On the Predictive Validity (Generalizability) of IELTS Academic Tests

Goodarz Shakibaei[1]

Department of Humanities, College of English language, Ahvaz Azad University Branch, Islamic Azad University, Ahvaz, Iran

Mehran Memari

Farhangian Teacher Education University, Iran

Abstract

Predictive validity of high-stake tests has attracted the attention of applied linguists since a couple of decades ago. Despite the existence of a plethora of studies on reliability and validity issues of high stake tests, it seems that predictive validity of International English Language Testing System (IELTS), as a high stake test should be measured appropriately. This study set out to investigate the predictive validity of IELTS academic tests practiced in Iran. The participants included 70Iranian IELTS test takers who took IELTS academic test in 2014 to 2015. To carry out the study, Construct and Validation Self-Assessment Inventory for Learners of English for Academic Purposes was used to collect data. The questionnaire included 48items on a five-point Likert scale. Coronbach's alpha coefficient of this questionnaire indicated the reliability of the instrument. The test-takers' scores on the four skills of academic IELTS were also used as another portion of data in this study. The collected data were subjected to a series of statistical analyses in relation to the proposed questions in this study. The data were analyzed using Regression and Analysis of Variance. The results of the study revealed that IELTS four modules significantly predict the test takers' performance in academic settings. The findings can be used by IELTS centers to make a change in test format, rubrics, as well as uses.

1. Introduction

The entrance of students who want to continue their studying abroad into the

universities where the medium of studying is English is on the ascent. Hence,

registration directions of colleges and organizations of higher education in

universities where the medium of instruction is English require nonnative

---

[1] Corresponding author: Shakibaeienglish@gmail.com

students to meet the English language registration prerequisite. Aside from some national and local created English language proficiency tests, two famous tests (i.e., the International English Language Testing System (IELTS Academic), and Test of English as a Foreign Language (TOEFL) Internet-based Test (iBT) are taken by worldwide students to satisfy tertiary universities' language registration necessities.

Most institutions and colleges accredit both TOEFL and IELTS as a criterion of entry requirement. Most North American colleges accept TOEFL certificate as an entry and registration requirement, though IELTS is more prominent in European and Asia Pacific institutions and colleges. However, IELTS Academic is the first vital important requirement, which must be met by students from different nations who apply to Australian colleges for higher education. Both international tests (i.e., IELTS and TOEFL-iBT) measure the test-takers' English language ability in the four modules, listening, speaking, reading and writing. IELTS Academic reports test-takers' scores utilizing a band score of 0– 9 for each module and an average of four modules, while a scale of 0–30 and an overall maximum score of 120 are used in TOEFL-iBT reports of candidates' performance.

Moore and Morton (1999) compared IELTS writing task items with 155 assignments given in two Australian universities. They found that IELTS task 1 was representative of the target language use content, while IELTS task 2, which requires students to agree or disagree with the proposition, did not match exactly with any of the academic genres in the target language use domain as the university writing corpus was based on external sources as opposed to IELTS task 2, which was based on prior knowledge as a source of information. IELTS task 2 was more similar to nonacademic public forms of discourse such as letter to the editor; however, IELTS task 2 could also be

considered close to the genre essay, which was the most common of the university tasks.

Uysal (2010) reviewed the IELTS writing test. Her study aimed to provide a descriptive and critical review of the IELTS writing test by focusing particularly on various reliability issues such as single marking of papers, readability of prompts, and validity issues such as the definition of the international writing construct, without considering variations among rhetorical conventions and genres around the world. She also discussed the validity-impact issues of writing module. She concluded that in terms of improving the validation efforts with regard to the IELTS writing module, future research should be performed to explore whether the characteristics of the IELTS test tasks and the TLU tasks match, not only in the domain of UK and Australia, but also in other domains.

Test makers and test users mainly concern about validity of high-stakes English proficiency tests such as IELTS academic tests. Modification of the tests by test makers and making better decisions by test users are two recognized outcomes of the evidence provided through empirical studies. The studies done during the development of the PTE academic tests, gathering, and analyzing the related data are the evidence for validity of PTE Academic test.

