Interpreting the Validity of a High-Stakes Test in Light of the Argument-Based Framework: Implications for Test Improvement[1]

Ali Darabi Bazvand[2] *& Alireza Ahmadi*[3]

Received: 14/12/2018 Accepted: 15/09/2019

Abstract

The validity of large-scale assessments may be compromised, partly due to their content inappropriateness or construct underrepresentation. Few validity studies have focused on such assessments within an argument-based framework. This study analyzed the domain description and evaluation inference of the Ph.D. Entrance Exam of ELT (PEEE) sat by Ph.D. examinees (*n* = 999) in 2014 in Iran. To track evidence for domain definition, the test content was scrutinized by applied linguistics experts *(n* = 12). As for evaluation inference, the reliability and differential item functioning (DIF) of the test were examined. Results indicated that the test is biased because (1) the test tasks are not fully represented in the Ph.D. course objectives, (2) the test is best reliable for high-ability test-takers (IRT analysis), and (3) 4 items are flagged for nonnegligible DIF (logistic regression [LR] analysis). Implications for language testing and assessment are discussed and some possible suggestions are offered.

Keywords*:* Argument-Based Validity; Differential Item Functioning (DIF); ELT

1. Introduction

Over the past few decades, there has been a growing use of high-stakes educational assessments across the globe with the information yielded by such assessments being increasingly relied upon by policymakers for a multitude of decisions. The quality of these decisions may be contingent upon the appropriateness

[1] Please cite this paper as follows:

Darabi Bazvand, A., & Ahmadi, A. (2020). Interpreting the validity of a high-stakes test in light of the argument-based framework: Implications for test improvement. *Journal of Research in Applied Linguistics*, *11*(1), 66-88.

[2] English Department, College of Languages, University of Human Development, Sulaimani, Iraq; *alidarabi1350@gmail.com*

[3] Corresponding Author, Department of Foreign Languages and Linguistics, Shiraz University, Shiraz, Iran; *arahmadi@shirazu.ac.ir*

of the interpretations made on the assessment results. As documented in the literature, high-stakes standardized instruments are associated with some technical expressions, such as levers for change (Alderson, 1986), powerful learning (Stiggins, 1990), standards of teaching (Kiany, Shayestefar, Ghafar Samar, & Akbari, 2013), and upward social mobility (Ross, 2008).

Nonetheless, it has been reported that the relative effectiveness of such instruments is a long- standing and highly polarized debate; they offer insufficient insights into different abilities which ultimately influence student success in college programs (Armstrong, 2000). Standardized instruments have also been reported not to be socially responsive for graduate studies (e.g., Motamedi, 2006). That is, they may fall short of their intentions (e.g., Farhady, 1998; Green, 2007). As such, the alleged presence of such problems may suggest that the accurate and appropriate interpretation and uses of test scores are not ensured (Moss, 2007), or there exist concerns about positive consequences of such large-scale tests (Cheng & Sun, 2015; Shepard, 2000).

It is hypothesized these potential factors inherent in large-scale tests may be prone to threats to validity; what scholars nominated as construct irrelevant variance and construct underrepresentation (Mesick, 1989). Such potential threats may obscure the correspondence between test intentions and test effects (e.g., Cizek, 2012; Hubley & Zumbo, 2011). The former occurs when the measure does not reflect the construct to be assessed; rather, additional characteristics affect performance, whereas the latter happens when the test fails to include important aspects of the construct (e.g., Cubilo, 2014; Knotch & Elder, 2013; Xi, 2008). Large-scale assessments like university entrance examinations in Iran may be prone to the abovealluded threats to validity (Kiani et al., 2013). This may be particularly true within the current Ph.D. Entrance Exam series in Iran.

The Ph.D. Entrance Exams in Iran are large-scale tests, designed and developed to measure Ph.D. applicants' General English proficiency and subject matter knowledge. These exam series are developed by the National Organization for Educational Testing (NOET), in cooperation with the Ministry of Science, Research, and Technology (MSRT), often funded by the Government. They consist of academic talent, general proficiency, and subject matter blocks. Obviously, the subject matter subtests of such exam series have more weight in screening the applicants, as compared to other two blocks. Although other factors like the applicants' performance in the interview session, their educational and research background, along with their GPA scores may have their own effects (Azmoon.NET, 2014), it can be claimed all these factors are contingent upon the written exam. For instance, it may be the case that some Ph.D. applicants with good academic and research background and with a good ability in oral performance are not short-listed for oral exams because they have not passed the written exam. Therefore, efforts should be made to ensure these written exam series are fair to test-takers.

One important section of these exam series is the Ph.D. Entrance Exam of ELT (PEEE). The PEEE test consists of a test of academic talent, a General English proficiency test, and a subject matter test, all appearing in MC format. The present study investigated the subject matter block. The subject matter test designed to measure the applicants' expertise in the field of ELT is seemingly related to the topics already been covered in the M.A. or, even B.A., programs (NOET, 2014).

Given that such written exam series play a pivotal role in screening applicants for college programs, it is assumed the content of such tests represent the important skills, knowledge, abilities, and processes (Chappelle, Enright, & Jamison, 2010) needed for ELT in Ph.D. courses. Furthermore, an assumption can be made that such field relevant subtests are not prone to test characteristics lapses, such as unreliability of the test items and differential item functioning (DIF). To date, the validation studies conducted in this area are few and far between (e.g., Hamavandy, 2014; Kheirzade, 2015). Such studies have focused on the General English section of non-English majors. The subject matter sections of Ph.D. exam series have been held in disregard. In view of the above, the present study was an attempt to track appropriate backing for the abovementioned assumptions by investigating the validity of the subject-matter section of the PEEE test in light of an argument-based framework.

