International Journal of Assessment Tools in Education
2019, Vol. 6, No. 3, 330–343
https://dx.doi.org/10.21449/ijate.546709
Published at http://www.ijate.net http://dergipark.org.tr/ijate *Research Article*

On the Statistical and Heuristic Difficulty Estimates of a High Stakes Test in Iran

Ali Darabi Bazvand [1,*], **Shiela Kheirzadeh** [2], **Alireza Ahmadi** [3]

[1] University of Human Development, College of Languages, Department of English Language, Sulaimani, Iraq
[2] Sobhe-Sadegh Institute of Higher Education, Isfahan, Iran
[3] Shiraz University, Faculty of Literature and Humanities, Department of English Language, Shiraz, Iran

ARTICLE HISTORY

Received: 29 March 2019
Revised: 12 May 2019
Accepted: 09 July 2019

KEYWORDS

Classical true score theory,
Heuristic difficulty,
High stakes test,
Item response theory,
Statistical difficulty

Abstract: The findings of previous research into the compatibility of stakeholders' perceptions with statistical estimations of item difficulty are not seemingly consistent. Furthermore, most research shows that teachers' estimation of item difficulty is not reliable since they tend to overestimate the difficulty of easy items and underestimate the difficulty of difficult items. Therefore, the present study aims to analyze a high stakes test in terms of heuristic (test takers' standpoint) and statistical difficulty (CTT and IRT) and investigate the extent to which the findings from the two perspectives converge. Results indicate that, 1) the whole test along with its sub-tests is difficult which might lead to test invalidity; 2) the respondents' ratings of the total test in terms of difficulty level are almost convergent with the difficulty values indicated by IRT and CTT, except for the two subtests where students underestimated the difficulty values, and 3) CTT difficulty estimates are convergent with IRT difficulty estimates. Therefore, it can be concluded that students' perceptions of item difficulty might be a better estimate of test difficulty and a combination of test takers' perceptions and statistical difficulty might provide a better picture of item difficulty in assessment contexts.

1. INTRODUCTION

To enhance the quality of educational systems, assessment is gradually taking the central role in the higher education process (Brown & Glasner, 1999). As a result, increasing attention has been paid to the academic standards with regard to the association between the students' entry level and the outcomes of the assessment (van de Watering & van der Rijt, 2006). However, as stated by van de Watering and van der Rijt (2006), "little is known about the degree to which assessments in higher education are correctly aimed at the students' levels of competence" (p. 134). This might have happened due to the obscured correspondence between test intentions and test effects (e.g., Cizek, 2012; Hubley & Zumbo, 2011; Xi, 2008) which might be associated with two technical expressions coined by Messick (1989), "construct-irrelevant variance" (CIV) and "construct underrepresentation". The former, which might be relevant to the present

CONTACT: Ali Darabi Bazvand ✉ alidarabi1350@gmail.com Lecturer of Applied Linguistics, University of Human Development, Faculty of Arts and Humanities, Department of English Language, Sulaimani, Iraq

ISSN-e: 2148-7456 /© IJATE 2019

study, occurs when the measure does not reflect the construct to be assessed; rather additional characteristics affect performance, while the latter happens when the measure fails to include important aspects of the construct (Cizek, 2012; Hubley & Zumbo, 2011; Knoch & Elder, 2013; Xi, 2008).

As one type of CIV, the difficulty level of the test items might affect test applicants' performance and hinder them from achieving the best level of their abilities. This would possibly make the test not to tap into the construct being measured and might render it unreliable and invalid. Therefore, research undertaken on item difficulty and the way teachers and students perceive item difficulty is germane to the assessment issues (van de Watering & Van der Rijt, 2006).

It is worth noting that the difficulty of an assessment instrument or items included in it might decrease the reliability of the assessment in two ways. First, if the difficulty level of the items was much higher than the students' ability level, this would result in loss of concentration, anxiety, decrease of motivation, confusion, uncertainty, etc. and as a consequence, more errors happen in assessment. Second, there is always the chance of guessing while answering test items, especially in multiple-choice tests. So, if the items are more difficult, this implies more students would guess and this allows more random errors to enter the variance of the assessment score (Bereby-Meijer, Meijer, & Flascher, 2002).

Moreover, in line with Messick's technical expressions of test invalidity, it is reported then that the difficulty level of test items seems to be an overriding factor in contributing to the test lapses and might create a mismatch between test score interpretation and test score use (e.g, Chappelle, Enright, & Jamison, 2010; Johnson & Riazi, 2013). Such a factor, more often than not, is considered to be a major cause for confusion, anxiety, uncertainty, and demotivation among test takers, and might subsequently motivate them to rely on guessing (Stanley, 1971).

In general, across the content of PhD entrance exams in Iran, it is assumed that such test lapses might exist which might be symptomatic of the test invalidity. Therefore, investigating the difficulty level of the test, by getting insight from stakeholders' perceptions (test takers' perspective in the case of the present study) and statistical quality of the test, as analyzed via CTT and IRT, is relevant. Considering the abovementioned points, the present study aimed at estimating the difficulty level of PhD Entrance Exam of ELT (PEEE, henceforth), a high stakes test in Iran, by taking both statistical and heuristic difficulty estimates, and whether the difficulty information yielded by both stakeholders' perception and statistical analyses converge.