There are two categories of studies in this field, the studies done by test developers and independent researchers. As one of the main reports provided by test organization, Pearson (2010) gathered the data from the first developing stages of the test and benchmarking it to the Council of Europe Frame of Reference or CEFR (2001). The levels of CEFR have been used for assessing in various settings since it was developed. The participants of second research were both native and nonnative test-takers who participated in two rounds of field tests (August–October 2007) and (May–June 2008).

The collected data on PTE were analyzed and the conformity studies of PET Academic scores with international tests such as TOEFL and IELTS scores were done (Zheng & De Jong, 2011).

The international English language testing system (IELTS) is the world's most popular English language proficiency test. IELTS assesses all of test takers' English skills – reading, writing, listening and speaking. There are two versions of IELTS; IELTS Academic and IELTS General Training. IELTS is designed to assess English language skills at all levels using a 9-band scale.

One of the fundamental qualities of IELTS is predictive validity. The term predictive validity can be described as the extent to which the results can be made general from the research sample to the whole population (Polit & Hungler, 1991). Predictive validity refers to making forecasts used based on a repeated experience. If something happens repeatedly, it is predicted that it will continue to have the same thing in following days. Investigators employ similar kind of logic when making general about the results of their studies. Once researchers have collected sufficient data to support a hypothesis, a premise regarding the behavior of that data can be formulated, making it generalizable to similar circumstances. In other words, it is the degree to which the findings of a research are applied to the individuals and circumstances beyond those studied.

When the alignment of learning standards and assessments is sound, then the extent to which one test score can predict another is elevated. The relationship between the two test scores can be called predictive or criterion validity (Bachman & Palmer, 1996). Predictive validity involves testing a group of subjects for a certain construct, and then comparing them with results obtained at some point in the future. Most educational and employment tests are used to predict future performance, this way, predictive validity is regarded as essential in these fields (Shuttleworth, 2009).

Due to the importance of making appropriate prediction and inference about test takers' performance in their academic/specific target language use based on their scores on an Administered IELTS, in this study the predictive validity is considered as one necessary part of construct validity in IELTS academic tests which needs further appropriate explorations. Therefore, the following questions are formulated:

1. How does the IELTS test takers' speaking module correlate to their speaking performance in an academic setting?
2. What is the significant relationship between the IELTS test takers' writing module and their writing performance at an academic level?
3. What is the correlation between the IELTS test takers' listening module and their listening comprehension in an academic context?
4. How is the IELTS test takers' reading module significantly correlated to their reading comprehension in an academic setting?

2. Review of the Literature

Bachman and Palmer (1996) and Bachman (1990) assisted to the definition of construct through their framework of language ability. A new term and manner of conceiving validity (Chapelle, 2012) was also introduced by them which is test usefulness. The evaluation of the test based on the use for which it is initially prepared is done through test usefulness. Predictive validity deals with the extent to which inferences made of a test predict the next performance on a course or a particular task. It is usually related to proficiency tests like IELTS and TOEFL. The test is usually given to the candidates in their home country, then as they arrive in the host country they are given a similar test, and the relationship between the findings on both tests would indicate the degree of predictive validity. Nevertheless, this type of validation has been always criticized for what is known as the problem of truncated sample because of the fact that candidates who fail the test in their home country do not sit for the equivalent test in the host country, which

increases the spread of students' scores and has negative effects on the validity coefficient (Alderson, Clapham, & Wall, 1995).

Another problem of predictive validity is that it neglects the abilities being assessed given it is hard "to identify and measure all the abilities and factors that are relevant to the criterion and to be clear about the type and strength of the relation between the predictors" (Bachman, 1990, p. 252). Empirical studies on the predictive validity of language proficiency tests have been inconclusive regarding their ability to forecast academic achievement. The following summarizes several researches that reported effective results supporting the predictive validity of IELTS, and some others that reported negative results showing its weakness. For example, Cotton and Conrow (1998) examined the predictive validity of IELTS by investigating the correlation between IELTS scores and academic achievement of a group of international students at the University of Tasmania, Australia. They measured the correlations between IELTS scores and students' first and second semester using three measures of academic performance; Grade Point Average (GPA), academic staff ratings of students' performance (including course tutors, student advisors, and English support tutors), and students' assessment of their own academic performance in the first semester and the second semester. The results indicated that there was not a significant correlation between the three measures and IELTS overall scores, but there was a moderate positive correlation between reading component of IELTS and academic scores. Based on the findings, Cotton and Conrow (1998) concluded that the reading subtest of IELTS had the greatest ability to forecast upcoming academic performance, whereas other subtests had very weak or no relationship with academic achievement.