1.1. A Word on Argument-Based Validity

Enlightened by Messick's formulation of test validity, some researchers proposed argument-based approaches to validity (Kane, 1992). These approaches tend to "specify the process used to prioritize, integrate, and evaluate evidence collected using various methods" (Xi, 2008, p. 177).

The argument-based approach to test validation is mainly based on a chain of evidence presented to support the interpretations and the use of test scores. According to Kane (1992), in this model, it is not the test score, *per se*, that accounts for the legitimacy of test validity, but the probable meanings assigned to that particular test score. This model is mainly based on the construction of two kinds of arguments: interpretive argument and validity argument. Interpretive argument clarifies the proposed interpretations and uses of the results of assessment by presenting a chain of inferences and assumptions guided from observed performance to the conclusions (Kane, 2006). In the validity argument, as explained by Kane (1992), the plausibility of the inferences and assumptions of the interpretive argument is critically analyzed. For example, after the test design and test operationalization are over, the test developer can provide "theoretical and empirical evidence that constitute support for the inferences and assumptions of the interpretive argument" (Chappelle, Enright, & Jamison, 2008, p. 5).

In view of the above, it can be fairly safely claimed that validation is a continuous process involving not only the accumulation of evidence to test score interpretation and uses (Kane, 2006) but judgment about the plausibility of those interpretations (Knoch & Elder, 2013; Xi, 2008). Therefore, instead of a collection of separate quantitative or qualitative evidence, validity is an argument construed by an analysis of theoretical and empirical evidence (Bachman, 1990; Chappelle, Enright, & Jamison., 2008, 2010).

Some researchers propose the structure of argument validity framework to include some bridges, such as domain definition, evaluation, generalization, explanation, extrapolation and utilization and their corresponding methods of collecting evidence to support those inferences (Bachman, 2005; Bachman & Palmer, 2010). As part of an argument-based validity framework, the present study was an attempt to synthesize domain definition and evaluation inferences.

1.1.1 Domain description and evaluation inferences

In the very processes of test development, test developers should bear in mind that specifying the test purpose and test use, as well as integrating the future needs of test-takers are integral to test validity. They should seriously consider, "what need the test will serve, what underlying construct will be applied, who will take the test, and how it will be used and by whom" (Span, 2006, p. 71). One way that test developers can do this is to set some criteria, which may come from various resources like the previous research, the construct itself, stakeholders, and audiences (Ryan, 2002). For example, a detailed and complete analysis of the construct need to be done, if we are to produce a language test that taps into the academic language needs of the students (Butler, Lord, Stevens, Borrego, & Bailey, 2004). This can fall under the category of domain definition as part of the structure of argument-based validity (Chappelle, Enright, & Jamison et al., 2010; Kane, 2011, 2013, 2015).

The domain description inference (see Table 1) is based on the warrant that performances in the target domain are related to the observations of performance in the test domain, which has been identified based on the test purpose (Chappelle, Enright, & Jamison, 2010). This warrant rests on the assumptions that assessment tasks represent important skills, knowledge, abilities, and processes needed for academic domain as identified by experts and/or proposed syllabi. Furthermore, these assessment tasks are assumed to include language skills critical to success in the academic domain. Possible backings for such assumptions may be derived from "the analytic process of domain analysis and task construction as well as from evaluations of the success of the outcomes of these processes” (Chappelle, Enright, & Jamison, p. 21).

Another important inference in the argument-based validity framework is evaluation (see Table 1). This inference is based on the assumption that the (statistical) characteristics of the test items as well as the test administration conditions are appropriate for academic L2 or content abilities (Johnson & Riazi, 2013). Likewise, factors like unclear instructions, inappropriate item characteristics (e.g., problems with difficulty levels of the items or gender-bias items), and implausible answer options, which may bring about misunderstanding for the examinee, may substantially influence outcomes, and may cast some doubt on the evaluation inference and overall validity of the instrument (Xi, 2010).

Evidence to support the evaluation claim may be collected through soliciting the insights of various stakeholders (possibly via questionnaires and interviews). Moreover, investigating the characteristics of the instruments through DIF techniques like IRT or logistic regression (LR) may provide required evidence to support or rebut the assumptions proposed in evaluation inference.

1.1.1.1. Differential item functioning (DIF)

One way for test designers to ensure their test enjoys fairness is through subjecting test items to DIF detection procedures like LR. DIF exists when examinees from different groups (e.g., males and females), but of equal ability, have different probabilities of successful performance in an item under the question. One possible source of DIF may be the effect of gender (male and female) on group differential performance. For example, in a hypothetical test, some reading comprehension texts may be partially biased for females because they have more background knowledge of those texts as compared with males. Therefore, they may perform better on the related items. When such items are subjected to DIF detection, they have the potential to be flagged for gender DIF.