1.1. Classical Test Theory (CTT)

Since the early 20th century, CTT has been used in estimating test/item difficulty. In relation to this theory, the knowledge/ability (represented by the true score of the test-takers) is defined as the expected score obtained by a student in a given test (Conejo, Guzmán, Perez-De-La-Cruz, & Barros, 2014).

The major assumptions underlying the CTT are: the mean of the test-takers' error score is zero; true scores and error scores are not correlated, and error scores obtained on the parallel tests are not correlated (Hambleton & Jones, 1993). According to Magno (2009), the assumption of classical test theory is that each test taker's score is a true score (unobservable) obtained if there were no errors in measurement. However, because the test instruments used are not perfect, the observed score of each test-taker might differ from his true ability.

In this theory, items are described by two parameters: the difficulty parameter, i.e., the proportion of the students who answered an item correctly, and the discrimination parameter, which will be estimated by the correlation between the item and the test score. As an early approach to estimate test/item difficulty, it suffers certain limitations such a considering all

errors as random (Bachman, 1990); however, CTT is easy to use in several situations and it requires fewer number of testees, compared with other methods such as IRT.

1.2. Item Response Theory (IRT)

Item response theory is a probabilistic model that is to explain an individual's response to an item (Hambleton, Swaminathan, & Rogers, 1991). This theory is based on two main principles: (a) students' performance in a test would be explained by their level of knowledge, measured as an unknown numeric value h. (b) the students' performance estimated by the level of knowledge in answering an item would be probabilistically predicted and displayed using a function called the Item Characteristic Curve (ICC) (Hambleton, Swaminathan, & Rogers, 1991).

According to Birnbaum, (1968), there are three different models of IRT, namely one-parameter logistic model, two-parameter logistic model, and three-parameter logistics model. The one-parameter logistic model indicates the probability of a correct response as a logistic distribution where items differ merely regarding their difficulty and this model is used on multiple-choice (MC) or short response items which are dichotomous and do not allow for guessing (Birnbaum, 1968). The two-parameter logistic model, as stated by the same author, generalizes the one-parameter logistic model and allows items to differ not only regarding their difficulty but also differ in discriminating among individuals of various proficiency levels. Similar to the one-parameter logistic model, the two-parameter logistic model assumes that the probability of guessing is zero. Birnbaum also stated the three-parameter logistic model extends the two-parameter logistic model by including a guessing parameter which represents the probability of testees with low ability level correctly answer an item since for low ability testees, guessing is an influential factor in test performance.

To estimate item difficulty, the one-parameter IRT model using a single item parameter (i.e., difficulty parameter) is more frequently used (Van der Linden & Hambleton, 1997). The one-parameter model designates the probability of answering an item correctly through a logistic function indicating the difference between the proficiency level and the item difficulty. In justifying IRT use for difficulty estimation, Pardos and Heffernan (2011) stated, "Models like IRT that take into account item difficulty are strong at prediction" (p. 2). It should be mentioned that the one-parameter IRT model was used for the present study since the aim was merely estimating the difficulty of the items.

1.3. Local context

Since the evidence of item difficulty for the present study is provided by the PhD Entrance exam in Iran, it seems imperative to briefly introduce it here. High stakes tests in Iran have been considered as predominate tools to measure applicants' general and domain-specific knowledge and skills for the purpose of admission to higher education. Nevertheless, empirical studies have found that such tests, as levers of entering higher education, have fallen short of their expectations. That is, they have not been without their fair share of negative consequences (Farhady, 1998; Razmjoo, 2006). Specifically, findings from validity studies have shown that university entrance examinations in Iran are not socially responsive for graduate studies (Hajforoush, 2002; Shojaee & Gholipour, 2005).

As part of university entrance examinations, PhD entrance exams in Iran play a great role in the admission decisions of postgraduate studies. These high-stakes exams consist of a series of centralized written exams designed to screen PhD applicants (with different academic majors) to enter PhD programs. Since 2011, these exams superseded the traditional university-based examination sets in Iran. Administered by the National Organization for Educational Testing (NOET), they all appear in MC format with four-option items often consisting of three blocks: a general competence section, an academic talent test, and a field-specific section. For this

study, the field-specific section of the PhD exam of ELT administered in 2014 was considered. More information on this exam is provided in the method section.