In contrast to the empirical researches reported, some studies indicated the strong predictive validity of IELTS test. For instance, Ingram and Bayliss

(2007) inspected the language behavior of nonEnglish-speaking pupils in Australian university. This research compared the participants' language behavior in various tertiary situations with their IELTS entry scores to discover whether there was any matching, and if their proficiency level, as measured by their IELTS scores, was enough to cope with the academic language tasks in their studies. To do this study, 28 international students who were in their first semester at two Australian tertiary institutions were selected. The obtained results uncovered that 25 the respondents displayed language behavior that equaled or exceeded what IELTS scores forecasted. This study came to the conclusion that IELTS scores can exactly predict students' language behavior in the first semester of their studies. In another study, Paul (2007) focused on four participants from the population of the previous study, and compared their language behavior in speaking and writing with their first semester language performance in various university programs. Paul (2007) utilized discourse analysis of classroom and IELTS spoken and written tasks to investigate how aspects of language including complexity, accuracy and fluency in academic contexts change from that produced under IELTS test conditions. The results of the four case studies indicated that language production was generally similar or improved in writing for three of the respondents, and it was generally identical as well in speaking though two of the participants indicated lower level in some aspects of academic tasks. The study concluded that the participants exhibited similar level of language in classroom tasks as that shown in IELTS test, and therefore, IELTS scores generally forecasted students' language levels in academic settings.

Gagen (2019) conducted a research on predictive validity of IELTS scores in Western university, in Canada. The researcher examined the predictive validity of IELTS scores on academic success, measured as GPA.

The findings of the study show that an approaching small effect size of $r =$ .22 was found for the overall relationship between the two variables. Additional analyses show that there is unlikely any bias from IELTS funding of results, that no single subscore is as strong a predictor of success as the overall score, and that few conclusions can be drawn about differences among levels, field, country of study, top-up courses, or time point.

Hartwell, Johnson, and Posthuma (2019) studied predictive validity of structured interview in the United States. The researchers gathered the data from 303 adult job applicants hired as employees of a state government agency in the southern United States. All data were gathered within a six-year period during which structured interviews were utilized. All applicants hired within that period constituted the study sample. Results reveal that all question types except job knowledge questions had validity in predicting subsequent job performance ratings. The insignificant result for job knowledge questions was surprising, as knowledge has been shown to be a top predictor of performance. However, the test of predictive validity of job knowledge was approaching marginal significance.

Yet in another study, Rumsey (2013) examined the predictive validity of the Common Educational Proficiency Assessment (CEPA) that is utilized as a vital requirement for admission to one of the three government tertiary institutions in United Arab Emirates. The test scores are utilized to put the candidates in the right level of the prebachelor program called foundation which takes one year and is used to prepare students with limited English, Math and computer skills to cope with the demands and challenges of the diploma and bachelor programs. She carried out her research at the Higher Colleges of Technology and showed that prior to her study, CEPA had been reported to have had relatively high relationship with students' English marks at the end of the one-year foundation. She aimed to investigate the predictive

validity of CEPA and to recognize other variables that help having a strong predictive validity. The researcher used a mixed method including qualitative and quantitative data collecting instruments from 347 students, foundation supervisors and coordinators, the central administrator and the CEPA supervisor. The results revealed that CEPA was a good indicator of students' English level prior to their admission to the college, and it was a good indicator of students' subsequent performance at the foundation year.

Milena, Severac, Baroffio, and Pelaccia (2019) evaluated the construct and predictive validities of the French version of the Strength of Motivation for Medical School-Revised questionnaire (SMMS-R-FR). The sample comprised 372 students at three French-speaking medical schools, who filled in the SMMS-R-FR and the Revised two-factor Study Process Questionnaire (R2-SPQ). The results confirmed positive similarity between our validation and the original version. The results of the study confirmed the reliability and validity of the SMMS-R's original three-factor structure. In addition, predictive validity was good for the Total Strength of Motivation scale and for all three of the SMMS-R-FR's subscales, although correlation strengths varied across scales and learning approaches.

Gebril (2009) examined predictive validity and test method. He investigated how reading-to-writing tasks are comparable to independent tasks in terms of score predictive validity. The researcher wanted one hundred and twenty five students from Egypt to write on two independent tasks and two reading-to-writing tasks. The results showed that the reading–to-writing tasks yielded as a reliable scores as independent tasks. Moreover, the findings indicated that the predictive validity is very low when using one task due to large variance component. The results also provided support for testing programs that employ integrated tasks.

Akbari and Abednia (2010) investigated second language teachers' sense of self-efficacy. The researchers used the new questionnaire Second Language Teaching Efficacy Scale (SLTES), consisted of 34 items, a combination of the 24 items of TSES (its long form) and the 10 added items related to language teachers. The questionnaire was administered to 272 English language teachers for construct validation. Factor analysis of the collected data with principal axis factoring showed the same factor structure of TSES, that is, efficacy in instructional strategies, efficacy in student engagement, efficacy in classroom management, which were redefined in light of the items, added to each.

Gebril (2006) did his PhD dissertation on predictive validity and concluded that the highest predictive validity and dependability coefficients were obtained by having four tasks and four raters in both tasks while the lowest estimate was obtained by using only one task and one rater.

In another study, Vafaee and Yaghmaeyan (2015) studied predictive validity of a speaking placement test scores. Their study employed univariate predictive validity theory to investigate the different sources of variance in the test scores and the dependability of the scores obtained from the Columbia University placement (CUP). The results showed that the dependability of the scores of CEP speaking test is high enough to be taken as a consistent measure of the speaking ability of the test takers.

Rezaei, Barati, and Youhanaee (2016) studied teaching practices and strategies employed in IELTS preparatory courses taught via strategy-based vs. non-strategy-based instruction in Isfahan. They selected two preparatory courses: strategy-based vs. non-strategy-based instruction. Communicative Orientation of Language Teaching (COLT) was employed to observe the courses which is a standard observation scheme. The findings showed that positive washback was observed in both strategy-based and nonstrategy-

based classes and candidates in strategy-based group outperformed their counterparts in nonstrategy-based group in reading and writing section of IELTS. In contrast, nonstrategy-based group performed significantly better in speaking section.

To summarize, the review of literature on IELTS academic tests has revealed that the predictive validity quality of IELTS academic tests has not been reviewed thoroughly. The review focused on this quality of IELTS academic tests investigated in recent studies and the gap of the research have indicated that the study could be significantly provided the IELTS academic practitioners with invaluable information and applicable findings.

3. Methodology
3.1 Participants
The participants were 70 test takers who took IELTS academic exam in 2014-2015. The participants were selected through convenience sampling. In order to select the participants, the researcher went to some IELTS preparation centers and asked for contact information of those who attended the test language institutes. Two hundred contact information entries were obtained. The researcher tried to either sent email to or call those who attended the language institutes. Among the list of contacts, only 120 responded and out of this number 70 claimed that they took IELTS academic test and passed the test. The participants were of different majors such as biology, civil engineering, medicine, chemistry, so on.

3.2 Instrumentation
To collect the required data for the study, two different instruments were used: Self-assessment inventory developed by Alibakhshi (2013) and learners' IELTS self-report. Self-assessment inventory scale was constructed and validated by Alibakhshi (2013). This scale consists of 41 items measured on a Likert scale. The 41 items were loaded on four components and the loading factor of each item exceeded the cutoff score which was set at 0.4.

The reliability of the instrument was calculated through running Coronbach's alpha. The Coronbach's alpha of 0.90 indicated that the adapted questionnaire enjoyed a high level of internal consistency.

The second instrument was developed by the researcher. It consisted of two parts. The first part contained items, which delineated the participants' majors, age, and gender. The second part contained four self-report items through which the participants reported their scores on different modules of the IELTS.