Among the methods of DIF detection, LR has been widely considered as one of the best statistical methods that have been proposed to examine DIF (Swaminathan & Rogers, 1990; Zumbo, 1999). The reason is that LR uses a binomial distribution, rather than a normal distribution, and can provide a better framework for analyzing and evaluating DIF. In LR, both uniform and nonuniform DIF can be detected in a binomial distribution. These terminologies are very important in gender DIF detection. To help clarify these concepts, suppose, for example, that we have two groups of males and females taking a hypothetical test: If the male group outperforms the female group within the entire range of proficiency, uniform DIF would be the result (Hauger & Sireci, 2008). Nonuniform DIF, on the other hand, would be present when the performance difference among the groups is systematic across the matched groups, “but the direction of the difference changes at different points along the range of proficiency” (Hauger & Sireci, 2008, p. 239):

Table 1. *Articulating the Validity Framework for the Context of PEEE*

<table>
<tr><td colspan="2">Inference 1: Domain Definition:
Warrant: Observations of performance via PEEE reveal relevant knowledge, skills, processes, and strategies representative of those required for Ph.D. Programs.</td></tr>
<tr><td>Assumptions:
1.1. Critical subject matter skills, knowledge, and processes needed for study in postgraduate programs can be identified
1.2. PEEE assessment tasks requiring important skills and representing the academic domain can be simulated as test tasks</td><td>Backings:
1.1 As part of domain analysis, the researchers identified critical skills and abilities required for academic tasks.
1.2. Expert judgment was used to judge the congruence of assessment task with the course objectives as academic use tasks</td></tr>
<tr><td colspan="2">Inference 2: Evaluation:
Warrant: Observations of test-takers’ performance on the PEEE tasks as well as the characteristics of tasks themselves are evaluated to provide observed scores informative of target academic domain</td></tr>
<tr><td>Assumptions:
2. The statistical characteristics of the test items are appropriate for providing evidence of academic target content abilities</td><td>Backings:
2.1 Test score data were analyzed with regard to DIF using LR
2.2 Test score data were analyzed with regard to reliability using 1-parameter IRT</td></tr>
</table>

2. Objectives and Research Questions

This study was an attempt to theoretically and practically address the existing gaps in literature with regard to validity investigations across postgraduate subject matter tests within an Iranian context. In particular, the present study’s aim was to support the domain definition and evaluation inferences as parts of validity argument along with their corresponding assumptions. Table 1 outlines the relationships among the assumptions and types of backings underlying the warrant for these two inferences. As such, the following research questions were formulated to serve backing for the proposed inferences.

1. What are the most important knowledge, skills, abilities, and processes needed for ELT in Ph.D. courses, as identified in the mandated syllabi by applied linguistics experts? And, to what extent do they think the present

PEEE test samples such important skills, tasks, and abilities? (Domain definition)

2. To what extent are the statistical characteristics of the PEEE test items appropriate for screening Ph.D. applicants? (Evaluation)
3. What are applied linguistics experts' suggestions for the betterment of the current content and the present policy of the PEEE test?

3. Methodology

3.1. Participants

The participants comprised two groups of stakeholders: The first group consisted of all the Ph.D. applicants who had taken PEEE in January 2014 (*N* = 999). They were divided into two groups of females (*n* = 602) and males (*n* = 397) to investigate their possible group differential performance on PEEE.

The second group of participants was a sample of university professors (*n* = 12), as applied linguistics experts in their own specialty. They virtually had some experience teaching some Ph.D. courses of ELT it Ph.D. programs. Given that this group of stakeholders had been directly involved in the current practices of PEEE and had also observed the radical changes in the centralized tests in Iran, they were supposed to provide us with important information about Ph.D. evaluation. Further, to their busy schedule and mere unwillingness to cooperate, due to the fact that the number of university professors with the required characteristics (e.g., those teaching Ph.D. courses) was very limited, such a small sample took part in the study. However, given that the overall population of the professors with the required characteristics was limited (about 100), the sample invited to take part in the present study (*n* = 12) may be a considerable percentage of the total population. Among them, 10 were males, whereas only two were females. Whereas nine were associate professors of applied linguistics, three were assistant professors. With regard to the length of their teaching experience, it ranged between 10 and 28 years—the average being 18 years. All were full-time university professors. They were requested to analyze the congruence of a revised syllabus with the content of the test tasks included in PEEE. Their suggestions for the betterment of the current content and present policy of PEEE were also analyzed.

3.2. Instruments and Data Collection

Three types of instruments were employed: The first instrument was a revised syllabus mandated by MSRT for postgraduate programs. The participants were requested to analyze the content of this syllabus against which the content of PEEE was to be compared. For the betterment of the quality of responses, a copy of the PEEE test (the 2014 administration) was provided to the participants. In the revised syllabus, the course objectives were specified for each postgraduate discipline. It was based on the outcome of analysis and investigation of the curriculum of most celebrated and well-known universities in Canada, United States of America, and some other universities. For planning such syllabus, the recent developments in ELT were also studied. In the revised syllabus, five obligatory domains like issues in language teaching, language testing and assessment, research in education, FLA and SLA, and discourse were included. Optional domains like sociolinguistics, syntax, and ESP were also added to such a syllabus. Critical information taken from this syllabus guided the domain description inference.

The second instrument was a structured telephone interview conducted with a sample of university professors (*n* = 12). Because the university professors were selected from different universities across the country, it was difficult to hold face-to-face interviews; therefore, they were all interviewed through telephone. For the betterment of the quality of responses, a copy of the PEEE test (the 2014 administration) was provided to the participants. The interview took place in the course of January 2015 and early May 2015. Each long-guided interview lasted from 30 to 45 min and solicited the participants' views with regard to any possible change in the current content and present policy of PEEE. The language of the interviews was English, and the interview data were recorded on an audio-cassette and, subsequently, were transcribed. The interview items were aimed at soliciting university professors' opinion regarding: (a) relevance of the PEEE test tasks to the Ph.D. courses (the first research question), and (b) their possible suggestions for the betterment of PEEE (the third research question). The validity of the qualitative interpretations was established using member checks.