2. PREVIOUS STUDIES ON TEST DIFFICULTY

Previous research has highlighted various factors that might influence item difficulty, for instance, word knowledge (e.g., Rupp, Garcia, & Jamieson, 2001), negative stem (e.g., Hambleton & Jirka, 2006) and background knowledge of the topic (e.g., Freedle & Kostin, 1999). The purpose of conducting such studies was to make the item-writing process more efficient through academically publishing more detailed guidelines and item level descriptors to help item writers (Kostin, 2004). However, besides sensitivity to guidelines and item descriptors, in Bachman's (2002) words, "difficulty does not reside in the task alone but is relative to any given test-taker" (p. 462). Therefore, who would take the test and answer items would definitely influence the way items are designed and developed. Hambleton and Jirka (2006) recommended asking experts in the field of test development and scoring to estimate the task difficulty. However, even these experts are not necessarily accurate in their predictions of task difficulty in both first language (L1) (Bejar, 1983; Hambleton & Jirka, 2006) and second language (L2) tests (Bachman, 2002; Elder, Iwashita, & McNamara, 2002).

Bejar (1983) concluded that a group of four test developers could not make a reliable difficulty estimation for L1 writing tasks. As with L2 tests, Alderson (1993) reported that experienced item writers and raters were somewhat better than inexperienced ones on predicting item difficulty; however, the significance of this difference was not estimated. Hamp-Lyons and Mathias (1994) reported a considerable agreement between expert judges (two raters familiar with the test and two L2 writing experts); however, there was an astonishingly reverse relationship between the difficulty predicted by experts and raters and the actual difficulty of the test. Therefore, as suggested by Lee (1996), students might be able to estimate difficulty more accurately than teachers. Nevertheless, teachers/experts' estimation has received more attention than students' estimation.

Wauters, Desmet, and van Den Noortgate (2012, p. 1183) compared six different estimations of the difficulty: "proportion correct, learner feedback, expert rating, one-to-many comparison (learner), one-to-many comparison (expert) and the Elo rating system" with the IRT-based calibration. Results revealed that proportion correct showed the strongest relation with IRT-based difficulty estimates, followed by student estimation. The participants of the study included 13 teachers and 318 students (secondary level) in the field of Linguistic and Literature. The researchers concluded that student estimations were somewhat better. To explain the difference in the rating of the two groups of the participants, the researchers referred to the much larger sample size of the students, compared to the teachers.

In a more recent study, Conejo, Guzmán, Perez-De-La-Cruz, and Barros (2014, p. 594-595) named three test/task difficulty estimation approaches.

- Statistical, that is, estimating the difficulty from a previous sample of students.
- Heuristic, that is, by human ''experts'' direct estimation.
- Mathematical, given a formula that predicts the difficulty in terms of the number and type of concepts involved in the task

To estimate difficulty using statistical approaches, the definition of the concept of difficulty need to exist. Therefore, this approach is commonly associated with using CTT or IRT in the assessment. From the heuristic standpoint, teachers or course designers are commonly experts that estimate the difficulty; however, students might also be considered as experts in this approach. In mathematical approaches, difficulty would be estimated by a formula that uses a number of item/task features, e.g., complexity or the number of concepts involved. As such, the

focus of the present study is to estimate the statistical (CTT and IRT estimations) and heuristic (test-takers' standpoint) difficulty of test items and investigate the extent to which findings from the two perspectives are congruent.

3. METHOD

3.1. Participants

The participants in the current study included PhD applicants and first semester PhD candidates of Iran majoring in ELT. Test score data for a population of 999 PhD exam applicants (397 females and 602 males) participating in January 2011 administration of this test was analyzed in terms of the difficulty level. Performance data for this population was provided by the National Organization for Educational Testing (NOET) at the request of Shiraz University, Iran. No information regarding their age, names, average score, and the socioeconomic status was provided by this organization.

The second group of participants was a sample of 103 PhD candidates of ELT who had been admitted to the PhD programs. Their ages ranged between 25 and 40, with 46 of them being female and 57 of them being male. They were recruited to respond to the survey questionnaires. Since it was not feasible to obtain a complete list of all the participants from whom to make a random selection, a snowball sampling procedure was preferred. This particular sample was targeted, since tracking them to administer the questionnaire was less likely to be problematic. In addition, they were in a better position to recollect their test-taking experience than those who had taken it earlier. They received the questionnaires through email. Upon their views, they provided evidence with regard to the characteristics of PEEE in terms of its difficulty level. A brief summary of the participants' self-reported background is provided in Table 1.

Table 1. Background Information Reported by PhD Students (n=103)

Variable	Level	F (%)
Gender	Male	57(55.3%)
	Female	46(44.7%)
	Total	103(100%)
Age	25-27	13(12.6%)
	28-30	38(36.9%)
	30-39	41(39.8%)
	40+	11(10.7%)
Times taking exam	First	10(9.7%)
	Second	60(58.3%)
	Third	24(23.3%)
	Fourth	9(8.7%)
Field-specific test scores	Less than 30%	5(4.9%)
	30-40%	31(30.1%)
	40-50%	48(46.6%)
	50+	42(40.8%)
General English test scores	Less than 30%	6(5.8%)
	30-40%	24(23.3%)
	40-50%	31(30.1%)
	50+	42(40.8%)

3.2. Instruments and Data Collection

Two types of instruments were used to collect the data for this study, namely PEEE test score data and PhD students' questionnaires. PEEE is a field-specific exam which is aimed at measuring the PhD candidates' expertise in the field of English Language Teaching (ELT) and is supposed to be related to the courses students have passed in the MA or even BA program.

