ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 1, No. 10, pp. 1304-1310, October 2011
© 2011 ACADEMY PUBLISHER Manufactured in Finland.
doi:10.4304/tpls.1.10.1304-1310

External Validity of TOEFL Section of Doctoral Entrance Examination in Iran: A Mixed Design Study

Goudarz Alibakhshi
Department of applied linguistics, Yasouj University, Iran
Email: alibakhshi_goodarz2000@yahoo.com

Hassan Ghand Ali
Payamenoor University, Iran

***Abstract*—External (generalization) validity is one aspect of construct validity (Messick, 1996) which deals with the inferences made on the basis of the test takers' scores on a test. External validity is of particular importance in high stake tests such as tests of English for academic/specific purposes which are used to evaluate the test takers proficiency in general English for academic purposes. This study was an attempt to investigate the generalization validity of TOEFL tests administered at Iranian universities to select Ph. D candidates. In doing so, a mixed design study was applied. The data for the quantitative part was collected through a self assessment instrument consisting of personal information and 40 items designed on five point Likert scale. 450 doctoral students from different universities in Iran took part in the study. The data of the study were analyzed through descriptive and inferential statistics including principal component analysis, univariate analysis of variance and regression analysis (p=.05) approaches. The qualitative data was analyzed through content analysis. The results of the study indicated that there is a significant difference between the participants' mean scores on TOEFL test and their means on academic language skills. Moreover, TOEFL scores did not significantly predict the test takers' scores on the use of academic language in target language use situations. Therefore, TOEFL test developers should take the issue of generalizability into consideration while planning TOEFL tests.**

***Index Terms*—generalization validity, construct validity, TOEFL tests, target language use situations**

I. Introduction

Although in Iran almost all undergraduate and postgraduate courses are taught in Persian, learners particularly master and doctoral candidates are urged to also read other sources and professional journals in English. To help students improve their English undergraduate and postgraduate students have to take language for academic purposes (EAP). The EAP courses not unlike the other subject matters are parts of master and doctoral entrance examinations which are held nationally and locally, respectively. The test takers' knowledge on EAP is usually tested through a general language test consisting on vocabulary, writing, and reading items. Such a kind of test is administered by almost all universities running Ph.D courses. The participants' score is a certificate for taking part in Ph.D examination. This test is of much significance to all Ph.D candidates. Therefore, like the other tests, TOEFL section of Ph.D entrance examination should have some characteristics such as validity.

Traditionally testers have distinguished different types of validity: content, predictive, concurrent, construct and face validity. Messick (1994; 1996) challenged this view and argued that construct validity is a multifaceted but unified and overarching concept which can be researched from a number of different perspectives. To simply put it, Messick (1989) captured the essence of construct validity into six distinguishable aspects, namely content, substantive, structural, generalizability, external, and consequential. In effect, these six aspects function as general validity criteria or standards for all educational and psychological measurement.

In the generalizability (external) aspect of construct validity, the concern is that a performance assessment should provide representative coverage of the content and processes of the construct domain. That is, to ensure that the score interpretation not be limited to the sample of assessed tasks but be generalizable to the construct domain more broadly (Messick, 1996).Evidence of such generalizability depends on the degree of correlation of the assessed tasks with other tasks representing the construct or aspects of the construct. This issue of generalizability of score inferences across tasks and contexts goes to the very heart of score meaning. Indeed, setting the boundaries of score meaning is precisely what generalizability evidence is meant to address.

Generalization validity is emphasized in two senses, namely, reliability and transfer. Generalizability as reliability (Feldt and Brennan, 1989) refers to the consistency of performance across the tasks, occasions, and raters of a particular assessment, which might be quite limited in scope (Messick, 1996). In contrast, generalizability as transfer requires consistency of performance across tasks that are representative of the broader construct domain. That is, transfer refers to the range of tasks that performance on the assessed tasks facilitates the learning of or, more generally, is predictive of (Ferguson, 1956 cited in Messick, 1996). The review of literature, indicate that the construct validity particularly general validity of TOEFL tests has not been studied appropriately.

The main objective of the present study is to investigate external validity of TOEFL section of doctoral entrance examinations which are developed and administered both nationally and locally by Iranian state universities running doctoral courses.

II. Research Question

To state the aim of the study, the following research questions were raised:

1-To what extent do the currently practiced TOEFL tests administered at Iranian universities have the characteristic of external validity?

2- How do Iranian Ph.D candidates view external validity of such proficiency tests?

III. Review of Literature

The review of literature indicates that the external (generalization) validity- the validity of inferences which are made on the basis of the tests takers' score on proficiency TOEFL test of local TOEFL tests has not been studied appropriately. However, the other two most popular tests: TOEFL and IELTS which provide the evidence of proficiency in the English language for non-English-speaking-background (NESB) students have been investigated in terms of their generalizability.

In the past 20 years, studies dealing with the relationship between language proficiency and academic achievement have been conducted. Lee (2006) investigating the dependability of scores on speaking assessment consisting of integrated and independent tasks through, generalizability theory (G-theory) procedures stated that it would be more efficient to increase the number of tasks rather than the number of ratings per speech sample in maximizing the score dependability. The multivariate G-theory analyses also revealed that (1) the universe (or true) scores among the task-type subsections were very highly correlated and that (2) slightly larger gains in composite score reliability would result from increasing the number of listening – speaking tasks for the fixed section lengths.

Van Moere (2006) investigated a group oral test administered at a university in Japan - one component of an in-house English proficiency test used for placing students, evaluating their progress, and making informed decisions for the development of the English language curriculum -to find if it was appropriate to use scores for higher stakes decision making.. Rasch analysis showed rater fit within acceptable levels considering the length and nature of the test; however, at correlations of .74 inter-rater agreements were lower than has been reported in research on commercially available interview tests. Candidates' scores on the two different test occasions correlated at .61. A generalizability study showed that the greatest systematic variation in test scores was contributed by the person-by-occasion interaction. Topic, or prompt, was not a significant factor. Candidates' performances, or how raters perceive an individual candidates' ability, could be affected to a large degree by the characteristics of interlocutors and interaction dynamics within the group.

Tonkyn (1995) asserts that there is plenty of evidence that language proficiency is a significant issue regarding the academic performance of overseas students, and that students who score higher on a Standard English test have a greater chance of future academic achievment.

Berieter and Scardamalia (1982) concluded that the problems of learning to speak and learning to write resulted in academic failure. Cummins and Swain (1986) established a theoretical framework concerning the difference between academic language and daily-life language. Demie, Taplin and Butler (2003) examined the relationship between bilingual students' level of English fluency and academic achievement and stated that bilingual students who were not fluent in English tended to do less well in school, and those who were fully fluent in English generally outperformed their English-speaking peers significantly. Kato, Albus, Liu, Guven and Thurlow (2004) reported that there was a strong relationship between the comprehensive assessment and the academic English test. In addition to IELTS and TOEFL tests, construct validation of performance assessments has been studied. Performance assessment typically asks students to show the processes of their thinking and reasoning so that educators can make direct inferences on the nature and depth of students' understanding (Lane, Liu, Ankenmann, & Stone, 1996; Messick, 1994). Linn, Baker, and Dunbar (1991) further stated that both logical and empirical evidence should be presented in order to draw valid inferences from performance assessment. They specified consequential validity and fairness as necessary criteria for evaluating performance assessments. Examining the Maryland School Performance Assessment Program (MSPAP), Yen and Ferrara (1997) found that the reading, writing, language, and math assessments of MSPAP show a substantial correlation (a = .54 to .78) with the reading, language, and math assessments of the Comprehensive Tests of Basic Skills, Fourth Edition (CTBS-4).

Performance assessments can also have fairly strong predictive validity on future achievements. Davis, Caros, Grossen, and Carnine (2002) found that the score components of a writing benchmark assessment significantly predicted achievement in SAT-9 and High School Exit Exam (HSEE) scores. The function, based on the score components, correctly identified 77% of students in the upper or lower 50th percentiles on the SAT-9 Writing score distribution and 67% of students in the upper or lower 50th percentiles on the HSEE Writing score distribution. Although review of literature indicates that the geneneralizabily of international proficiency tests has been studied to a great extent, the local TOEFL/EAP tests such as those practiced at Iranian universities have not been investigated in terms of inference validity.

IV. Methodology

A. Participants

The participants of the study were 450 doctoral students majoring in chemistry, biology, geography, civil engineering, Persian literature and geography. The participants were selected through multi-stage sampling procedure. First, from different branches of science five disciplines were randomly selected. Then, from all universities running master and Ph.D courses seven big universities (Isfahan, Mashhad, Tarbiat Modares, Tehran, Allama, Shiraz, Shahid Chamran, Tarbiat Moalem and Shahid Beheshti,) were randomly selected. In order to know if the sample size is large enough to represent the population, we consulted Krejcie and Morgan (1970), which offer a table for estimating sample size by giving figures for populations ranging from 10 to 1000000 and the corresponding figures for the required sample size. The appropriate sample size for the population of the present study was found to be 390 doctoral students who were selected through convenient sampling procedures. To be on the safer ground, we selected 450 test takers.

B. Instrumentation

The main instrument used in this study was a *Self-assessment Questionnaire.* The questionnaire consisted of two parts. The first part dealt with the participants' general information such as major, education level, and their scores on TOEFL part of doctoral entrance examination. The second part consisted of four components: listening, speaking, writing, and reading with 10 items on each skill. Respondents assessed their English proficiency on a 5-point scale ranging from very week (1 = very week) to very good (5 = very good). The ordinal scale then was converted into interval scale. Therefore, the participants' score on each skill ranged between 10 -50.

In addition to the questionnaire, a Cued-Recall Interview was used to tap into the participants' knowledge of the generalization validity of such tests. That is, at first a leading question was asked to see what perceptions the test takers may have about the possible generalizations which they can make about such tests. Then, the participants' answers were followed by some other questions to explore the main possible merits and demerits of such tests in terms of generalization which can be made on the basis of the scores of such tests.

C. Data Analysis

This was a mixed design study and therefore a specific procedure was need. As a first step, the needed questionnaires were administered to the participants either directly by the researcher or through some colleagues and some were e-mailed to them. After collecting the questionnaires, they were analyzed by the researcher and the scores of each participant on each of the measures were calculated. The participants' score on TOEFL tests (independent variable) and theirs on academic reading, listening, speaking, and writing skills (dependent variables) were gathered. The data of study were analyzed through different statistical procedures including principal component analysis in order to extract the irrelevant items, univariate analysis of variance to compare the participants' mean scores on the five tests, Cronbach alpha for estimating internal consistency, and regression analysis to explain the effects of TOEFL on the prediction of the test takers' scores on different academic skills.

Then, the qualitative data was collected. Data were gathered during face-to-face in-depth interviews. The researchers informed the participants of the purpose of the research and obtained their written consent. The researchers also obtained the participants' permission to audiotape each interview for purposes of content analysis and audit trail. The interviews were conducted in both an unstructured and a semi-structured manner. The interviews lasted on average for about 30 minutes. Interviewing took place during all days over a five-month period, until the data collected were being consistently duplicated. No new information was gained from the last three interviews, thus data saturation was considered to have been achieved. The interview data were immediately transcribed verbatim and analyzed using qualitative content analysis

V. Results

A. Results of the Quantitative Phase

1. Results of factor analysis

As the instrument consists of four variables, four different factor analyses with extraction method of principal component analysis were run. The initial Eigenvalues for all components were above 6. The loading factors for all items of each component were above .6. Therefore, it was confirmed that all items of each variable constitute one factor. The reliability of the instrument was above .9 which indicates the instrument has a very good internal consistency.

TABLE 1.
UNIVARIATE ANALYSIS OF VARIANCE

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	67113.392	4	16778.348	570.219	.000
Intercept	1187256.131	1	1187256.131	40349.361	.000
TESTS	67113.392	4	16778.348	570.219	.000
Error	57230.477	1945	29.424		
Total	1311600.000	1 950			
Corrected Total	124343.869	1949			

As the results in the above table indicate there is a significant difference between the participants' means on different tests (F= 570.219, df= 4, Sig. = .000).

2. Results of Regression Analysis

Four different regression analyses were performed to examine whether the learners' scores on TOEFL tests had any effects on their academic listening, speaking, reading, and writing proficiency in target language use situations. The results are presented in the following Table:

TABLE 2:
MODEL SUMMARY FOR LANGUAGE SKILLS

Model	B	Se(B)	Sig.	R	R-Square	Adjusted R square
Listening	.090	.073	.21	.009	.004	.004
Speaking	.156	.037	.12	.160	.019	.015
writing	.091	.035	.059	.091	.008	.006
reading	.062	.038	.000	.622	.39	.36

a. Predictors (constant) TOEFL, b. dependent variables: listening, speaking, writing, and reading

As can be seen, TOEFL scores had no significant effect on the prediction of the learners' scores on academic listening skill, speaking, and writing skills. However, the results indicate that TOEFL scores had significant effects on the prediction of the learners' scores on academic reading skill.

B. Qualitative Analysis

20 doctoral students and 10 applied linguists were interviewed. Five themes were extracted from the interview data using qualitative content analysis. Here, we show the themes with direct quotations to exemplify them.

1. External invalidity pollutes any decision made on the test

Almost all participants believed that a good test is the one which predict the performance of the test takers' in real life situations. The TOEFL section of doctoral examination is administered to evaluate the test takers' English for general purposes. Therefore, high scores on this test means that the test takers are able use language in real academic situations. However, the participants believed that despite their high scores on this test they are not able to meet their academic needs. A doctoral student who got a very good test on TOEFL test states:

It is really naïve to think that a test taker who passed this test can make use of language in target language use situations. The items of the proficiency test that we took are not representative of the tasks in real life situations. This test only measures our knowledge of vocabulary, grammar, and reading comprehension; whereas, we do quite different tasks in target language use situations. So it is quite clear that the results of the mentioned test cannot be generalized to the target situations.

2. External invalidity is due to this test indirectness

Another extracted theme was grammar, vocabulary, and reading comprehension are tested through multiple choice questions; whereas, in target academic situations doctoral students read the texts for reporting the main findings, to summarize, etc. Consequently, it could be strongly argued that no one can make generalization on the basis of a set of indirect items. To illustrate this theme, the following quotations from the participants are reported.

A Ph.D candidate in chemistry at one of the universities in Iran argued:

Answering the items on TOEFL test was a piece of cake for me. I practiced a couple of sample tests and I got aware of test methods and the contents. I got 90 out of 100. But now while writing an abstract, or summarizing a paper in English I do really find it hard and beyond my ability. I think generalizations made on the tests consisting of indirect items on grammar and vocabulary cannot be as much valid as one expects.

3. Generalization follows authenticity

All most all participants stated that the inferences and decisions made about their abilities on the basis of their scores on selection TOEFL tests are not valid because there is a very low similarity between the tasks and contents of these tests and real academic life non-test tasks. In other words, the students' high achievements in these tests do not guarantee their success in real academic life. The following direct quotation from a Ph.D candidate of law illustrates this theme.

Another problem with this test is that it onl measures limited skills and sub-skills through discrete point items. Whereas, in academic situations do we need to use the skills integratively. So the mismatch between the test items in tests and academic situations will expose the tests to indirectness and invalidity as well.

Another participant majoring in Geography states:

The TOEFL tests administered at the university in which I am studying now is limited in scope and contents. Vocabulary is tested through multiple choice items and the reading passages are not related to my own field of study. The tasks I need to do in target language use situation are more complex and complicated in terms of the discourse, genre, and lexicon. Therefore, the scores on the TOEFL test cannot predictive my performance the test takers' performance on in target language use situations.

4. Generalization invalidity leads to negative consequences

Another extracted theme was that if a test does not have generalization validity the inferences which are made on the basis of its scores will certainly lead to negative social, personal, individual, and financial consequences. These negative consequences are more detrimental when the administered test is a high stake test. The following direct quotations illustrate this theme.

One Ph.D candidate in applied mathematics argues:

Two or three years in row I was deprived of taking part in content part of doctoral examination just because of my low scores on TOEFL section. Although I finally passed this test, I have many problems in English. I don't know what the use of such invalid test is…….

5. External validity influences washback validity

A majority of the participants stated that they do not try to practice the language skills which they really need. They also argued that studying general and technical reading is sufficient for them to pass the test. That is why; they hardly ever study the journals and textbooks related to their field of studies. Therefore, such tests do not produce significant washback effects on learning. One of the participants majoring in Persian literature states:

One needs to learn finite grammar rules and the most essential words for TOEFL so that s/he can pass this test. Therefore, s/he does not attempt to study the other sources such as the journals, reference books, or other related books. Therefore, no innovation in language teaching and learning methods is made by the learners. I myself just use traditional strategies such as memorization. I think the results of this test do not lead to positive impacts on learning and teaching.

VI. Discussion

TOEFL test as a part of doctoral examinations are used throughout the Iranian universities both to control the entry of students into post graduate studies and to diagnose the test takers' proficiency. Language tests require technical expertise in their construction and application, in order to make the inferences that we draw from test results interpretable and supportable. The aim of this study was to determine the relative impact of the independent variable, the test takers' score on TOEFL test, on their performance on target language use situation tasks. It was also an attempt to study the external validity of such test from test takers' points of views.

To do so, the students were asked to assess their use of academic language in TLU situations. 450 students returned the questionnaire. The reliability of the questionnaire was above .9. Construct validity of the questionnaire was calculated through principal component factor analysis. The results of factor analysis indicate that the initial Eigenvalues for each component was above 5.3 and the loading of each factor was above .65. A brief look at the loadings shows that almost all of the loadings are high enough to conclude that all ten items of each component constitute one factor. The assumption of the study was that the students' performance on TOEFL tests was responsible for the greatest share of variance for all academic language skills. The results of data analysis, however, rejected this assumption and revealed the students' proficiency in TOEFL tests had by far the greatest share of variance for reading skill.

As the participants had five independent tests, the univariate analysis of variances was the best approach for comparing their means. This statistical approach is used when one group of participants have more than two different tests. As it could be seen, there is a significant difference between the participants' mean scores on TOEFL test and language the skills of language for academic purposes. The results of post hoc test (Tukey) also confirms that the participants' mean score on TOEFL test is significantly different from their means on listening, reading, writing, and speaking. The results also indicate that there is no difference between the means of listening, speaking and writing. However, the participants mean on reading test was significantly different from the means on listening, speaking, and writing. The descriptive analysis shows that mean scores on TOEFL test was the highest and reading was next to it (33.14 & 30. 33)

Four regression analyses predicted the relative impact of TOEFL test scores on students' overall academic listening, speaking, reading, and writing performances. The findings indicated that TOEFL proficiency accounted for.8% of the listening variance1%of speaking, .6% of writing, and38% of reading. Due to the vital importance of these tests and roles which they play in acceptance/non-acceptance of the candidates, it is firmly believed that test users should safely trust in the validity of such tests so that they can make inference and generalize the students' scores to non-test and real situations. Surprisingly enough, the TOEFL tests developed to screen doctoral candidates of chemistry, Persian literature, biology, geography, and civil engineering in 2009 and 2010 lacked generalization validity. That is, except for reading skill, there was no significant correlation between the test takers' scores on TOEFL tests and their use of academic listening, speaking, and writing tasks in target use situations. It could also be argued that, although TOEFL scores accounted for 38% of the reading variance, the results of unvariate analysis of variance indicated that there was a significant difference between the mean scores on reading scores and TOEFL test. As the results, such variance share cannot be considered as ideal. Therefore, it could be argued that TOEFL proficiency tests are not good predictors of doctoral students' proficiency in academic listening, speaking, and writing skills. That is, test developers and university authorities cannot generalize the scores on TOEFL to target language use situation tasks.

The generalization invalidity occurs as the result of lack of correspondence between the content of TOEFL tests and target language use (TLU) tasks. In line with advocates of communicative testing, it is argued that maximal authenticity – the degree of correspondence between a given test task and a target language use task (Bachman and Palmer 1996) – and directness – the extent to which a test entails a candidate performing precisely the skill(s) we intend to measure– should be fundamental considerations in test design. The influence of sociolinguistics and pragmatics on the construct of communicative language ability (Bachman 1990) dictates the integration of skills in meeting test task demands. However, the findings of this study contradict Bachman and Palmer (1996) notion of generalizability indicating that the score interpretation should not be limited to only the sample of assessed tasks but be generalizable to the construct domain.

Another reason for generalization invalidity of locally administered TOEFL tests is deeply rooted in their function. Where a test is used for selection, as are Iranian TOEFL tests, those who seek access will attempt to gain the skills they believe necessary to succeed on the test. Some of these skills are generally considered to be desirable, as they are required in the target language use domain. However, as all tests are limited in how much of the domain they can sample and involve a certain amount of measurement error, there is inevitably scope for the misrepresentation of test takers' abilities. The skills required to pass a test are not necessarily or comprehensively the skills required in a target language use domain (Bachman and Palmer 1996). The content analysis of the TOEFL tests indicates that they only include test items on technical reading and vocabulary; whereas, post graduate students need all language skills to cope with target language use situation tasks (Alibakhshi, etal). Therefore, it could be strongly discussed that the construct of TOEFL tests is underrepresented. The target language use situation tasks exceed the domain of reading and technical vocabulary. The post graduate learners' academic language needs are not covered by the TOEFL test contents. In addition, test tasks are to a great extent, different from the academic tasks which Iranian learners will face. They have to summarize the texts, take notes, paraphrase, describe a technical problem, take not from live lectures, etc. However, the test tasks are only multiple choice items which the candidates are instructed to select the best. It could be argued that such a mismatch between real life tasks and test tasks jeopardize the validity of the TOEFL tests as well as the inferences which are made on the basis of these tests results.

VII. Conclusion

The results of the study indicated that TOEFL tests practiced at Iranian universities do not have generalization validity. In order to increase the generalization aspect of construct validity in TOEFL tests it is recommended that test designer pay attention to the real academic needs of master and doctoral students. That is, the content of TOEFL tests should be authentically representative of the learners' needs. In addition, test tasks should be representative of TLU situations. That is, the scores of tests which are both authentic and direct could be generalized to target language use situation tasks. It could also be concluded that authentic tests do have generalization validity and well certainly have positive washback effects on teaching and learning TOEFL at Iranian universities**.**

References

[1] Alibakhshi, G., Kiani, G.R., & Akbari, R. (2010). On the authenticity of ESAP tests. *Journal of Asian ESP. 6(2),65-93.*

[2] Bachman, L.F. (1990). Fundamental considerations in language testing. New York: Oxford University Press

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

[4] Bereiter, C., & Scardamalia, M. (1982). "From conversation to composition: the role of instruction in a developmental process," in R Glasser, Ed., *Advances in instructional psychology.* Volume 2 pp 79-88. Lawrence Erlbaum Associates, Hillsdale, New Jersey

[5] Cummins, J., & Swain, M. (1986). Bilingualism in Education. Harlow: Longman Group UK Limited.

[6] Davis, B., Caros, J., Grossen, B. & Carnine, D. (2002). Initial stages in the development of benchmark measures of success: Direct implications for accountability (Research Report No. RR-11). Washington, DC: Special Education Programs (ED/OSERS). (ERIC Document Reproduction Service No. ED469290).

[7] Demie, F., Taplin, A., & Butler, R.(2003). "Stages of English Acquisition and Attainment Of Bilingual Pupils: Implications for Pupil Performance in Schools, in *Race Equality Teaching.* