe in essay exam". Another participant confessed that "Knowledge is not enough! A good exam should assess the potentiality of applicants to be PHD candidate".

PhD students also believed that they should be introduced to the universities based on their specialized field of interest (e.g., testing, research, teaching, discourse, and so on). This was a common suggestion by the participants. "We should construct our test according to the capacities and interest of candidates and those who wish to participate in our program should know in advance that our department is, for instance, discourse oriented" confided one of the participants.

PhD students also claimed that the content of IPEET test items should be based on the content of MA courses or MA syllabi. In this way, they believed, the validity of this test would be enhanced. "It should be drawn from the books that we have read in M.A, but not in the book that we will read next", suggested one of the participants.

Finally, the unanimous perceptions of PhD students were that the content of IPEET should be revised and improved. As one of the participants argued "the content should be reshuffled" - a suggestion corroborated by another PhD student: "I think the questions are not designed properly and you can find some faults in them".

Looking to the responses of the PhD students (see Table 14), they suggested almost the same rate for the revision and improvement of the content of IPEET as opined in the qualitative section. The observed proportion of the responses for the test improvement ranged from 66% to 98%. Some participants were not fully satisfied with discarding the items related to BA courses (66%). However, this number was reported to be 80% for university professors, suggesting that the items included in IPEET should not be based on the content of BA courses. However, responses of PhD students with regard to the decision quality (see items 11 to 15) shows that the observed proportion ranges from 51% to 62%, indicating that PhD students suggest a minor revision for the decision quality (as compared with the responses from university professors), but they propose a substantial change for the content quality of this test.

Table 14. Binomial test of PhD students' suggestions for ultimate actions taken to change the content and decision quality of IPEET

Questionnaire statement	Group	N	Observed Prop.	Test Prop.	Sig. (2-tailed)
1. As a sub-test, academic writing should be added to the content of PhD Entrance Exam of ELT	agree*	79	.77	.50	.000
	disagree+	24	.23		
	Total	103	1.00		
2. The questions included in PhD Entrance Exam of ELT should be based on the content of MA courses.	agree*	92	.89	.50	.000
	disagree+	11	.11		
	Total	103	1.00		
3. The items included in the PhD Entrance Exam of ELT should be task-based.	agree*	89	.86	.50	.000
	disagree+	14	.14		
	Total	103	1.00		
4. The items included in PhD Entrance Exam of ELT should test PhD applicants' application of knowledge rather than their memorized knowledge	agree*	98	.95	.50	.000
	disagree+	5	.05		
	Total	103	1.00		
5. The questions related to some courses at BA level are not logical to be included in PhD Entrance Exam of ELT.	Agree*	35	.66	.50	.001
	Disagree+		.34		
	Total	103	1.00		
6. The PhD Entrance Exam of ELT should have a more significant role in admission process than the oral interview	Agree*	68	.81	.50	.000
	Disagree+		.19		
	Total	103	1.00		
7.Part of IPEET scores should be allocated to research background	Agree*	101	.98	.50	.000
	Disagree+	2	.02		
	Total	103	1.00		
8. Academically celebrated professors from both center and periphery universities should have an equal hand in developing questions for PhD Entrance Exam of ELT	Agree*	90	.87	.50	.000
	Disagree+	13	.13		
	Total	103	1.00		
9. At the top of the answer sheet distributed among PhD applicants should be a short check list requiring PhD applicants to tick their field-specific interest	Agree*	94	.91	.50	.000
	Disagree+	9	.09		
	Total	103	1.00		
10. Overall, the current content of PhD Entrance Exam of ELT should be changed or improved.	Agree*	95	.92	.50	.000
	Disagree+	8	.08		
	Total	103	1.00		
11. The decisions made by policy makers should be systematic (should not change from one year to another	Agree*	58	.56	.50	.237
	Disagree+	45	.44		
	Total	103	1.00		
12. Decisions made by policy makers should be based on a collective judgment	Agree*	60	.58	.50	.114
	Disagree+	43	.42		
	Total	103	1.00		
13. More attention should be paid to	Agree*	53	.51	.50	.844

quality rather than the quantity of articles as research backgrounds.	Disagree+	50	.49		
	Total	103	1.00		
14 14.The evaluation of PhD applicants' educational background should be based on a centralized system	Agree*	64	.62	.50	.018
	Disagree+	39	.38		
	Total	103	1.00		
15. The current policy of PEEE should be changed or improved.	Agree*	39	.62	.50	.018
	Disagree+	64	.38		
	Total	103	1.00		

* Combined 'Agree' and 'Strongly Agree' responses
+ Combined 'Disagree' and 'Strongly Disagree' responses
Assumption of minimum 5 participants in each cell not met

Thus, it is axiomatic from the results reported for both telephone and focus group interviews as well as from observed proportions demonstrated for questionnaire responses that university professors and PhD students are almost unanimous in their perceptions that IPEET test needs substantial revision and a radical change for this test is urgent. All in all, it can be argued that the proposed assumptions for the inference of ultimate action are somehow rebutted.

5. Discussion and Conclusion

The present study aimed to investigate the content of IPEET in light of argument- based validity and theory of action. The overall results of the present study demonstrate that the IPEET test instrument suffers from validity requirement. Findings from the present study provided evidence of the rejection of all the proposed claims.

With regard to evaluation inference, the LR results showed negligible DIF items; however, before jumping to any conclusion we should caution that in this study the size of the reference group (602 females) was almost twice as much as the size for the focal group (397 males). This might pollute the validity of DIF interpretation; therefore the degree of certainty in a strong conclusion is limited in this regard.

However, the aggregate findings from both interview and questionnaire analyses with regard to the difficulty level of the items as well as the adequacy of the number of those items strongly showed that more than minimal CIV was introduced into the test scores, hence a possible evidence to rebut the assumption of test characteristics. According to Johnson and Riazi (2013), little concern can be detected in the literature regarding standardized instruments in terms of test characteristics. The findings of their study (on test characteristics) on an English placement system confirmed much concern for the writing sample subtest but not for the Accuplacer Companion (AC). Thus, with regard to test characteristics, the findings of the present study are somehow consistent with Johnson and Riazi's study at least in their writing sample investigation.

Further to test characteristics, mixed method results revealed inappropriate test administration conditions for IPEET. One possible explanation can reside on Xi (2010) and Kunnan (2000, 2003), arguing that inconsistent test administration, lack of accommodation for test takers with disabilities and raters' bias are among the factors that, may act as the construct-irrelevant variances and render the test invalid. As regards the timing issue, participants were not satisfied, claiming that morning session is more appropriate. This finding was convergent with research in literature (Monk, 1990; Wise, Kingsbury, Hauser, & Ma, 2010).

As regards the proctors' issue, a great majority of focus group participants were not satisfied with the test proctors of IPEET.