In fact, it assesses the students' domain-specific knowledge in areas which are the prerequisite for entering the PhD programs since the PhD program is built on such areas of knowledge. It consists of 100 items including questions on Linguistics (15 items), Teaching Methodology (15 items), Research Methods (15 items), Language Testing and assessment (15 items), Theories of SLA (30 items), and finally Discourse & Sociolinguistics (10 items).

PhD students' questionnaire comprised of 24 items and categorized into two parts to provide information on students' background and their perceptions with regard to test characteristics. For test characteristics part, the options included *very difficult*, *difficult*, *average*, *easy* and *very easy*. The reliability of the whole questionnaire was reported to be .73, as estimated through Cronbach's alpha. The validity of questionnaires was established using expert judgment.

3.3. Data Analysis

For the data analysis, both questionnaire and test score data were analyzed. With regard to the questionnaire, stakeholders' perceptions were analyzed for the difficulty level of the test. For this reason, 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. Concerning the PEEE test score data, CTT, IRT and Cronbach's alpha were applied to estimate the difficulty level and the reliability coefficients of the whole test and its subtests, respectively.

4. RESULTS

For investigating the difficulty level of the test, the study benefitted from heuristic analysis, i.e., stakeholders' perceptions (via questionnaire) and statistical analysis, i.e., CTT and IRT analysis. The details are explained below.

4.1. Heuristic difficulty of the items

PhD students' responses to questionnaires revealed some important findings. They, almost all, did express the same collective opinion with regard to the level of difficulty of the items. As shown in Table 2, of 103 respondents, about half of them (58%, p =.114), answered that the total test is difficult. It is also reported that some subtests like Teaching Methodology (49%) and Linguistics (46.1%) designed based on the BA courses are moderately difficult and some others like Theories of SLA (64%), Language Testing and assessment (73%), and Discourse & Sociolinguistics (63%) which were based on MA courses are reported to be significantly difficult.

Table 2. Stakeholders' Perceptions of Difficulty of PEEE

PEEE and its Subtests	Category	N	Observed Prop.	Test Prop.	Sig. (2-tailed)
Total test	Easy*	43	.42	.50	.114
	Difficult +	60	**.58**		
Linguistics	Easy*	55	.54	.50	.000
	Difficult+	48	**.46**		
Teaching Methodology	Easy*	52	.51	.50	000
	Difficult+	51	**.49**		
Theories of SLA	Easy*	37	.36	.50	.006
	Difficult+	66	**.64**		
Language Testing and assessment	Easy*	28	.27	.50	.000
	Difficult+	75	**.73**		
Research Methods	Easy *	44	.43	.50	.001
	Difficult+	59	**.57**		
Discourse & Sociolinguistics	Easy*	38	.37	.50	.010
	Difficult+	65	**.63**		

* Combined 'Easy' and 'Very easy' responses

+ Combined 'Difficult' and 'Very difficult responses

4.2. Statistical difficulty of the items

4.2.1. *CTT difficulty*

In addition to the analysis of stakeholders' perceptions with regard to the level of difficulty, the test was also subjected to statistical item analysis. In this procedure, the difficulty index (referred to as a p-value) was estimated as the proportion of examinees correctly answering each item. As such, items shown to have demonstrated values above .80 or below .40 were considered to be too easy or too difficult, respectively (Apostolou, 2010); therefore, their difficulty level is not desired. With regard to the present study, all the subtests of PEEE were subjected to item analysis.

The first specialized subtest included in the PEEE was "Teaching Methodology" consisting of 15 items. As indicated by item analysis, the results from Table 3 reveal that the difficulty level of this subtest amounts to .39 with the difficulty values of individual items ranging from .52 to .07. As it is reported, of 15 items included in this subtest, 12 items do not fall within the above criterion range of difficulty, revealing that this subtest is somehow difficult.

Table 3. Difficulty Level of the Total Test and its Subtests

Subtest	Number of items	Mean Difficulty
Total Test	100	.24
Teaching Methodology	15	**.39**
Linguistics	15	.32
Research Methods	15	.25
Language Testing and assessment	15	.15
Language Skills	10	.21
Theories of SLA	20	.19
Discourse & Sociolinguistics	10	.23

The second subtest subjected to item analysis was "Linguistics" subsisting of 15 items. With regard to this subtest, Table 3 displays that the difficulty value of the whole subtest (p =.32) is not desired. Hence it provides evidence that this subtest is difficult.

The third subtest analyzed for difficulty index was "Research Methods". Like the first two sections, this subtest consists of 15 items. As Table 3 demonstrates, the difficulty value of the whole test is .25, falling far below the acceptable estimate of the desired difficulty. This finding is also true for individual items. Of the total of 15 items analyzed, 13 of them demonstrated difficulty values lower than the least acceptable criteria of the desired difficulty, suggesting that this subtest is also difficult.