3.3 Procedure

The study was conducted to answer four formulated questions on predictive validity of IELTS skills and academic setting. As a first step, the required questionnaires were administered to the candidates either directly by the researcher or via some colleagues and some of them were sent via e-mail to participants. The participants were informed of the purpose of the research and their consent was obtained. After gathering the questionnaires, the researcher analyzed the scores of each participant on each of the instruments. Each questionnaire was given a numerical code to know the respective associated score on each of the measure.

After eliciting the scores from the measures, they were entered into SPSS data sheet. In order to answer the questions of the research about the predictive validity of IELTS academic test, descriptive statistics of the participants including frequencies, percentage, and mean scores were calculated of the data elicited form predictive validity questionnaires. At the end the test takers' self-reported scores on IELTS modules were reported as predictors (independent variables) and their scores on the self-assessment modules were rated and computed as the dependent variables.

4. Results

The results of the study including four different regression analyses are presented in the following section:

4.1 IELTS Academic Tests and Listening Skill

To see whether, the test takers' scores on listening module of IELTS significantly predicts their listening skill scores in an academic situation, a regression analysis was run. Results are shown in Table 1.

Table 1
Summary of Simple Regression Analysis for Predicting Listening Skill in Academic Situations

Equation	Model Summary					Parameter Estimates	
	R^2/	F	df1	df2	Sig.	Constant	b1
Linear	.24	14.748	1	70	.001	3.7	.52

Simple Regression Analysis was used to test if the participants' scores on listening module of IELTS significantly predicted their listening performance in academic settings. The results of the regression indicated the predictor explained 24% of the variance (R^2= .24, F (1, 70) = 14.75, p = .001< 0.05). That is, the listening module of the IELTS The listening module of IELTS test could explain only 24% of the variance of listening tasks in real academic life of the test takers.

4.2 IELTS Academic Tests and Speaking Skill

Regression analysis was used to see if the learners' scores on IELTS academic could predict their scores on academic speaking skill. Results are shown in Table 2.

Table 2
Summary of Simple Regression Analysis for Predicting Speaking Skill in Academic Situations

Equation	Model Summary					Parameter Estimates	
	R^{22}	F	df1	df2	Sig.	Constant	b1
Linear	.40	17.748	1	70	.001	4.7	1.2

Simple Regression analysis was used to test if the participants' scores on speaking module of IELTS significantly predicted their speaking performance in academic settings. The results of the regression indicated the predictor explained 40% of the variance (R^2 = .40, F(1, 70) = 17.74, p =

On the Predictive …

.001< 0.05).The speaking module of IELTS test could explain only 40% of the variance of speaking tasks in real academic life of the test takers.

4.3 IELTS Academic Tests and Writing Skill

A Regression was performed to compute the relationship between writing in target situation and IELTS academic writing skill (Table 3).

Table 3
Summary of Simple Regression Analysis for Predicting Writing Skill in Academic Situations

Equation	Model Summary					Parameter Estimates	
	R^2	F	df1	df2	Sig.	Constant	b1
Linear	.53	19.748	1	70	.001	4.2	1.5

The results of the Table 3 show that the participants' scores on writing module of IELTS can significantly predict their scores on writing tasks in target language use situation (R^2 = .53, $F(1, 70)$ = 17.74, p = .001< 0.05). It can also be seen that the writing module of IELTS test could explain 53 percent of the variance of writing tasks in real academic life of the test takers.

4.4 IELTS Academic Tests and Reading Skill

Simple Regression analysis was used to test if the participants' scores on reading   module   of   IELTS   significantly   predicted   their   reading comprehension in academic settings. Results are presented in Table 4.

Table 4
Summary of Simple Regression Analysis for Predicting Reading Skill in Academic Situations

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	.60	25.748	1	70	.001	4.00	1.3

The results of the Table 4 show that the participants' scores on reading module of IELTS can significantly predict their scores on writing tasks in target language use situation (R^2 = .60, $F(1, 70)$ = 17.74, p = .001< 0.05). That is, the writing module of IELTS test could explain 60 percent of the variance of writing tasks in real academic life of the test takers.

5. Discussion

This study aimed at investigating the predictive validity of IELTS academic tests. To answer the questions posed in this study, the self-assessment questionnaire was used to collect the required quantitative data on predictive validity of IELTS academic tests. The relative impact of IELTS academic test scores on students' overall academic listening, speaking, reading, and writing performances was predicted through four regression analyses. Regarding the four posed questions of the study, four different regression analyses were run. As the findings show, all modules significantly predict the test takers' performance in academic settings.

Further detailed analysis showed that listening module of IELTS predicted only 24% of the variance of the test takers' performance in listening tasks in academic target situations, despite the fact that the correlation was significant. Results also revealed that speaking module predicted 40% of speaking, the scores on writing module predicted 53% of the variance of the scores on writing performance of the test takers in academic target situations while the scores on reading module predicted 60% of reading scores in the target academic situations.

The findings of this study are not consistent with the findings of Alibakhshi (2013), Dauglas (2001), and Spolsky (1995). The results of the study carried out by Alibakhshi(2013) indicated that test takers' scores on the entrance examination to universities were not good predictors of the test takers' performance in academic setting. However, based on the results of this study, it could be argued that the predictive validity of listening module was not satisfactory while the predictive validity of reading and writing modules were above the average but not perfect. Therefore, it could be argued that the tasks of listening are not the same as the tasks used in the academic situations.

Therefore, as Douglas (2001) believes, it is very difficult to predict the nontest performance in the real life target situations. In other words, it is not easy to predict whether those who get their favorite band scores on IELTS academic scores can meet their needs in real life situations or not. It could also be argued that not all four modules of IELTS academic tests could have the same predictive validity. It could also be argued that lack of appropriate predictive validity of IELTS academic tests skills could be resulted from the fact that it is impossible to adequately sample even the most particular target language use tasks. Such a deficiency of IELTS in predicting the test takers' performance in real life situations as Alibakhshi (2013) argues, can be deeply associated with the fact that in IELTS academic tests we are able to measure the candidates' language proficiency involving four main language skills in terms of accuracy, complexity and fluency and assess their general ability in L2, but as Paul (2007) argues, it is not known whether the IELTS test takers can do the academic tasks in real academic situations, which require the use of language skills cannot learn a lot about their language abilities to carry out the tasks in real life situations.

The results of the study carried out by Moore and Morton (1999) showed that IELTS writing task 2 did not match exactly with any of the academic genres in the target language use domain as the university writing corpus was based on external sources as opposed to IELTS writing task 2, which was based on prior knowledge as a source of information. IELTS writing task 2 was more similar to non-academic public forms of discourse such as letter to the editor; while the results of the study showed that the participants' scores on writing module of IELTS can significantly predict their scores on writing tasks in target language use situation.

The findings of the study are also not in line with the results of the study done by Cotton and Conrow (1998) who examined the predictive validity of

IELTS by investigating the correlation between IELTS scores and academic achievement of a group of international students at the University of Tasmania, Australia. The researchers came to the conclusion that listening, writing and speaking subtests had very weak or no relationship with academic achievement. It can be due to the fact that, as Douglas (2001) and Paul (2007) argues, predicting academic performance is difficult because it is not known whether the IELTS test takers can do the academic tasks in real academic situations. As Spolsky (1995) argues, the tests produced within this new 'modern' paradigm did little to embrace the criterion of validity and objective reading tests may be reliable, they may not be delivering valid information on the abilities we seek to measure.

In line with the findings of the study, it could be concluded that as IELTS as a high stake test might have negative consequences for the stakeholders. One of the negative consequences is the predictive validity, which should be guaranteed by the test developer, test users, and test interpreters. Moreover, due to the vital importance of IELTS academic tests and their roles in acceptance/non-acceptance of the test candidates in universities in which English is used as medium of instruction, it is strongly suggested that test users should safely trust in the validity of IELTS academic tests so that they can generalize the students' scores and results of their IELTS band scores to real life situations.

Due to the practical limitations, it was not possible to develop a test to assess the test takers' performance on language skills in real academic situations and the test takers' self-reported scores on self-assessment inventory were considered as their scores in the target language situations; therefore, results might be different a real standardized test is used for assessing the test takers' performance. The other interested researchers are recommended to replicate the study using an authentic performance test

consisting of real academic situations tasks to see whether the IELTS test has

psychometrics characteristics such as reliability and construct validity.

References

Abbiati, M., Severac, M., Baroffio, A.,& Pelaccia, T.(2019).Construct and predictive validity of thestrength of motivation for medical school-revised (SMMS-R) questionnaire: A French validation study. *Canadian Medical Education Journal, 10*(3), 32-48.

Akbari, R.,& Abedni, A. (2010). Second language teachers' sense of self-efficacy: A construct validation. *Journal of Teaching English Language,4*(1),69-101.

Alderson, J. C. (2005).*Diagnosing foreign language proficiency*. London: Continuum.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Alibakhshi, G. (2013). Construction and validation self-assessment inventory for learners of English for academic purposes. *Journal of Research in Applied Linguistics,4*(2), 93-106.

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Chapelle, C. A. (2012). Conceptions of validity. In G. Fulcher & F. Davidson. (Eds.). *The Routledge handbook of language testing* (pp.21-33). Oxon: Routledge.

Cotton, F., & Conrow, F. (1998). An investigation of the predictive validity of IELTS amongst a group of international students studying at the University of Tasmania. *IELTS Research Reports, 1*, 72-115.

Council of Europe. (2001). *Common European framework of reference for languages: Learning teaching, assessment*. Cambridge: Cambridge University Press.

Douglas, D. (1997). Language for specific purposes testing. *Language testing and assessment, 7*, 111-119.

Douglas, D. (2001). Language for specific purposes assessment criteria: Where do they come from? *Language Testing, 18*(2), 171-85.

Elder, C. (1993). Language proficiency as a predictor of performance in teacher education. *Melbourne Papers in Language Testing, 2*(1), 68-89.

Gagen, T. (2019). *The predictive validity of IELTS scores: A meta-analysis.* The University of Western: Canada.

Gebril, A. (2006). *Independent and integrated academic writing tasks: A study in generalizability and test method.* The University of Loma.

Shakibaei & Memari

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing*, *26*,507–531.

Hartwell, J., Johnson, D., & Posthuma, A. (2019). Are we asking the right questions? Predictive validity comparison of four structured interview question types. *Journal Business Research, 100*, 122-129.

Ingram, D., & Bayliss, A. (2007). IELTS as a predictor of academic language performance, part1. *IELTS Research Reports, 7*, 137-204.

Moore, T., & Morton, J. (1999). *Authenticity in the IELTS academic module writing test: A comparative study of Task 2 items and university assignments* (IELTS Research Reports no. 2.). Canberra: IELTS Australia.

Paul, A. (2007). IELTS as a predictor of academic language performance, Part 2. *IELTS Research Reports, 7*, 205-271.

Pearson (2010). *PT E Academic score guide.* Retrieved from http://pearsonpte.com/PTEAcademic/scores/Documents/PTEA_Score_Guide.pdf.

Polit, D., & Hungler, B. (1991). *Nursing research: Principles and methods*. New York: JB Lippincott.

Rezaei, O., Barati1, H., & Youhanaee, M. (2016). Evaluation of IELTS Preparatory Courses in Iran: Teaching practices and strategies in focus. *Journal of Teaching English Language, 10*(2), 47-71.

Rumsey, L. (2013). *Shedding light on the predictive validity of English proficiency tests in predicting academic success* (Doctoral dissertation). ETHOS.

Shuttleworth, M. (2009). *Peer review process.* Retrieved from http:// experiment-resources.com/peer-review-process.html.

Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford: Oxford University Press.

Uysal, H. H. (2010). A critical review of the IELTS writing test. *ELT Journal, 64*(3), 314-320.

Vafaee, P., & Yaghmaeyan, B. (2015). Providing evidence for the generalizability of a speaking placement test scores. *Iranian Journal of Language Testing 5*(2), 78-95.

Zheng, Y., & De Jong, J. (2011). *Research note: Establishing construct and concurrent validity of Pearson Test of English Academic (PTE Academic).* Retrieved from http://pearsonpte.com/research/Pages/ResearchSummaries.aspx.