The third instrument was the PEEE test score data. This data came from the 2014 administration of PEEE. It subsisted of an academic talent test, a General English test, and a knowledge test. For the present study, only the knowledge test was investigated. First, the test score data were subjected to reliability analysis, using 1-parameter IRT. Then, the data were analyzed for DIF, employing LR. The findings from such analyses informed evaluation inference.

3.3. Data Analyses

Two types of analyses were undertaken for this study: Qualitative analysis of stakeholders' suggestions for the improvement of the content of PEEE and quantitative analysis of the test score data.

As for the analysis of stakeholders' perceptions (i.e., qualitative data analysis), we relied on Glaser and Strauss's (1967) method of constant comparison, which includes reading through transcriptions, developing a general category scheme, and aggregating similar codes together to develop themes and identifying categories and sub categories.

For the analysis of the PEEE, the test score data were subjected to 1-parameter IRT reliability and LR DIF. Each is described below:

One way for test designers to ensure their test productions are fair enough across exam candidates is through subjecting test items to the statistical analysis of reliability. Unfortunately, methods of reliability estimation like Cronbach's alpha are highly affected by the number of items (Bachmann, 1990). This necessitates a better model which accounts for both difficulty level of the test items and ability of test-takers (Bachman, 1990). IRT models have such characteristics and, for this reason, the present study relied on such models to investigate reliability.

Given that using LR requires a stepwise procedure which may lead to minimizing the probability of Type I and Type II errors (French & Maller, 2007), three steps were followed to identify uniform and nonuniform DIF: In the first step, the conditioning variable (i.e., total score) was entered to account for baseline proportion of variance or matching variable. In the second step, the grouping variable (i.e., gender) was entered into the equation (i.e., total score + gender). Finally, the interaction variable (i.e., total score + gender + interaction) was entered in the third step. As such, with 1 *df*, the difference between steps 1 and 2 and between steps 2 and 3 was used to determine practically significant uniform and nonuniform DIF, respectively.

However, because the interpretation on LR binomial results is affected by the size of the sample, effect sizes were also analyzed for each item. As such, DIF items were classified as following based on the criterion proposed by Jodoin and Gierl (2001):

- Negligible or A-level DIF: $R2 < 0.035$
- Moderate or B-level DIF: Null hypothesis rejected AND $0.035 \leq R2 < 0.070$
- Large or C-level DIF: Null hypothesis rejected AND $R2 \geq 0.070$

4. Results and Discussion

The present study aimed to interpret the domain definition and evaluation inferences of PEEE as parts of an argument-based validation framework.

4.1. Domain Definition Inference

The first research question aimed to investigate how the construct of ELT, as defined in the syllabus, includes relevant knowledge, abilities and skills, and to what extent such abilities are represented on PEEE test tasks and items. To find a logical answer to the first part of the research question, the revised version of ELT for Ph.D. program mandated by the High Council of MSRT was scrutinized by applied linguistics experts. Among the important goal statements of this syllabus, the following skills and abilities required for ELT discipline of Ph.D. program were identified:

1. Teaching general and specialized ELT courses in B.A., M.A., and Ph.D. levels
2. Researching issues related to L2 learning and teaching
3. Designing programs for ELT
4. Designing English textbooks for ELT and testing
5. Designing specialized ELT textbooks
6. Educating and training English language teachers

Followed from the above, the first critical ability highlighted in the goals is teaching general and specialized courses at academic contexts. This ability required of Ph.D. candidates falls outside of the category of ELT. The second ability prioritized in the revised syllabus was doing research in the field of ELT, for example, being able to publish ELT articles in high-qualified journals. This ability is so critical because the big challenge that most postgraduate students face in their academic life is dealing with research related activities, such as writing a research paper as a course requirement. Another important skill emphasized in the goal statement of the revised syllabus was designing materials and programs for ELT and language testing. The emphasis on this ability shows that designing and developing materials is of utmost importance in ELT instructional programs. One more critical ability highlighted in the syllabus was related to teacher education and teacher training. In view of the above objectives and the identified knowledge, abilities and skills, it was found that about 50% of the goal statements is devoted to the principles of curriculum development and, especially designing ESP materials, indicating that these two areas are crucial across the discipline of ELT.

As for the second part of the abovealluded research question, the applied linguistics experts were asked to investigate how the critical knowledge, skills, and abilities within the revised syllabus were congruent with the tasks and items tested in PEEE for each domain. For this purpose, the informants were consulted to investigate such congruence.

The respondents unanimously agreed that the PEEE tasks are not fully commensurate to the objectives and requirements of Ph.D. courses of ELT, claiming that some of the items are not represented in the materials taught in Ph.D. courses for this subject and some are beyond the Ph.D. students' level of competence. For example, the findings from such analysis showed that although having a critical knowledge of curriculum development was strongly highlighted and attended to in the mandated syllabus, this ELT course was not among the materials to be tested for the PEEE, suggesting that this course is underrepresented on PEEE. Moreover, they pointed out that though candidates' ability in designing ESP courses was greatly emphasized in the goal statement, but surprisingly, it was observed that ESP items and tasks were not represented on PEEE. One of the participants confirmed:

- T5: *Actually I checked the questions in details. I found just around 50% of the questions are represented in the courses in the Ph.D. program.*

Another respondent opined that:

- T8: *So as far as the content is considered, well, I think that they were partially representing the content of M.A. courses and Ph.D. courses and you see that there is a broad area of investigation.*

Some claimed of a construct-irrelevant easiness (Mesick, 1989) manifest in the present PEEE in Iran. One participant stated:

- T1: *I disagree with including the introductory linguistic subtest in PEEE since this subject is related to B.A. materials and is not based on Ph.D. courses. We don't have classes held on introductory linguistics in Ph.D. courses.*

Some others referred to construct underrepresentation (Messick, 1989), threatening the present screening test. Among the basic materials and subjects that are more or less measured in the present PEEE are research methodology, language assessment, principles/theories of language teaching, sociolinguistics, and discourse analysis. Whereas some of these subjects like discourse analysis have been introduced as critical in the course objectives, it is evident that they are disappointedly underrepresented in the test; for this subject, only five items were included in 2014 administration of PEEE. Said, an expert in the field of discourse analysis, said:

- T6: *Much to my surprise, there is a subject, em… in the test material development at Ph.D. program of ELT in the university, my university, at least. This part is not covered at all in PEEE.*

It is claimed if the content of the test does not fully represent the target academic domain, then candidates screened through such a test may not have the ability to manage the course requirements (Johnson & Riazi, 2013). This may be the reason why university professors were not fully satisfied with the achievement of Ph.D. students on required courses, claiming that their performances were far from being perfect and, indeed, were rife with different gaps and blocks, big and small. This claim is corroborated by a university teacher, arguing that no or little future success can be traced in the content of the PEEE test: As one of the participants observed:

- T3: *Unfortunately, no you know I understand that they have been selected based on their memorized knowledge and they are really weak in their application of knowledge in our classes, they are really weak. So this shows that the test is not functioning well. So this MC test cannot show the full potential of the candidates.*

Another monolithic block perceived by the university professors was the drudgery of dealing with students with mixed abilities. A professor with specialty in discourse pointed out:

- T2: *Well, we have students who have been admitted to our department and actually; they would not be able to present themselves not even in their writing which requires spontaneous abilities to be manifested.*

Among the skills, knowledge abilities required for postgraduate students' success, as suggested by experts, academic writing occupied the pride of place. Thanks to both their inability in academic writing and language proficiency, they are not qualified, a claim corroborated by one of the professors remarking:

- T7: *As mentioned before in the past few years the level of proficiency of Ph.D. candidates has reduced quite dramatically which is actually, something quite dissatisfactory, when compared with the students accepted via traditional system.*

These findings are, somehow, consistent with what is presented in the literature (though on different test application contexts), which refer to the similar standardized instruments as failures with regard to their relatedness to the target content/language use domain (James & Templeman, 2009). Thus, the lack of correspondence between test content and academic target use domain may be partially associated with indirect assessment of language skills, the closed-ended nature of such tests, and the complex requirements of the instructional courses into which applicants are to be placed (Murphy & Yancey, 2008; Williams, 1990). This may be true as part of an explanation for the findings of the present study

Another possible explanation for the alluded above mismatch may rest on the fact that tests' simulation to language use in target situation is partially limited and, in this case, they are used as tools for readiness to enter the academic world, rather than the degree of the mastery students have over the required academic skills (Taylor, 2007). This can also be confirmed by Goldman (2004), arguing that readiness to fulfill the course requirements cannot be measured through a limited test when students are supposed to do much more complex tasks in the target academic context. This apparent mismatch between the intentions of the test content and that of target content use domain may be true for the present PEEE and may be considered as one reason for why the present PEEE fails to predict success for applicants.

Still, another explanation can be attributed to the fact that Ph.D. programs are more research-based, an area on which Ph.D. candidates have not had ample opportunities to work neither at their B.A. nor at M.A. levels. Moreover, such programs are more focused on application of knowledge and applicants 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 was not the case for the current PEEE. These factors may be considered as other sources of paradox, in addition to the ones mentioned so far.

It is suggested, then, that test developers design tests which are, more or less, based on the real academic courses for which applicants are to be prepared. Before designing tests, they should critically evaluate the syllabi mandated for a particular postgraduate course of instruction.

4.2. Evaluation Inference

The second research question was raised to investigate the extent to which the PEEE test characteristics were appropriate to screen postgraduate applicants for Ph.D. programs. Therefore, analyzing the reliability and differential functioning of the items were relevant.

4.2.1. Reliability

Among the methods of reliability analysis, the IRT model is more logical because it can account for both the difficulty level of the items and the ability of test-takers (Bachman, 1990). As displayed in Figure 1, the highest amount of information is displayed for very high levels, suggesting that the test is best reliable for such a level. Furthermore, a closer look at item information functions reported for individual items well indicated that most items are reliable if used for individuals with very high levels of ability. Nevertheless, when these curves of information functions are matched with the test and item characteristic curves, indicating the ability of the individual test-takers, it can be easily noticed that this test does not enjoy reliability for the target test-takers. Therefore, with this piece of evidence, it would be fairly safe to rebut the reliability assumption of PEEE. As such, it can be argued that such poor reliability may put the scoring validity of the test under question. In line with this claim, researchers agree that the validity of tests associated with inappropriate instrument technical quality (Bennett, 2010) may introduce more than construct-irrelevant variance (CIV) in observed scores (Mesick, 1989). This is particularly true with tests of, for example, poor reliability, as it is the case for the present study. It is true, then, that improving the psychometric characteristics of such tests leads to appropriate and valid test score interpretations (Chappelle, Enright, & Jamison, 2010). Thus, it is suggested, then, that the Ph.D. Entrance Exam series in Iran should be piloted with appropriate samples to critically evaluate their reliability (especially via IRT models) and DIF magnitudes before they are used for different practical purposes:

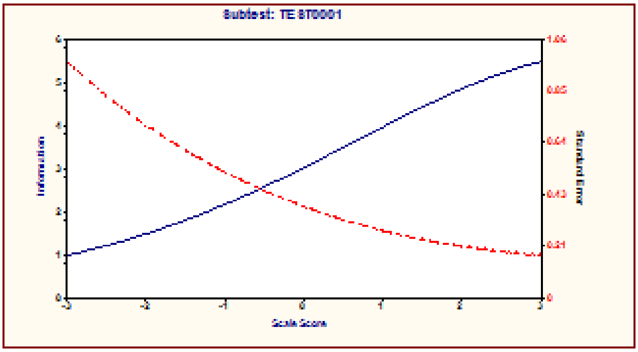

Figure 1. Reliability Information for PEEE

4.2.2. Differential item functioning (DIF)

One source of evidence to inform the evaluation inference is DIF. To examine this evidence, the study employed LR model to detect potential DIF items. Details of the results are presented below:

Based on the information reported in Table 2, it is observed that 12 (12%), out of 100 items, were flagged for DIF. However, a safe interpretation of the results cannot be warranted without considering the magnitude of effect size. If this magnitude is not taken, the results may lead to inflated Type I error (e.g., Hauger & Sireci, 2008). The items were classified as showing negligible DIF ($R2 < 0.035$), moderate DIF ($0.035 \leq R2 < 0.070$), or large DIF ($R2 \geq 0.070$), according to the criteria recommended by Jodoin and Gierl (2001). As displayed in Table 2, it is observed that all the obtained R2 values manifested a negligible DIF magnitude (category A), that is, they were lower than .035. However, when the DIF magnitude of flagged items was compared against the odds ratio criteria recommended by Dorans and Holland (1993), four items were flagged for nonnegligible DIF (3 uniform & 1 nonuniform). This finding is in keeping with previous research (Monahan et al., 2007), when they followed the same procedure by using the odds ratio criteria recommended by Dorans & Holland (1993):

Table 2. *Uniform and Nonuniform DIF and Effect Results*

			*R*2 Effect Size						
Items	Subtest	Favored	UDIF	NUDIF	DIF	ϰ2	Category *R*2	Odds Ratio	Category *OR*
20	L	M	....	.005	.005	4.488	A	.880	A
23	L	F	.008	....	.008	7.953	A	.639	A
26	L	M	.012	....	.012	10.231	A	1.790	C
28	L	F	....	.005	.005	4.19	A	1.140	B
36	R	F	.008	....	.008	7.312	A	.674	A
40	R	M	....	.005	.005	4.163	A	.859	A
46	T	M	.007	....	.007	6.31	A	1.476	C
49	T	F	....	0.019	.019	12.821	A	.733	A
78	SL	F	.016	....	.016	9.706	A	.532	A
95	D	F	....	.000	.000	22.506	A	.517	C
97	S	M	.015	....	.015	11.425	A	1.899	C
99	S	M	....	.006	.006	4.431	A	.640	A

Notes. L = Linguistics; R = Research; T = Testing; SL = SLA; D = Discourse; S = Sociolinguistics; M = Male; F = Female; UDIF = Uniform DIF; NUDIF = Nonuniform DIF; A = Negligible DIF

Given that effect size is proved to be somehow influenced by large differences between focal and reference groups in terms of sample size (e.g., Jodoin and Gierl, 2001), safe interpretations on DIF detection are not warranted. The same may be true for the present study with female group (*n* = 602); being almost twice as much as the male group (*n* = 397), this casting some doubts on the interpretation of the magnitude of DIF. However, Herrera and Gomez (2008) found that differences in group sample size may not affect the interpretations on DIF results. Given the ideas are different and may be, somehow, confusing, it can be fairly safely claimed that DIF findings in this area well indicate the complexity of DIF conceptualization. As such, future studies using different DIF detection procedures are warranted to analyze the influence of equal or unequal group sample sizes on the magnitude of DIF interpretation. Given the subject matter test of ELT in Iran plays a key role in the admission of postgraduate applicants to Ph.D. programs, across the present research, replication studies are highly needed to analyze and interpret gender DIF in this context.

4.3. Suggestions for Improvement

The third research question dealt with the informants' suggestions for the betterment of the current content and the present policy of PEET. The study took benefit from the opinions of applied linguistics experts with regard to test change or test improvement. They were asked if they perceived any problems regarding the content and decision of the test and requested to propose suggestions for the betterment of such instruments. The analysis of the results gave birth to six general themes:

1. Significant role for PEEE
2. Application of knowledge
3. Academic writing
4. Collective development of questions
5. Specialized interest
6. Overall change

It is in the above order that the detailed findings are presented. The informants proposed that all the Ph.D. applicants should be informed of the purpose, potential effects, and other processes involved in the design and development of PEEE. For example, Ph.D. applicants should know in advance the weight assigned to both written and oral exams. This would definitely result in dedication of more time and attention on the part of applicants.

One more theme emerging from the respondents was that the content of PEEE should test the candidate's application of knowledge. A participant confirmed:

- T4: *I really believe in the fact that we should have essay-type items because in this way we can understand their application of knowledge.*

Another participant observed:

- T10: *The questions, well, have to be challenging questions, average and above average. Likewise, more control and supervision on the content of the test is needed.*

Furthermore, the experts acknowledged that top-tier decision makers should make a radical change in the content of this test. They proposed that, as a subtest, academic writing should be added to the content because the Ph.D. program is more research-based and students should be able to show their abilities not only in completing a doctoral dissertation but in writing some high quality papers. This is confirmed by one respondent:

- T9: *What I mean is that examinations should be oriented toward the capacities expected of the students and this makes the tests more specialized, rather than general capacities of our students or their language abilities or content abilities.*

In addition, the university professors proposed that the content of PEEE test items should be developed collectively by different applied linguistics experts from different universities. As such, it would not be biased for some groups of test-takers having some courses with some test developers from specific universities. One of the participants suggested:

- T12: *It would be better if more actually, universities, let's say, are involved in test development but now it is not the case.*

As another respondent argued:

- *If top-tier decision makers want to make some amendments to the current procedure in terms of boosting the quality of test content, they should actually listen to suggestions from other university professors.*

Another important amendment for the content of PEEE suggested by stakeholders was that the participants should respond to test items that are, somehow, based on their specialized interest so that (after they pass) they would be introduced to the universities that run those specialized courses. For example, the English Department of Shahid Chamran University (Iran) is more discourse-oriented, thus, those Ph.D. applicants who are interested in discourse and show their abilities or capacities in this regard would be introduced to that particular university for being interviewed. This would be fairer and much in line with the desires and needs of Ph.D. students in terms of their specialty in their own areas of interest. This would lead to fairer decisions or brings the task of decision-making to a sort of even-handedness. The following is a critical suggestion from the participants in this regard:

- T7: *If we construct our test according to the capacities that we expect of students to enter our department and those students who wish to participate in our program should know in advance that our department is discourse oriented, so consciously they decide whether they join us or not.*

Adapting the content of PEEE to the capacities and interests of Ph.D. applicants is welcomed by participants. One of the professors confided:

- T3: *Examinations should be oriented toward the capacities expected of the students and it makes the test more specialized, rather than testing general capacities.*

One of the participants also pointed out:

- T5: *I don't whole heartedly reject the status quo of the testing well, while, at the same time, opt for well suggestions for the betterment of these modes of testing.*

As such, it can be claimed that the current procedure is far from being perfect, but it still needs improvement:

- T8: *The idea of MC is good but its quality should be improved.*

Another one added:

- T12: *I think, actually, the structure of the questions should be changed. We should have a good mc test, enough number of items, items, let's say of high quality.*

Thus, if the quality of PEEE is appropriately improved, decision-making will be enhanced accordingly.

These proposed suggestions are expected to be taken into account by test users within and beyond testing agencies by practicing the reiterative investigation of the design validity, scoring validity, content validity, and consequential validity of the present Ph.D. Entrance Exams in Iran. We hope future studies can reveal further problems and continue to provide more insights for the improvement of the present content and current policy of PEEE.

5. Conclusion and Implications

The present study aimed to interpret the validity of PEEE within the domain definition and evaluation inferences. The results provide disconfirming evidence of the PEEE test validity; the assumptions proposed for domain definition and evaluation inferences were rebutted by counter evidence of test validity. The content of the test was not fully represented in the postgraduate revised syllabus and Ph.D. course objectives. Statistical characteristics of the test also referred to the unreliability of the test items, as well as some items flagged for DIF.

Such findings highlight several important implications for test developers and test users (i.e., policymakers): As far as the personal and social ramifications of PEEE are concerned, test developers and test users need to highly examine the content of the present gender DIF findings. For example, after reviewing the problematic items (e.g., those flagged for DIF), test developers can revise and improve the content of such items in future practices. Consequently, they can exercise care in fair test practice by dedicating a bona fide effort to develop and produce more unbiased tests and make more quality decisions on test scores accordingly. Another important insight the present study can provide for policymakers (i.e., NOET and MSRT) is related to the appropriate decisions-making process. For instance, the present findings revealed that most interview participants believed in a collective judgment as a basis for decision-making in Ph.D. admittance and evaluation processes. This may be a starting point for test users and test writers.

Further, it is felt that the present study can provide insights into the relative strengths and weaknesses of Ph.D. centralized exams in Iran. This may possibly identify some avenues for the betterment of the present decisions and current policies of such large-scale assessments and may result in some actions by policymakers.

Low academic achievement in postgraduate programs can also be associated with the problematic technical quality and social inappropriacy of the screening instrument (Kiany et al., 2013). In the case of present study, the findings revealed such inadequacies and may, accordingly, radiate such potential problems to policymakers and test practitioners to appropriately design, develop, administer, score, and interpret the Ph.D. Entrance Exam, and may motivate these responsible agencies to ensure that those who are responsible for advancing the technical quality of this test have the qualifications and competencies to design it. Likewise, they can improve the quality of such large-scale assessments and identify problems associated with the current system of Ph.D. programs and seek ways to rectify them accordingly. This can promote the validity of interpretations and decisions across educational and psychological assessments, as far as testing organizations in Iran are concerned.

As is often the case for nearly all studies focusing on validity issues, the current study is associated with some limitations: One significant limitation was full access to first-semester Ph.D. candidates' performance data on the written exam, the oral interviews, and the Ph.D. courses. In addition to the written exam, other factors like the applicants' performance in the interview session, their educational and research background, and their GPA scores may have their own effects on the admission processes of Ph.D. evaluation in Iran. Having accessed to these types of information, we could have added to the forcefulness of validity interpretations. Therefore, more research is urgently needed to investigate these areas, as well.

Funding

The authors received no direct funding for this research.