The fourth subtest subjected to item analysis was "Language Testing and Assessment". Concerning this subtest, the results from Table 3 show that with a difficulty index of .15, this subtest might have been much too difficult for the applicants. Of particular interest is that no individual item displayed a difficulty value greater than the least desired difficulty of .40; such finding reveals that this subtest is problematic and might have introduced substantial CIV into the test scores.

The fifth subtest analyzed in terms of difficulty level was "Language Skills" consisting of 10 items. With regard to test difficulty, Table 3 shows a low index of difficulty (p =.21). As for individual items, it is reported that no items demonstrated a difficulty value more than the least desired yardstick (p =.40); therefore, introducing substantial CIV in the test scores.

The sixth subtest of PEEE analyzed for item difficulty was "Theories of SLA" subsisting of 20 items. As displayed in Table 3, the difficulty index reported for the whole subtest is .19. As for

individual items, no difficulty value was reported to exceed the least acceptable criterion, showing that the test is unduly difficult.

The last area of investigation for item analysis was "Discourse & Sociolinguistics", both of which being considered as one subtest and consisting of 10 items. As it is evident in Table 3, the estimated difficulty value reported for the whole test was .23 which was far too low as measured against the least desired yardstick. Like other subtests, in this section, the difficulty indices for all of the individual items were shown to be dramatically lower than the acceptable criteria, indicating that this subtest is also very difficult. In a nutshell, the overall results from the item analysis refer to the PEEE as being substantially difficult for PhD students.

4.2.2. *IRT difficulty*

In addition to the statistical analysis of CTT and stakeholders' perceptions with regard to the level of difficulty, the test was also subjected to IRT analysis. In this procedure, the theoretical range of item difficulty falls within the range of -∞ to + ∞ on the ability scale, but in practice, the empirical range falls within the area of -2 to +2 (Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991). Items are shown to demonstrate *b* values (difficulty estimates) near -2 correspond to very easy items that are at the left or the lower end of the ability scale and items displaying *b* values near +2 are considered as very difficult that fall at the right or higher end of the ability scale. In order to have a better understanding, Baker (2001) defined the difficulty level of an item in verbal terms with their corresponding empirical ranges of *b* parameter as follows:

Table 4. Difficulty Parameter Values (from Baker, 2001, p. 12)

Verbal Label	Range of *b* values
Very easy	– 2.0 and below
Easy	– 2.0 ~ – 0.5
Medium	– 0.5 ~ + 0.5
Difficult	+ 0.5 ~ + 2.0
Very difficult	+ 2.0 and over

With regard to the present study, all the subtests of PEEE were subjected to IRT difficulty analysis. In the interest of brevity, only the difficulty values for the overall test as well as subtests are presented here. As indicated by test difficulty analysis, results from Table 5 reveal that the difficulty level of Teaching Methodology subtest amounted to 5.18. Based on the yardstick reported in Table 4, this subtest was considered *very difficult* and among the 15 items included in this subtest, 12 items fell beyond + 2.0, as the criterion range of *b* value; that is, they were *very difficult* and the remaining 3 items fell within the range of + 0.5 ~ + 2.0, being considered as *difficult*. Worthy of note is that the difficulty value for some items amounted to 10, showing that they were much beyond the ability level of examinees.

The second subtest subjected to *b* parameter analysis was Linguistics subsisting of 15 items. With regard to this subtest, Table 5 displays that the difficulty value of the whole test ($b = 4.05$) which was beyond + 2.0, demonstrated that it was *very difficult*. Regarding the individual items, 10 items were considered as *very difficult*, 2 as *difficult*, one item as *medium,* and 2 items as *easy*.

The third subtest analyzed for difficulty index was Research Methods. Like the first two sections, this subtest consisted of 15 items. As Table 5 demonstrates, the difficulty value of the whole test fell beyond + 2. This finding was also true for most of the individual items. Of the total of 15 items analyzed, 14 of them demonstrated difficulty values beyond + 2.0, and one item fell within the range of *difficult* items.

Table 5. Results of Test Difficulty Parameter in IRT Model

Subtest	Mean Difficulty	SD
Total Test	3.86*	1.70
Teaching Methodology	5.18*	2.87
Linguistics	4.05*	3.27
Research Methods	3.90*	.52
Language Testing and assessment	3.45*	.93
Language Skills	3.87*	1.30
Theories of SLA	3.41*	1.34
Discourse & Sociolinguistics	3.13*	.99

* Larger than + 2.0. (Very difficult)

The fourth subtest subjected to item analysis was Language Testing and Assessment. Concerning this subtest, the results from Table 5 show that with a b value of 3.45, this subtest was *very difficult* for the applicants. Of particular interest was that 14 items display a *b* value greater than the least value for *very difficult* items and one item covered the range of *difficult* items; this finding reveals that this subtest was substantially difficult.

The fifth subtest analyzed in terms of difficulty level was Language Skills consisting of 10 items. With regard to test difficulty, Table 5 shows a greater index of difficulty (b = 3.87) than + 2.0. As for individual items, it was found that 9 out of 10 items demonstrated a difficulty value more than the yardstick for *very difficult* items. Only one item was reported as difficult.

The sixth subtest of PEEE analyzed for test difficulty was Theories of SLA subsisting of 20 items. As displayed in Table 5, the difficulty *b* parameter reported for it was 3.41, symptomatic of *very difficult* tests. With regard to the individual items, it was found that 17 items displayed difficulty values larger than + 2.0, suggesting that they were *very difficult*, with the remaining 3 items fell under the category of *difficult* items.

The last area of investigation for item analysis is Discourse and Sociolinguistics, both of them were considered as one subtest and consisted of 10 items. As it is evident in Table 5 above, the estimated *b* value reported for this subtest was 3.13, indicating the test was *very difficult*, as measured against the yardstick of + 2.0. Like other subtests, in this section, the difficulty indices for almost all of the individual items were shown to be dramatically larger than the yardstick labeled for *very difficult* items.

Finally, as indicated in Table 5 as well as in Figure 1, the total test was shown to be *very difficult (*3.86*)*. As such, it can be argued that, based on the results from the IRT difficulty analysis, the PEEE test is prone to unreliability.

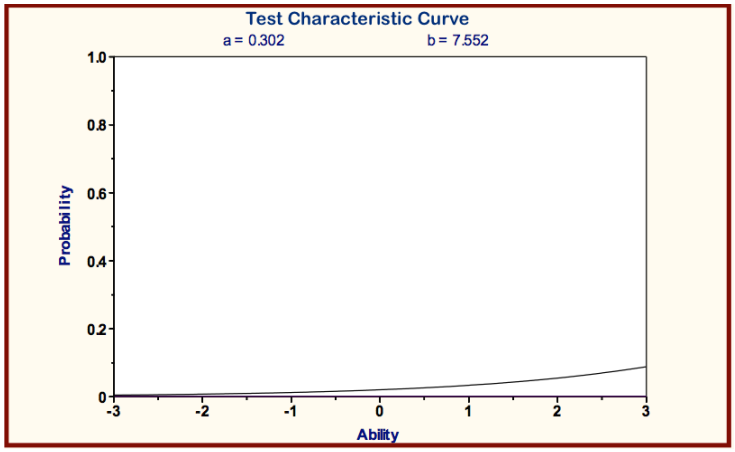

Figure 1. Total Test Difficulty: Test Characteristic Curve

4.2.3. *Comparison between heuristic and statistical difficulty*

As demonstrated in Table 6 below, results revealed that the respondents' rating of the total test in terms of difficulty level was almost convergent with the difficulty values indicated by IRT and CTT difficulty analyses, with reference to the same subtests. However, there were some specific cases of inconsistency between the results from heuristic difficulty and statistical difficulty; the results reported for the heuristic difficulty showed moderate difficulty values for Linguistics and Teaching Methodology subtests, while the findings from IRT and CTT difficulty demonstrate *very difficult* description for the same subtests. To recapitulate, when comparing the heuristic and statistical results for difficulty level, most of the subtests in the heuristic difficulty classification displayed the label of *difficult* and *very difficult*, while most of the subtests in the statistical category demonstrated the label of *very difficult*. This finding might lead to the overall conclusion that PEEE is a difficult test. As such, inappropriate test difficulty level was considered as evidence for invalidity of PEEE.

Table 6. Comparison between Heuristic and Statistical Difficulty

Test	Heuristic Difficulty	Statistical Difficulty	
	Questionnaire	CTT	IRT
Total Test	Difficult (58 %)	Very difficult (.24*)	Very difficult (3.86*)
Teaching Methodology	Moderate (49 %)	Difficult (.39*)	Very difficult (5.18*)
Linguistics	Moderate (46 %)	Difficult (.32*)	Very difficult (4.05*)
Research Methods	Difficult (57 %*)	Very difficult (.25*)	Very difficult (3.90*)
Language Testing	Very difficult (64 %*)	Very difficult (.15*)	Very difficult (3.45*)
Language Skills	Very difficult (64 %*)	Very difficult (.21*)	Very difficult (3.87*)
Theories of SLA	Very difficult (63 %*)	Very difficult (.19*)	Very difficult (3.41*)
Discourse & Sociolinguistics	Very difficult (73 %*)	Very difficult (.23*)	Very difficult (3.13*)

* Heuristic difficulty values above % 50 (difficult & very difficult)
* CTT difficulty values below 0 .40 (difficult & very difficult)
* IRT difficulty values larger than + 2.0 (very difficult)

5. DISCUSSION and CONCLUSION

The present study investigated the statistical and heuristic difficulty of PEEE in Iran. Findings of the study demonstrated that the statistical and heuristic difficulty investigations converge, indicating that the PEEE test is unduly difficult for test applicants. Results of the analysis of questionnaire items showed that for most of the PhD students (58%), the total test was very difficult.

IRT analysis of test difficulty also showed that all subtests were labeled as *very difficult* as compared with the criterion (+2.0 and beyond for very difficult items) recommended by researchers (Baker, 2001). Specifically, some items displayed values as large as 9.0, suggesting that they were much beyond the ability level of test applicants. The overall results from the *b* parameter IRT analysis of the PEEE subtests, and in most cases, their individual items together with the results from stakeholders' perceptions denote the PEEE test as very difficult. This finding can be regarded as good evidence for invalidity of this test (at least in terms of difficulty level). Moreover, the findings from the comparison between statistical and heuristic analysis showed that they were almost convergent, though there were some minor contradictions. One possible explanation might rest on the fact that, for the main part, the content of PEEE test is not based on the courses PhD applicants have passed but on those sources that they are not aware of or at least a few applicants have the chance to make use of. This could make the test difficult and might systematically introduce CIV into observed scores.

In a similar study, Rezvani and Sayyadi (2016) investigated the validity of new Iranian TEFL

PhD program entrance exam by asking PhD instructors and students. The result of their study revealed that, "the new exam was perceived to demonstrate defective face, content, predictive, and construct validities" (p. 1111). Razavipur (2014) studied the substantive and predictive validity facets of the university entrance exam for English majors by asking the ideas of 111 English major university students. He found that a large number of construct-irrelevant items exist in the exam along with a number of items that make no unique contribution to the exam. Furthermore, this finding was supported by research, though on a different testing application context. For example, in Apostolo's (2010) study, candidates' heuristic task difficulty in the KPG listening tests was found to correlate to a great extent with the results of item analysis.

The findings of the present study might be somewhat consistent with Hamp-Lyons and Mathias (1994) who reported an astonishingly reverse relationship between the difficulty predicted by experts and raters and the actual difficulty of the test. As it was the case in the present study, the present so-called standard exam turned out to be a highly difficult one both heuristically and statistically. In the words of Nickerson (1999), when one decides to assess others' knowledge and information, he requires to make a mental model of what they might know and if he has no access to specific information regarding those target group, a faulty mental model would be formed.

As such, any indiscriminate dealing with these tests regarding their interpretation and use might generate negative impacts on different stakeholders, across different testing contexts. Therefore, test practitioners should exercise high care when dealing with these gatekeeping tests in terms of item writing, test construction and test administration and also, as stated by Bachman (2002), difficulty is not just due to the tasks but it is a relative concept that varies across test-takers. As stated by Elder, Iwashita, and McNamara (2002, p. 350),

> *If test-takers can predict what makes a task difficult, it may be wise for us to access their views during the test design stage to determine whether they correspond to the hunches of test-developers and with existing theories about what makes a task more or less complex. It is conceivable that test-takers may be able to identify additional features of the task, or additional challenges involved in performing such tasks other than those visible to the test-developer or to the rater.*

Finally, the findings might be discussed from the social projection perspective, i.e., ascribing, generalizing and projecting what we know (the item developers in the case of the present study) to others (test-takers). In this regard, Nickerson (1999) stated that high familiarity with the particular topic might lead to over-ascription of what one knows to others. Also as stated by Goodwin (1999), judges (or item designers as it is the case in the present study) are typically experts in their fields. Since they might be much more knowledgeable in the related field, they might not be able to put themselves in the place of students adequately. Furthermore, their expectations of the examinees are possibly too high and they might also have difficulty differing between the proportion of examinees who should have answered an item correctly and who could have answered an item incorrectly.

In other words, the item writers might differ in their backgrounds and levels of experience with students. Item writers might tend to overestimate the performance of students. They might have based their judgments on "what they think students ought to know" (Verhoeven et al., 2002, p. 865). Such claim was supported by Impara and Plake (1998) who stated that even though judges (the expert in the field who design items/tests) are in close contact with the educational program, there is still a large difference in cognitive levels between them and the students. As it might be the case in the present study, the designers of PEEE exam, due to their familiarity with the subject matter, overgeneralized it to the test takers whose result was a very difficult test from the perspective of test takers.

Therefore, as the focus of the present study was a high-stakes test, the designers of such tests

are recommended to consider the learners' characteristics and various possible learning environments in mind while developing items since very difficult test/items result in loss of concentration, anxiety, decrease of motivation, confusion, and uncertainty on the side of the test-takers. Such implication is in line with Bachman (2000, as cited in Brindley & Slatyer, 2002) who stated that, as soon as one considers what makes items difficulty, one immediately realizes that difficulty is not a reasonable question at all. A given task or item is differentially difficult for different test takers and a given test taker will find different tasks differentially difficult. Ergo, difficulty is not a separate quality at all, but rather a function of the interaction between task characteristics and test taker characteristics. When we design a test, we can specify the task characteristics, and describe the characteristics of the test takers, but getting at the interaction is the rub. Therefore, future researchers are recommended to work on item-writing guidelines used by item writers to see if these guidelines match the expectations, needs, and requirement of the target populations taking the test, especially in the case of high-stakes tests.

ORCID

Ali Darabi Bazvand https://orcid.org/0000-0002-2620-4648
Shiela Kheirzadeh https://orcid.org/0000-0003-4665-0554

6. REFERENCES

Alderson, J. C. (1993). Judgments in language testing. In D. Douglas & C. Chapelle (eds.), *A new decade of language testing* (pp. 46–57). Arlington. VA: TESOL.

Apostolou, E. (2010). Comparing perceived and actual task and text difficulty in the assessment of listening comprehension. In *Lancaster University Postgraduate Conference in Linguistics & Language Teaching* (pp. 26-47).

Bachman, L. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*, 453–476.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford university press.

Baker, F. (2001). *The basics of item response theory.*, College Park: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland.

Bejar, I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7*, 303–310

Bereby-Meijer, Y., Meijer, J., & Flascher, O. M. (2002). Prospect theory analysis of guessing in multiple choice tests. *Journal of Behavioral Decision Making, 15*, 313–327.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord. & M. R. Novick (Eds.), *statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing, 19*, 369-394.

Brown, S., & Glasner, A. (1999). *Assessment matters in higher education*. Buckingham: SRHE and Open University Press.

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

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

Conejo, R., Guzmán, E., Perez-De-La-Cruz, J. L., & Barros, B. (2014). An empirical study on the quantitative notion of task difficulty. *Expert Systems with Applications, 41*, 594-606.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer?. *Language Testing, 19*, 347-368.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah: Erlbaum.

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

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing, 16*, 2-32.

Goodwin, L. D. (1996). Focus on quantitative methods: Determining cut-off scores. *Research in Nursing & Health, 19,* 249–256.

Hajforoush, H. (2002). Negative consequences of entrance exams on instructional objectives and a proposal for removing them. Proceedings of *the Isfahan University Conference on Evaluating the Issues of the Entrance Exams*.

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on: Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*, 38-47.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hambleton, R., & Jirka, S. (2006). Anchor-based methods for judgmentally estimating item statistics. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 399–420). Mahwah, NJ: Erlbaum.

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing, 3*, 49–68.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*, 219.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35*, 69–81.

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*, 31-58

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

Kostin, I. (2004). *Exploring item characteristics that are related to difficulty of TOEFL dialogue items* (TOEFL Research Rep. No. 79). Princeton, NJ: ETS.

Lee, F. L. (1996). *Electronic homework: an intelligent tutoring system in mathematics*. (Doctoral Dissertation). The Chinese University of Hong Kong. Hong Kong, China.

Lee, F. L., & Heyworth, R. M. (2000). Problem complexity: a measure of problem difficulty in algebra by using computer. *Education Journal, 28*, 85–107.

Magno, C. (2009). Demonstrating the difference between Classical Test Theory and Item Response Theory using derived test data. *The International Journal of Educational and Psychological Assessment, 1,* 1-11.

Nickerson, R. S. (1999). How we know-and sometimes misjudge-what others know: Imputing one's own knowledge to others. *Psychological Bulletin, 125*, 737–759.

Pardos, Z. A., & Heffernan, N. T. (2011). KT-IDEM: Introducing item difficulty to the knowledge tracing model. In J. Konstan, R. Conejo, J. L. Marzo, & N. Oliver (Eds.), Proceedings of the *19th international conference on user modeling, adaptation and personalization* (Vol. 6787, pp. 243–254). Lecture Notes in Computer Science.

Razavipur, K. (2014). On the substantive and predictive validity facets of the university entrance exam for English majors. *Research in Applied Linguistics*, *5*, 77-90.

Razmjoo, S. A. (2006). A content analysis of university entrance examination for English majors in 1382. *Journal of Social Sciences and Humanities, Shiraz University, 46,* 67-75.

Rezvani, R., & Sayyadi, A. (2016). Ph. D. instructors' and students' insights into the validity of the new Iranian TEFL Ph. D. program Entrance Exam. *Theory and Practice in Language Studies*, *6*, 1111-1120.

Rupp, A. A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing*, *1*, 185-216.

Shojaee, M. & Gholipoor, R. (2005). *Recommended draft of applying university student system survey and designing acceptance model of university student*. Research Center of the Parliamnet, No. 7624.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 356-442). Washington, DC: American Council on Education

van de Watering, G., & van der Rijt, J. (2006). Teachers' and students' perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. *Educational Research Review*, *1*, 133-147.

van der Linden, W., & Hambleton, R.K. (1996). Item response theory: Brief history, common models, and extensions. In W. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item-response theory* (pp. 1–28). Berlin: Springer-Verlag.

Verhoeven, B. H., Verwijnen, G. M., Muijtjens, A. M. M., Scherpbier, A. J. J. A., & Van der Vleuten, C. P. M. (2002). Panel expertise for an Angoff standard setting procedure in progress testing: Item writers compared to recently graduated students. *Medical Education, 36*, 860–867.

Wauters, K., Desmet, P., & van Den Noortgate, W. (2012). Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education, 58*, 1183–1193.

Xi, X. (2008). Methods of test validation. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of Language and Education, 2nd edn, vol. 7: Language testing and assessment* (pp. 177–196). New York: Springer.