Concerning, testing venue, the qualitative and quantitative findings are difficult to reconcile. Findings from focus group data showed partially appropriate conditions. However, the questionnaire results indicated that the majority of PhD students (about 70%) considered 'test venue' conditions appropriate. On the whole, the collective findings indicated that test performance is affected by minimal CIV, being in line with findings from Shulman, Boster, & Carpenter (2011) and Douglas (2014) in which they argued that if testing venue is inappropriate, test takers may not perform to the best of their abilities.

The second research question formulated in this study sought to verify the degree to which Cronbach alpha coefficient is .7 or higher for IPEET test. Results indicated that Cronbach reliability for the total test was reported to be .873 which is beyond .7 as the rule of thumb criterion. However, when it came to subtests, the reliability estimates dramatically decreased to below .70. This finding is in keeping with Johnson and Riazi (2013) who found a high reliability value for Accuplacer test used to place non-native candidates in appropriate instructional courses. However, any tentative conclusion with regard to high reliability estimates is unwarranted and literature refers to reliability criteria as insufficient (Weir, 2005) or even worthless (Bachman, 1990; Wood, 1993).

Possible explanations for the constraints on reliability estimates can be found in the words of Sawilovsky (2000) who proclaimed: "Statements about the reliability of a certain test must be accompanied by an explanation of what type of reliability was estimated, how it was calculated, and under what conditions or for which sample characteristics the result was obtained" (p.159). This explanation refers to reliability as an estimate which is mostly sample dependent; that is, reliability is not a feature of the test itself but a characteristic of the population who sit the test. One other line of explanation for the limitations of reliability which is somehow related to the first one, is put forward by Weir (2005), arguing that "candidates of widely ranging ability are easier to rank reliably, and so will produce higher reliability indices than groups that are more equal in level where all the scores tend to bunch together [lower standard deviation and lower variance]" (p, 32). In line with this argument, it

can be concluded that since high-stakes tests like the present PhD exams in Iran are sat by a large number of candidates with a wide range of abilities, one can expect higher reliability indices to be reported for these tests. This can be one reason for the high reliability value (r=.87) reported for the present IPEET.

With regard to subtests, the reliability indices were depressed for IPEET (see Table 5). One reason may rest on the fact that Cronbach alpha values are quite sensitive to the number of items in the scale (e.g, Huges, 2003; Weir, 2005) so that with short scales (e.g. scales with fewer than ten items), it is common to find quite low Cronbach values (e.g. 5). As shown in Table 5, we saw that for the total test a reliability value of .87 was observed. However, when it came to advanced testing subtest which consists of 15 items, a reliability index of .68 was reported. Still, when the number of items decreased to as few as 5, as it was the case for discourse, then the reliability estimate decreased to .30 as well. Any tentative interpretation may be difficult here. However, resting on the reason that all the subtests included in IPEET measure different constructs, meaning they are unrelated to each other, a high reliability value reported for the total test which includes all these subtests may not be a good source for the desired reliability of the test. Likewise, with these contradictory information rebutting the generalization claim based on internal consistency may not seem to be logical.

However, modern theories of validity, further to considering statistical analyses, suggest investigating the sources of unreliability. It is claimed that one source of unreliability might be the content of the test itself. With regard to the present study, as it was revealed in the evaluation inference, the content was somehow biased, as it showed items flagged with gender DIF as well as those being displayed as very difficult. Moreover, the test administration conditions were reported to be problematic. These factors are to some extent, in contradiction with the high reliability value reported for the total PEEE test.

Thus, with these types of evidence, though not that much forceful, one can be inclined to rebut the generalization inference. However, more investigations are warranted to clarify this somehow dark area.

The third research question aimed to scrutinize the extrapolation inference. With regard to relevance of IPEET test content and its predictive power for success, mixed method data provided by experts revealed that IPEET was neither related to target content domain nor did it fully predict success for PhD students. Findings for the representation of the content of IPEET as revealed in the present study are consistent with what Kane purports that “expert evaluations of test items do not generally provide strong support for extrapolation to the target domain” (Kane, 2006, p. 57). As regards IPEET context, one line of explanation is that this test consists of some subtests with a limited number of items not being enough to measure the full potential of PhD students' ability. Moreover, the weak prediction of power attributed to this test may be associated with the fact that PhD programs are more research-based, an

area on which PhD candidates have not had ample opportunities to work neither at their BA nor MA levels. Even good performance on the IPEET test has not, at the very least guaranteed success for PhD students to fulfill the requirements of PhD courses. Moreover, PhD students admitted to Iranian PhD programs are expected to be able to apply their subject area knowledge. As such, the instruments through which these applicants are screened should test their production and application of this knowledge. But we see it is not the case for IPEET in Iran.

To recapitulate, it seems that the content of IPEET has not been carefully and critically examined. And this may be one of the reasons for the test to prone to unintended consequences. A claim corroborated by Haertel (2013) arguing that careful attention should be paid to test content; otherwise, intended positive effects are not realized and unintended effects would not be avoided. All in all, the overall findings of the first part of this study (which is based on Kane's framework) were not consistent with what Haertel emphasized, demonstrating that the content of IPEET is problematic and suffers from validity requirement. Except for the generalization claim which was supported via reliability coefficient, the evaluation and extrapolation claims as articulated in the present study were rejected.

With regard to the fourth research question which sought to examine the inference of intermediate action, stakeholders proposed that some actions should have been taken by policy makers to improve the content and decision quality of IPEET, the most important of which were policy makers' responsibility to make systematic decisions, reporting score descriptors which are clear, understandable and representative, and finally basing the admission decisions on a collective opinion of different stakeholders. Mixed method data showed it was not the case for IPEET. These findings indicate that types of decisions made by responsible agencies are a sort of hasty ones, ignoring the ideas of all stakeholders and experts. What can be concluded from the overall perceptions of participants was that the present PhD exam and the decisions made on it is a sort of trial and error program. As such these hasty and unsystematic decisions may eventuate in unintended consequences, letting some unqualified PhD students enter the PhD programs or some qualified ones fail entering the program.

The fifth research question or inference articulated in the theory of action argument sought to investigate the ultimate effects happening as a result of the use of IPEET test in Iran. It was assumed that this test has a positive washback and predicts success for PhD students. Findings showed that neither this test promoted good instructional practice, nor did it predict success for PhD students, leading to the rejection of the assumptions. One possible explanation for the negative washback is that PhD courses in Iran cover major areas and topics in applied linguistics whereas this MC test, pseudonymed as IPEET, mostly measures a limited range of abilities, suffering from both construct-

irrelevant variance and construct under-representation as two plagues bothering test validity. Therefore it is an inappropriate instrument to screen test takers for PhD programs.

As mentioned before, the last research question articulated for this study sought to examine stakeholders' suggestions for the ultimate actions proposed to be taken to redress the unintended consequences as revealed in the present study. Some of these suggestions include: adding academic writing to the content of IPEET, collective development of the items by applied linguistics experts from different universities, IPEET test measuring test takers' application of knowledge, developing items based on applicants' specialized interest, and discarding BA items from the content of IPEET. These recommendations are symptomatic of some problems with the present content and the current policy of PhD evaluation in Iran, since this system is supposed to be more research-based and students should be able to demonstrate their abilities not only in completing a doctoral dissertation but in writing some high quality papers. But we see the instrument (the entrance exam) through which PhD applicants are screened is unrepresentative and inappropriate. As such, top-tier decision makers should make a radical change in this regard. We hope that these possible suggestions as ultimate actions would contribute to the betterment of PhD entrance exam in general and the PhD entrance exam of ELT in Iran in particular.

References

Ary, D., Jacobs, L. C. & Sorensen, C. (2010). *Introduction to research in education* (8th Ed.). New York, NY: Wadsworth.

Azmoon.Net. (2014). PhD entrance examination news. Retrieved 2014, October, 15th from www. Phd.Azmoon.Net. www. PhD Test.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives,* 8, 70-91.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (sixth Ed.) London: Routledge.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? New directions for testing and measurement: Measuring achievement over a decade. In *Proceedings of the 1979 ETS Invitational Conference* (pp. 99-108). San Francisco, CA: Jossey- Bass.

Douglas, D. (2014). *Understanding language testing*. Oxon.Hodder Education.

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

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *The Journal of Applied Psychology, 72,* 19–29.

Farhady, H., Jafarpur, A. J., & Birjandi, P. (2014). *Testing Language Skills from Theory to Practice.* Tehran: SAMT.

Glaser, B. G., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. New York: Aldine.

Green, A. (2007). Washback to the learners: Learners and teacher perspectives on IELTS preparation course expectation and outcomes. *Assessing Writing*, *11,* 113 -134.

Haertel, E. (2013). How is testing supposed to improve schooling? *Measurement: Interdisciplinary Research and Perspectives, 11*(1-2), 1-18.

Johnson, R.C., & Riazi, M. (2013). Assessing the assessments: Using an argument-based validity framework to assess the validity and use of an English placement system in a foreign language context. *Papers in Language Testing and Assessment. 2(1),* 31-58.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527-535.

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

Kane, M. T. (2006). Validation. *Educational Measurement, 4*, 17-64.

Kane, M.T. (2011). Validating score interpretations and uses. *Language Testing 29(1)*, 3– 17.

Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement. 50(1)*, 1–73

Kiany, R., Shayestefar, P., Ghafar Samar, R., Akbari, R. (2013). High-rank stakeholders' perspectives on high- stakes University entrance examinations reform: priorities and problems. *High Educ 65*, 325–340

Kline, P. (2000). *The handbook of psychological testing* (2nd Ed.). London: Routledge.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan, (Ed.). *Fairness and Validation in Language Assessment*: *Selected Papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 1-14). Cambridge: Cambridge University Press.

Kunnan, A. J. (2003). Test fairness. In M. Milanovic & C. Weir (Eds.), *Select Papers from the European Year of Languages Conference, Barcelona*. Cambridge: Cambridge University Press.

Maxwell, J. A. (1996). Qualitative Research Design: An Interactive Approach. Thousand Oaks, California: Sage Publications.

Monk, T H. (1990). The relationship of chronobiology to sleep schedules and performance demands. *Work and Stress, 4*(3), 227-236.

NOET. (2013). PhD entrance examination news. Retrieved 2013, December, 20th from http://www.eao.ir/eao/Full Story.aspx? gid=1&id=730

Shulman, H C., Boster, F J., & Carpenter, C J. (2011). *Do data collection procedures influence political knowledge test performance?* Paper presented at the annual meeting of the Midwestern Political Science Association in Chicago, IL. Oaks, CA: Sage.

Sireci, S.G., & Rios, J.A. (2013). Decisions that make a difference in detecting differential item Functioning. *Educational Research and Evaluation, 19*, 170–187. DOI: 10.1080/13803611.2013.767621.

Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing, 17*, 323–40.

Teddlie, C. & Tashakkori, A. (2003).Major Issues and Controversies in the Use of Mixed Methods in the Social and Behavioral Sciences. In Tashakkori, A. & Teddlie, C. *Handbook of mixed methods in social and behavioral research*. Thousand Oaks, CA: Sage.

Teddlie, Ch. & Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Research in Schools, 13 (1),* 12-28.

Weir, C. J. (2005).Language testing and validation. Hampshire: Palgrave McMillan.

Wise, S L., Kingsbury, G., Hauser, C., & Ma, L. (2010). *An investigation of the relationship between time of testing and test-taking effort*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147- 170.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense

Zumbo, B. D. (2008, July). Statistical methods for investigating item bias in self-report measures. *Florence Lectures on DIF and Item Bias*. Lectures Conducted from Universita degli Studi di Firenze, Florence, Italy.

APPENDICES

Appendix A: PhD Students' Questionnaire

Thank you for taking the time to complete this questionnaire. You are helping us improve the quality of Iranian PhD Entrance Exam of TEFL (IPEET). This questionnaire is designed for a PhD dissertation. Completing this questionnaire is completely voluntary and all possible measures will be taken to ensure the confidentiality of your personal information.

A1. Background Information: Please tick the appropriate answer.

1. I am a male ○ I am a female ○
2. I am 25-27 ○ 28-30 ○ 30-39 ○ more than 40 + years old ○
3. This is the first ○ second ○ third ○ fourth ○ time I take this exam. ○
4. My total score in specialized subtests.

less than 30% ○ 30 to 40% ○ 40 to 50% ○ more than 50% ○

5. My total score in general English.

less than 30% ○ 30 to 40% ○ 40 to 50% ○ more than 50% ○

A2: PhD students' questionnaire regarding administration conditions and characteristics of PhD Entrance Exam of ELT

Please evaluate the following items based on a 3-point Likert scale. (The purpose of this part is to investigate the quality of administration procedure for PhD Entrance Exam of TEFL)

6. Were the inspectors/proctors available at the exam session?

Yes ○ No ○ No Idea ○

7. Did the exam proctors behave well?

Yes ○ No ○ No Idea ○

8. Was cheating possible at the exam?

Yes ○ No ○ No Idea ○

9. Were you appropriately informed about the site (place) of the test?

Yes ○ No ○ No Idea ○

10. Could you commute easily to the testing venue (site)?

Yes ○ No ○ No Idea ○

11. Could you find your seat easily?

Yes ○ No ○ No Idea ○

12. Were the testing venues well-ventilated?

Yes ○ No ○ No Idea ○

13. Was there enough light in the testing venues?

Yes ○ No ○ No Idea ○

14 . Was the time allocation for each sub-test appropriate?

Yes ○ No ○ No Idea ○

15. Do you prefer the exam to be administered in the morning?

Yes ○ No ○ No Idea ○

16. Do you prefer the exam to be administered in the evening?

Yes ○ No ○ No Idea ○

17. Was there any time delay in the administration?

Yes ○ No ○ No Idea ○

A3. Comments

Please write any additional comments you would like to make about the improvement of administration procedure for PhD Entrance Exam of ELT.

A4: Please rank the different components of the PhD Entrance Exam of ELT (version 93) according to how difficult you found them.

	Very difficult	Difficult	Average	Easy	Very Easy
18. The overall test	1 ○	2 ○	3 ○	4 ○	5 ○
19. Linguistics (BA)	1 ○	2 ○	3 ○	4 ○	5 ○
20. Teaching (BA)	1 ○	2 ○	3 ○	4 ○	5 ○
21. Teaching (MA)	1 ○	2 ○	3 ○	4 ○	5 ○
22. Testing (MA)	1 ○	2 ○	3 ○	4 ○	5 ○
23. Research (MA)	1 ○	2 ○	3 ○	4 ○	5 ○
24. Socio and Discourse (MA)	1 ○	2 ○	3 ○	4 ○	5 ○

A5: PhD students' questionnaire regarding the quality of decisions made by policy makers

C1: Please evaluate the following items based on a 5-point Likert scale of agreement. If you have no ideas select undecided

25. Policy makers' admission decisions are based on a collective judgment.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

26. Policy makers report and present test scores and score descriptors in ways that are understandable to test takers

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

27. Decisions made by Policy makers are systematic (do not change from one year to another).

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

A6. Comments

Please write any additional comments you would like to make about the quality of decisions made by policy makers on the acceptance or non-acceptance of PhD applicants.

A7. PhD students' questionnaire regarding the improvement of IPEET

: Please evaluate the following items based on a 5-point Likert scale of agreement. If you have no ideas select undecided (the purpose of this questionnaire is to investigate the possible suggestions for the improvement of IPEET).

28. As a sub-test, academic writing should be added to the content of PhD Entrance Exam of ELT.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

29. The questions included in PhD Entrance Exam of ELT should be based on the content of MA courses.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

30. The items included in the PhD Entrance Exam of ELT should be task- based.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

31. The items included in PhD Entrance Exam of ELT should test PhD applicants' application of knowledge rather than their memorized knowledge.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

32. The questions related to some courses at BA level are not logical to be included in PhD Entrance Exam of ELT.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

33. The PhD Entrance Exam of ELT should have a more significant role in admission process than the oral interview.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

34. Part of IPEET scores should be allocated to research background.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

35. Academically celebrated professors from both center and periphery universities should have an equal hand in developing questions for PhD Entrance Exam of ELT.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

36. At the top of the answer sheet distributed among PhD applicants should be a short check list requiring PhD applicants to tick their field-specific interest.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

37. Overall, the current content of PhD Entrance Exam of ELT should be changed or improved.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

38. The decisions made by policy makers should be systematic (should not change from one year to another).

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

39. Decisions made by policy makers should be based on a collective judgment.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

40. More attention should be paid to quality rather than the quantity of articles as research backgrounds.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

41. The evaluation of PhD applicants' educational background should be based on a centralized system.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

42. The current policy of IPEET should be changed or improved.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

A8. Comment:

Please write any additional comments you would like to make about the improvement of PhD Entrance Exam of ELT

The End of the Questionnaire

Appendix B: Post Graduate University Professors' General Questionnaire

Thank you for taking the time to complete this questionnaire. You are helping us improve the quality of Iranian PhD Entrance Exam of TEFL (IPEET). This questionnaire is designed for a PhD dissertation. Completing this questionnaire is completely voluntary and all possible measures will be taken to ensure the confidentiality of your personal information.

B1. Background Information: Please tick the appropriate answer.

Your gender: Male ○ Female ○

Your age: below 25 ○ 26-35 ○ 36-45 ○ 46 or above ○

Your rank: Assistant professor ○ Associate professor ○ Professor ○

Years of teaching: less than 10 ○ 11-20 ○ 21-30 ○ 31 years or above ○

The credit courses you teach at university:

Teaching Methodology	○	SLA	○
Advanced Testing	○	FLA	○
Advanced Research	○	Discourse Analysis	○
Material Development	○	Syntactic Argument (Linguistics)	○

B2. University professors' opinion regarding the improvement of the content and decision quality of IPEET.

Please evaluate the following items based on a 5-point Likert scale of agreement. If you have no ideas select undecided (the purpose of this questionnaire is to investigate the possible suggestions for the improvement of IPEET).

1. As a sub-test, academic writing should be added to the content of PhD Entrance Exam of ELT.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

2. The questions included in PhD Entrance Exam of ELT should be based on the content of MA courses.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

3. The items included in the PhD Entrance Exam of ELT should be task- based.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

4. The items included in PhD Entrance Exam of ELT should test PhD applicants' application of knowledge rather than their memorized knowledge.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

5. The questions related to some courses at BA level are not logical to be included in PhD Entrance Exam of ELT.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

6. The PhD Entrance Exam of ELT should have a more significant role in admission process than the oral interview.

Strongly Disagree ○ Disagree ○ Undecided ○ Agree ○ Strongly Agree ○

7. Part of IPEET scores should be allocated to research background.

Strongly Disagree ⬭ Disagree ⬭ Undecided ⬭ Agree ⬭ Strongly Agree ⬭

8. Academically celebrated professors from both center and periphery universities should have an equal hand in developing questions for PhD Entrance Exam of ELT.

Strongly Disagree ⬭ Disagree ⬭ Undecided ⬭ Agree ⬭ Strongly Agree ⬭

9. At the top of the answer sheet distributed among PhD applicants should be a short check list requiring PhD applicants to tick their field-specific interest.

Strongly Disagree ⬭ Disagree ⬭ Undecided ⬭ Agree ⬭ Strongly Agree ⬭

10. Overall, the current content of PhD Entrance Exam of ELT should be changed or improved.

Strongly Disagree ⬭ Disagree ⬭ Undecided ⬭ Agree ⬭ Strongly Agree ⬭

11. The decisions made by policy makers should be systematic (should not change from one year to another).

Strongly Disagree ⬭ Disagree ⬭ Undecided ⬭ Agree ⬭ Strongly Agree ⬭

12. Decisions made by policy makers should be based on a collective judgment.

Strongly Disagree ⬭ Disagree ⬭ Undecided ⬭ Agree ⬭ Strongly Agree ⬭

13. More attention should be paid to quality rather than the quantity of articles as research backgrounds.

Strongly Disagree ⬭ Disagree ⬭ Undecided ⬭ Agree ⬭ Strongly Agree ⬭

14. The evaluation of PhD applicants' educational background should be based on a centralized system.

Strongly Disagree ⬭ Disagree ⬭ Undecided ⬭ Agree ⬭ Strongly Agree ⬭

15. The current policy of IPEET should be changed or improved.

Strongly Disagree ⬭ Disagree ⬭ Undecided ⬭ Agree ⬭ Strongly Agree ⬭

B3. Comment:

Please write any additional comments you would like to make about the improvement of PhD Entrance Exam of ELT

B4. University professors' opinion with regard to PhD students' relative abilities.

Please evaluate the following items based on a 5-point Likert scale of agreement. If you have no ideas select undecided (the purpose of this questionnaire is to investigate PhD students' qualifications in terms of their performance on required PhD courses).

16. Most of the PhD students of TEFL do not have problems with language proficiency.

Strongly Disagree ⬭ Disagree ⬭ Undecided ⬭ Agree ⬭ Strongly Agree ⬭

17. Most of the PhD students of TEFL do not have problems with the content of specialized courses like SLA, FLA, Discourse etc.

Strongly Disagree ⬭ Disagree ⬭ Undecided ⬭ Agree ⬭ Strongly Agree ⬭

18. When writing a research paper, most of the PhD students of ELT do not have problems with principles of academic writing.

Strongly Disagree ⬭ Disagree ⬭ Undecided ⬭ Agree ⬭ Strongly Agree ⬭

19. Most of the PhD students of TEFL do not have problems with basic principles of research.

Strongly Disagree ⬭ Disagree ⬭ Undecided ⬭ Agree ⬭ Strongly Agree ⬭

B5. Critical Comment: Please write any additional comments you would like to make about the *PhDstudents' qualifications in terms of their performance on required PhD courses*.

The end of questionnaire

Appendix C: Post Graduate University Professors' Specialized Questionnaire

C1: University Professors' Specilized Questionnaire of Teaching

Please take a look at the following sample of questions for the teaching subtest of PhD Entrance Exam of ELT. **Then, on a 4-point Likert scale of quantity that comes after the sample, please evaluate to what extent the following important principles and skills that PhD students of ELT should be familiar with in PhD programs, have been assessed in the sample of teaching questions attached below.**

Methodology

1- All of the following are examples of metacognitive strategies EXCEPT ----------.
1) overviewing and linking with already known material
2) reasoning deductively
3) identifying the purpose of a language task
4) paying attention

2- Nativists contend that ----------.
1) conditioning can account for the acquisition of language
2) human beings are socio-programmed
3) language competency develops in predetermined steps
4) all linguistic categories are universal

3- In Bachman's model, functions of language are directly subsumed under --------- competence.
1) sociolinguistic 2) pragmatic 3) illocutionary 4) textual

4- According to Schumann's acculturation theory, ---------.
1) a good learning situation is one in which the L2 learners' group is non-cohesive
2) motivation is either integrative or instrumental
3) a dominant L2 learners' group can help language learning
4) social distance is a metacognitive variable

5- A comprehensive theory of SLA, according to Long, should ----------.
1) recognize acquisition as a regular intake of generalizations
2) be social constructivist by nature
3) mainly focus on subconscious acquisition
4) account for universals

6- Halliday believes that ----------.
1) the linguistic aspect of language is illocutionary rather than locutionary
2) "Pronounce you guilty" has an instrumental function
3) functions of language are either personal or interpersonal
4) "The sun is hot" has a representational function

7- The current research on language learning strategies has already established all of the following EXCEPT their ----------.
1) transferability across languages 2) socio-cognitive nature
3) worldwide operational measures 4) teachability

8- All of the following themes refer to problematic areas in task-based language teaching EXCEPT --------.
1) causing hindrance to learners' intrinsic motivation
2) task difficulty and sequencing
3) cultural resistance and curriculum mismatch
4) being too structured

9- All of the following hypotheses cast doubt on the psycholinguistic validity of 'practice' as the building block of a grammar teaching course EXCEPT ---------.
1) transformational grammar 2) teachability hypothesis
3) natural order hypothesis 4) input hypothesis

10- The current principled approaches to language teaching build upon all of the following EXCEPT --------.

1) meaningful learning and anticipation of reward
2) scientific quantification and universal generalization
3) interlanguage and communicative competence
4) autonomy and self-confidence

11- The teacher who tries to help his students overcome low vocabulary size in reading comprehension through purposeful proactive attention is teaching ---------- strategies.

1) advance organizer 2) key word 3) self-monitoring 4) grouping

12- An example of a learning-centered method is ----------.

1) Suggestopedia 2) Total Physical Response
3) the Functional-Notional Approach 4) the Natural Approach

13- Consider the following exchange:

Teacher: What did you eat for dinner?
Student: I eat a sandwich.
Teacher: You ate a sandwich.

The type of correction made by the teacher is --------.

1) metalinguistic 2) explicit and deductive
3) recast 4) repair

14- The humanistic approach to language teaching ----------.

1) posits that a match between teachers' affection and that of students is of paramount importance
2) gives weight to both affective and cognitive factors
3) accentuates cognitive factors more than affective factors
4) highlights the priority of affection over intake

15- The strategy of relating new information to other concepts in memory is known as --------.

1) contextualization 2) inferencing 3) elaboration 4) transfer

1) Criteria for the critique of issues in language teaching.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

2 the critical analysis of method era.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

3) Critique of post method era.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

4) Critique of research method in language teaching.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

5) Language identity, professional identity and intercultural identity of teachers.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

6) Learners' language and intercultural identity.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

7) Critique of language teacher education.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

8) Critical pedagogy.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

9) Cultural and social issues of English as an international language

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

10) Critique of models of communicative competence.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

11) In the following box, please write any topics which you think are important, but not mentioned here and may be represented or not represented in the exam.

C2: University Professors' Specilized Questionnaire of Research

Please take a look at the following sample of questions for the research subtest of PhD Entrance Exam of ELT. Then, on a 4-point Likert scale of quantity that comes after the sample, please evaluate to what extent the following important principles and skills that PhD students of ELT should be familiar with in PhD programs, have been assessed in the sample of Research questions attached below.

Research Methodology

31- Unlike other kinds of triangulation, theoretical triangulation is mainly aimed at ----------.
1) drawing on different measures to investigate a particular phenomenon
2) enhancing the validity of the information
3) using multiple perspectives to analyze the same set of data
4) using multiple observations to obtain data

32- Which of the following is TRUE of quasi-experimental research?
1) Random assignment is not ensured.
2) All groups need to receive treatment.
3) Within-group rather than between-groups design is at work.
4) The correlation between or among variables is the basis of all prediction.

33- The concept of confirmability in qualitative research ----------.
1) needs to be analyzed through triangulation
2) refers to the three components of thick description
3) is based on the credibility of the finding to the research population
4) is analogous to reliability in quantitative research

34- In which of the following experimental types of research, control is exclusively achieved through replication?
1) Factorial designs
2) Time series designs
3) Quasi-experimental designs
4) Single subject designs

35- In multiple regression analysis, in which of the following cases will a predictor variable have maximum amount of unique variance?
1) It has a high correlation with the criterion variable.
2) It has zero correlation with the other predictor variables.
3) It has a high correlation with the criterion variable and zero correlation with the other predictor variables.
4) It has a high correlation with the criterion variable and low correlation with the other predictor variables.

36- Which of the following types of research can be used to discover the effect of one variable on another?
1) Survey research
2) Ex-post-facto research
3) Experimental research
4) Correlational research

37- To guard against wild samples and to cater for systematic variation in the population, it would be advisable to use ---------- sampling.
1) simple random
2) comprehensive
3) proportional stratified
4) extreme case

38- What is NOT true about case study?
1) The researcher should use a single procedure for data collection.
2) It should provide a detailed description of the case under investigation.
3) It should focus on a single unit, whether an individual or an organization.
4) It has a potential for theory-building and/or generalization to other cases.

39- Meta-analysis involves explicit criteria for including relevant studies as well as ---------.
1) quantitative measure of effect size
2) qualitative analysis of their findings
3) re-analysis of the data in other studies
4) synthesis of a wide range of topics

40- The use of time-series designs is recommended when ----------.
1) there is a systematic variation in the population
2) random assignment and having a control group is not feasible
3) treatment and control groups are different at the outset of the study
4) there is a danger of sensitizing the subject with pretest

41- If with t-observed = 3 , df = 35 , the null hypothesis is rejected at p<.05 , we may conclude that ----------

1) t-critical must be smaller than 3
2) the null hypothesis is also rejected at p<.01
3) directional hypothesis cannot be maintained with the same values
4) t-observed must be less than 3 for df=30

42- In qualitative research, detailed analysis of contextual factors, participants, and their roles in the social setting refers to ----------.

1) audit trail 2) triangulation 3) grounded theory 4) thick description

43- You want to examine the effect of experience on teacher's self-efficacy. You divide your sample into the following subcategories: 1-5 , 5-15, and beyond 15 years of experience. You check self-efficacy through a questionnaire. The appropriate statistical test would be ----------.

1) Analysis of Covariance (ANCOVA) 2) one-way ANOVA
3) multivariate ANOVA 4) three-way ANOVA

44- What is NOT true about mixed-methods research?

1) It incorporates blends of paradigm and philosophical positions.
2) It is clear whether qualitative or quantitative aspect is emphasized.
3) Multiple forms of data are used, both qualitative and quantitative.
4) Mixing can take place in any or all phases of the study.

45- What analytic technique is appropriate for the study below?

An MA student of TEFL intends to find out if paraphrasing or L1 translation of texts would make any difference in adult EFL students' level of reading comprehension. In so doing, she has to think of a number of variables such as the nature of the text, the participants' level of proficiency, the measuring instruments, etc.

1) t-test 2) correlation 3) think-aloud study 4) two-way ANOVA

1) Principles of research in language teaching.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

2 The basics of qualitative and quantitative research.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

3) Methods of data collection in qualitative and quantitative research.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

4) Cross-sectional and longitudinal research.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

5) Methods of data analysis and interpretation in quantitative and qualitative research such as factorial analysis, regression, directional analysis and interview analysis.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

6) Critical considerations of quantitative and qualitative research in language teaching.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

7) Ethical consideration in language teaching research.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

8) In the following box, please write any topics which you think are important, but not mentioned here and may be represented or not represented in the exam.

C3: University Professors' Specilized Questionnaire of Testing

Please take a look at the following sample of questions for the testing subtest of PhD Entrance Exam of ELT. **Then, on a 4-point Likert scale of quantity that comes after the sample, please evaluate to what extent the following important principles and skills that PhD students of ELT should be familiar with in PhD programs, have been assessed in the sample of testing questions attached below.**

Language Testing

46- Dynamic assessment from a sociocultural perspective ----------.
1) is primarily related to ZPD
2) needs to be non-gradual and given by peers
3) should mainly take place at the intrapsychological plane
4) is based on the distinction between object and human mediation

47- In the input-response relationship, ----------.
1) reciprocity negatively affects the expected response
2) both input and response are part of test method facet
3) adaptive relationship requires both feedback and interaction
4) the two options are either nonreciprocal or adaptive

48- In Bachman's (1990) model of test development, ----------.
1) quantifying test performance observation is part of operational definition
2) unlike language skills, general proficiency should be defined theoretically
3) quantifying test performance observation requires defining units of measurement
4) deciding on the scoring scale should be a prerequisite to the operational definition of a construct

49- In assessing the pragmalinguistic component of ESL learners' pragmatic competence, the rating rubric should ----------.
1) include the use of politeness marker
2) be derived from the norms of the Expanding Circle
3) be based on the learners' performance in real-life situations
4) focus on the consideration of social norms and conventions

50- "Many different kinds of evidence can be provided to support the intended interpretations and use of a test". This statement ----------.
1) is in line with validity as a unitary concept
2) is valid only if the multitrait-multimethod matrix is used
3) is based on the findings of confirmatory factor analysis
4) goes against Messick's conceptualization of construct validity

51- In the test performance research based on DIF, ----------.
1) items function differentially due to rater bias
2) testee variables such as gender and ethnicity count
3) each item is considered as an independent variable
4) rater severity is the main concern in understanding test information function

52- Systematic errors have all of the following characteristics EXCEPT ----------.
1) tending to decrease validity
2) introducing bias into measures
3) tending to decrease estimates of reliability
4) limiting the generalizability of test scores as indicators of universe scores

53- Which of the following statements relevant to Bachman and Palmer's (1996) notion of test usefulness is FALSE?
1) The individual qualities that affect test usefulness need to be evaluated independently.
2) The threshold level for practicality in any given situation would be one in which required resources do not exceed available sources.
3) Interactiveness is considered to be the extent and type of involvement of the test taker's characteristics in accomplishing a test task.
4) Authenticity is the degree of correspondence between the characteristics of the test task and the target language use task.

54- Which of the following definitions is FALSE?
1) The practice of teaching to the test in order to raise test scores is called test score pollution.
2) Formative assessment is using assessment information to provide feedback to the teaching/learning process.
3) Aggregation refers to the collapsing of the detailed performance profile for each individual into a single grade.
4) A test is systemically valid to the extent that it provides evidence confirming the assessment system being practiced.

55- Which of the following is NOT a potential problem with reliability estimates based on correlational analyses?
1) Short tests
2) Skewedness
3) Homogeneity of test takers
4) Linear relationships

56- Which of the following is used as a reliability estimate for NRTs?
1) Guttman split-half estimate
2) Threshold loss agreement statistics
3) Squared-error loss agreement coefficients
4) Domain score dependability estimates

57- Which of the following statements about Test Information Function (TIF) is FALSE?
1) It is the IRT analog of classical true score reliability.
2) It provides estimates of measurement errors at various ability levels.
3) It provides the least information for test takers at or near the level of the test.
4) It is independent of the particular sample of individuals taking the test.

58- Which of the following statements is FALSE about G-Theory?

1) If all test takers take every item in the test, it is called a crossed design and is symbolized as *p x i*.
2) When the number of conditions for a facet in a G-study includes all the conditions of the D-study, the facet is considered to be a fixed facet.
3) G-theory provides an estimation of an individual's level of ability, independently of the particular set of items used.
4) G-theory allows us to estimate the different variance components, except the highest-order interaction, which cannot be distinguished from the error variance.

59- Which of the following refers to an analytical process of test creation in which we analyze a test item to see what it is testing in order to infer the underlying guiding principles of the item, both to decide whether it is a useful item and to help generate similar items, if necessary?

1) Piloting 2) Field testing 3) Prototyping 4) Reverse engineering

60- Which of the following statements about classical true score measurement is FALSE?

1) Classical true score measurement considers all sources of error to be random.
2) Classical true score measurement fails to distinguish between different sources of variance.
3) The true score in classical true score measurement theory is the analog of universe score in G-theory and theta (θ) in item response theory.
4) In classical true score measurement theory, reliability is defined in terms of observed score variance.

1) Principles of assessment.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

2 Recent development in language assessment.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

3) Communicative and activity-based assessment.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

4) Language learning theories and assessment.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

5) Dynamic and non-dynamic assessment.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

6) Assessing language through self-assessment, teacher assessment, and homogeneous assessment.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

7) Assessment of language competence and pragmatics.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

8) Methods of analysis for the results of the test.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

9) Critical language assessment.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

10) Ethics of language assessment.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

11) Fairness and biasedness in language assessment

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

12) In the following box, please write any topics which you think are important, but not mentioned here and may be represented or not represented in the exam.

C4: University Professors' Specialized Questionnaire of SLA

Please take a look at the following sample of questions for the SLA subtest of PhD Entrance Exam of ELT. **Then, on a 4-point Likert scale of quantity that comes after the sample, please evaluate to what extent the following important principles and skills that PhD students of ELT should be familiar with in PhD programs, have been assessed in the sample of SLA questions attached below.**

Second Language Acquisition

71- In Activity Theory, ----------.
1) "internalization" is used instead of "appropriation"
2) activity is the motive behind actions
3) the surface behavior is called action
4) all needs are socially constructed

72- In the information processing model of SLA (Susan Gass), the "integration" stage refers to --------.
1) assimilating comprehended input to existing knowledge system
2) analyzing apperceived input
3) restructuring existing knowledge system
4) noticing and parsing the input

73- In terms of the Full Transfer/Full Access (FTFA) hypothesis, ----------.
1) UG role is limited to all parameters not principles
2) all features of L1 are transferred to L2 grammar
3) interlanguage and the native speaker's grammar are the same
4) L2 grammar is UG constrained

74- All of these features characterize the construct of L2 implicit knowledge EXCEPT --------.
1) early learning favored
2) primary focus on form
3) consistent responses
4) time pressure

75- All of the following are most likely to underlie the sociocultural theory of SLA EXCEPT ----------.
1) successful learning should lead to the appropriation of new knowledge
2) the most fruitful dialogic interaction is expert-expert
3) there should be a regulatory scale for error feedback
4) language is centrally a tool for thought

76- The main contribution of Dulay et al. to SLA studies was that ----------.
1) there is a similar order in the acquisition of L2 morphemes
2) children cannot learn a wide range of L2 rules
3) the L1 plays a major role in the L2 acquisition process
4) few L2 errors are developmental

77- The Minimalist Program does NOT support the idea that ----------.
1) languages are different from one another only because of lexicons
2) language faculty consists of computational lexicon and UG
3) "narrow syntax" is basically invariant across languages
4) "merge" is a computational principle

78- Which of the following features is appropriate in the Interaction Hypothesis of Long?
1) Automatic rather than controlled processing should be taken into account in learning L2.
2) Negative feedback is facilitative to learning L2.
3) Natural order of acquisition takes place in learning L2.
4) Form-focused instruction is mostly needed to learn L2 pragmatics.

79- The feature of perceptual saliency is that ----------.
1) the beginning and end of stimuli are easier to remember and to manipulate
2) stages of acquisition cannot be skipped through formal instruction
3) learners will first be able to move elements from outside to inside the sentence
4) underlying semantic relations should be marked overtly and clearly

80- Based on UG, second language learners ----------.

1) have available to them from the onset the full range of UG principles and set parameters
2) start off with the parameter settings of their L1
3) resort to first language parameter setting in the last stance
4) reset principles on the basis of input

81- From a Vygotskyan perspective, it would be argued that we witness microgenesis in the learner's second language system ----------.

1) through the appropriation of a new lexical item from the scaffolding talk of the native speaker
2) which appears to take place during scaffolded teacher-student talk
3) while the negotiated zone of proximal development is led to explicit feedback
4) in social settings and as a result of interaction within the ZPD

82- In connectionism, the real criticism is ----------.

1) lack of distinction between competence and performance
2) that it is based on language making capacity
3) that learning occurs based on associative processes
4) ignoring both property and transition theories

83- Which of the following is NOT among the characteristics of the information-processing approach?

1) Complex behavior is composed of simpler processes that are modular.
2) The mind is a limited-capacity processor.
3) Component processes cannot be isolated.
4) The mind is a symbol-processing system.

84- Which of the following states that the frequency of a feature in the materials is most likely to affect L2 learning?

1) Real-operating conditions principle
2) Given-to-new principle
3) Markedness hypothesis
4) Input-flooding strategy

85- Which of the following statements is TRUE of the Zone of Proximal Development (ZPD)?

1) It suggests that what we can do today with assistance is likely to be done independently later.
2) It is the same thing as scaffolding or assisted performance.
3) It is conceptually and theoretically similar to Prabhu's concept of reasonable level of challenge.
4) It is similar to Krashen's notion of *i+1*.

86- Which of the following hypotheses does NOT have a role in the updated version of the Interaction Hypothesis?

1) teachability hypothesis
2) noticing hypothesis
3) output hypothesis
4) input hypothesis

87- Why is focusing on pragmatic meaning of paramount importance?

1) It provides an opportunity for a focus-on-form approach.
2) It contributes to the learning of formulaic expressions.
3) It is likely to bring about change in acquisitional route.
4) It is intrinsically motivating and fosters fluency.

88- What is the difference between comprehensible and comprehended input?

1) Comprehended input fosters implicit knowledge, while comprehensible input develops explicit knowledge.
2) Comprehensible input is speaker-controlled but comprehended input is learner-controlled.
3) Comprehensible input is concerned with meaning but comprehended input deals with form.
4) Comprehended input is a dichotomous variable, while comprehensible input is not.

89- Which SLA theory does the principle "learners tend to process the first noun or pronoun they encounter in a sentence as the subject" belong to?

1) Pienemann's Processability Theory
2) Chomsky's Universal Grammar
3) Lantolf's Sociocultural Theory
4) VanPatten's Input Processing

90- The three distinct phases proposed in Dornyei's motivational cycle respectively follow as ----------.

1) choice motivation, executive motivation, and motivational retrospection
2) executive motivation, motivational retrospection, and choice motivation
3) executive motivation, choice motivation, and motivational retrospection
4) choice motivation, motivational retrospection, and executive motivation

1) Principles of theorizing in second language.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

2 The history of theorizing in second language.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

3) Traditional theories related to second language learning.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

4) Modern theories related to second language learning.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

5) The role of external factors like materials or learning environment.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

6) The role of internal factors like cognitive and affective variables of language learner.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

7) Research methods in second language studies.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

8) Research findings related to second language learning.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

9) In the following box, please write any topics which you think are important, but not mentioned here and may be represented or not represented in the exam.

C5: University Professors' Specilized Questionnaire of Discourse

Please take a look at the following sample of questions for the Discourse subtest of PhD Entrance Exam of ELT. **Then, on a 4-point Likert scale of quantity that comes after the sample, please evaluate to what extent the following important principles and skills that PhD students of ELT should be familiar with in PhD programs, have been assessed in the sample of *Discourse* questions attached below.**

Discourse Analysis

91- Felicity conditions are met when ----------.

1) communication is carried out by the right person in a right place at the right time
2) rules and principles in a communication are followed
3) communication is carried out in a particular context
4) the analysis of speech acts, implied meaning, and pragmatic routines are taken into account

92- The difference between conventional presupposition and pragmatic presupposition is that ----------.

1) the latter is context-independent and arises from the use of an utterance in a particular context
2) the former is based on politeness universals
3) the latter is typically linked to particular linguistic forms
4) the former is less context-dependent

93- Schegloff criticizes critical discourse analysis for ----------.

1) overemphasizing the context in which a text is produced
2) lack of attention to issues of power, inequality, and social status
3) lack of attention to wider historical, cultural, and political issues
4) overlooking how the participants take up what is said in the text

94- When speakers report on people's mental states, they often use expressions which identify the type of mental state. This is called ----------.

1) formulaic expression
2) domain restriction
3) propositional attitude
4) indexicality

95- Given that conceptualizations of face are rooted in conceptualizations of the social self, (Arundale, 2006) ----------.

1) face explains the actions of individuals as caused by internal needs
2) face is a social psychological phenomenon
3) face is a matter of the individual actor's public self-image
4) face is a relational phenomenon

1) Approaches to discourse analysis.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

2 written discourse analysis.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

3) Oral discourse analysis.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

4) Classroom discourse analysis.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

5) Basics and models of critical discourse analysis.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

6) Critical classroom discourse analysis.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

7) Media critical analysis.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

8) Cohesion and coherence.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

9) Research methods in discourse analysis

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

10) The analysis of speech and para speech symbols

not at all ⬭ slightly ⬭ some extent ⬭ to a large extent ⬭

11) In the following box, please write any topics which you think are important, but not mentioned here and may be represented or not represented in the exam.

C6: University Professors' Specilized Questionnaire of Sociolinguistics

Please take a look at the following sample of questions for the Sociolinguistics subtest of PhD Entrance Exam of ELT. Then, on a 4-point Likert scale of quantity that comes after the sample, please evaluate to what extent the following important principles and skills that PhD students of ELT should be familiar with in PhD programs, have been assessed in the sample of Sociolinguistics questions attached below.

Sociolinguistics

96- According to Malinowski, ---------- is a type of speech in which ties of union are created by a mere exchange of words. They are neither the result of intellectual reflection, nor do they necessarily arouse reflection in the listener.

1) institutional language 2) phatic communion
3) constructed discourse 4) language idling

97- Which of the following is NOT true of conversation analysis?

1) Language in its primitive function and original form has essentially encyclopedic character.
2) Historically, and almost certainly phylogenetically, talk came first.
3) Conversation analysis emerged from two intellectual streams in sociology.
4) Conversation is mutually constructed and negotiated.

98- As one of the processes which a typical standard language passes through, codification refers to ----------.

1) the function of the standard language as a unifying force for society
2) agreement on what is correct by everyone
3) the recognition of the language as one of the social dialects
4) the degree of mutual intelligibility

99- The term acrolect refers to ----------.

1) a dialect from which all standard dialects originate
2) a range of non-standard dialects
3) a creole's lexical source language
4) the basilect spoken by children

100- The expressions such as "This is probably wrong" and "I suppose …" as used more frequently by women mainly show ----------.

1) women's preference for the violation of maxim of quantity
2) women's tendency to use more hedges
3) the importance of the lexifier language
4) the falsity of genderlects

1) Approaches to discourse analysis.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

2 written discourse analysis.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

3) Oral discourse analysis.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

4) Classroom discourse analysis.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

5) Basics and models of critical discourse analysis.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

6) Critical classroom discourse analysis.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

7) Media critical analysis.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

8) Cohesion and coherence.

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

9) Research methods in discourse analysis

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

10) The analysis of speech and para speech symbols

not at all ⬭ slightly ⬭ to some extent ⬭ to a large extent ⬭

11) In the following box, please write any topics which you think are important, but not mentioned here and may be represented or not represented in the exam.

The end of questionnaire

Appendix D: Focus Group Interview Items

1. What's your idea about the characteristics of the PhD Entrance Exam of ELT in terms of adequacy of the number of items and their level of difficulty?
2. What's your idea about the conditions (time and physical conditions) under which the test instruments were taken?
3. To what extent do test practitioners at Educational Assessment Organization (EAO) inform university professors and PhD applicants of the type of decisions they will make on the admission of PhD applicants.
4.To what extent do you think these decisions are based on the collective judgments of a wide range of stakeholders?
5.To what extent do you think test practitioners at Educational Assessment Organization (EAO) report test scores in ways that are understandable to PhD applicants?
6.To what extent do you think the use of the test helps promote good instructional practice and effective learning in ELT instructional settings?
7. Do you think that the current procedure of selecting PhD candidates is appropriate for PhD program? If no, what are your suggestions for the possible ways of improvement for this procedure?

Appendix E: Telephone Interview Items

1. To what extent do you think the content of the tasks or items included in the instruments (Multiple-Choice Exam) represent the content of MA courses and relate to PhD courses at universities?
2. To what extent do you think the current decisions made by the policy-makers (EAO & MSRT) on the cut scores or on the classifications of PhD applicants are based on the collective judgments of a wide range of stakeholders?
3. What's your idea about the characteristics of the test instrument in terms of the adequacy of the number of items and their level of difficulty?
4. What is your idea regarding PhD students' abilities in terms of their performance on required PhD courses? Are you satisfied with them generally?
5. To what extent do you think the use of the test helps promote good instructional practice and effective learning in ELT instructional settings
6. Do you think that the current procedure (Semi-Centralized PhD Exam of ELT) for selecting PhD candidates is appropriate for PhD program? If no, what are your suggestions for the possible ways of improvement for this procedure?