References

Alderson, J. C. (1986). Testing English for specific purposes: How specific can we get? *ELT Documents, 127*, 16-28.

Armstrong, W. B. (2000). The association among student success in courses, placement test scores, student background data, and instructor grading practices. *Community College Journal of Research and Practice*, *24*(8), 681-695.

Azmoon.Net. (2014). *Ph.D. entrance examination news.* Retrieved October 15, 2014, from the World Wide Web: www.Phd.Azmoon.Net.www.PhD Test

Bachman, L. F. (1990). *Fundamental considerations in language testing*. UK: Oxford University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal, 2,* 1-34.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, UK: Oxford University Press.

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.

Butler, F. A., Lord, C., Stevens, R., Borrego, M., & Bailey, A. L. (2004). *An approach to operationalizing academic language for language test development purposes: Evidence from fifth-grade science and math.* CSE Report 626. US Department of Education.

Chappelle, C. A., Enright, M. K., & Jamieson, J. (2008). *Building a validity argument for the test of English as a foreign language*. New York. Routledge.

Chappelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement*: *Issues and Practice, 29*(1), 3-13.

Cizek, G.J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*(1), 31-43.

Cubilo, J. (2014). Argument-based validity in classroom and program contexts: applications and Considerations. *Shiken Research Bulletin, 18*(1), 18-24.

Cheng, L., & Sun, Y. (2015). Interpreting the impact of the Ontario secondary school literacy test on second language students within an argument-based validation framework. *Language Assessment Quarterly, 12*, 50-66

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.

Farhady, H. (1998). A critical review of the English section of the B.A. and M.A. University Entrance Examination. In the *Proceedings of the conference on M.A. tests in Iran* (1998). Ministry of Culture and Higher Education, Center for Educational Evaluation. Tehran, Iran.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement*, *67*, 373-393.

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

Goldman, S. R. (2004). Cognitive aspects of constructing meaning through and across multiple texts. In N. Shuart-Faris & D. Bloome (Eds.), *Uses of intertextuality in classroom and educational research* (pp. 317-351). Greenwich, CT: Information Age.

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

Hamavandy, M. (2014). *Validation of a high-stakes test of English in Iran in comparison with TOEFL and IELTS: An assessment use argument approach*. Unpublished doctoral dissertation, Department of English, Tarbiat Modares University, Tehran, Iran.

Hauger, J. B., & Sireci, S. G. (2008). Detecting differential item functioning across examinees tested in their dominant language and examinees tested in a second language. *International Journal of Testing, 8*, 237-250.

Herrera, A., N., & Gomez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel-Haenszel and logistic regression techniques. *Quality & Quantity, 42*, 739-755.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Springer, 103*, 219-230.

James, C. L., & Templeman, E. (2009). A case for faculty involvement in EAP placement testing. *TESL Canada Journal, 26*(2), 82-99.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating power and type I error rates using an effect size with the logistic regression procedure for DIF. *Applied Measurement in Education, 14,* 329-349.

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. 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

Kheirzade, Sh. (2015). *Fairness in a validity argument: The case of the General English section of the Ph.D. Entrance Exam for non-English majors in Iran.* Unpublished doctoral dissertation, Department of English, Al Zahra University, Tehran.

Kiany, R., Shayestefar, P., Ghafar Samar, R., & Akbari, R. (2013). High-rank stakeholders' perspectives on high-stakes university entrance examinations reform: Priorities and problems. *Higher Education, 65*, 325-340.

Knoch, U., & Elder, C. (2013). A framework for validating postentry language assessments (PELAs). *Papers in Language Testing and Assessment, 2*(2), 48-66.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.; pp. 13-100). Washington, DC: American Council on Education.

Motamedi, A. (2006). The effect of university entrance examination on general health, self-esteem and psychic disorders symptom of those who were not admitted to the university. *Quarterly Journal of Research and Planning in Higher Education,12*(2), 54-72.

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32,* 92-109.

Moss, P. A. (2007). Reconstructing validity. *Educational Researcher, 36*(8), 470-476.

Murphy, S., & Yancey, K. B. (2008). Construct and consequence: Validity in writing assessment. In C. Bazerman (Ed.), *Handbook of research on writing: History, society, school, individual, text* (pp. 365-385) New York: Lawrence Erlbaum Associates.

NOET. (2013). *Ph.D. entrance examination news.* Retrieved December 20, 2013, from the World Wide Web: http://www.eao.ir/eao/FullStory.aspx?Gid=1&id=730.WWW.Sanjesh Organization

Ryan, K. (2002). Assessment validation in the context of high-stakes assessment. *Educational Measurement: Issues and Practice, 21*(1), 7-15.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Span, M. (2006). Test and item specifications development. *Language Assessment Quarterly: An International Journal, 3*(1), 71-79.

Stiggins, J. R. (1990). Toward a relevant classroom assessment research agenda. *Alberta Journal of Educational Research*, *36*(1), 92-97.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Taylor, L. (2007). The impact of joint-funded research studies on the IELTS writing module. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 20-48). Cambridge, MA: Harvard University Press.

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

White, E. M. (1990). Language and reality in writing assessment. *College Composition and Communication*, *41*(2), 187-200.

Williams, K. L. (1990). Three new tests for overseas students entering postgraduate and vocational training courses. *ELT Journal*, *44*(1), 55-65.

Xi, X. (2008). Methods of test validation. In N. H. Hornberger (Ed.), *Encyclopedia of language and education* (pp. 2316-2335). Boston, MA: Springer.

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.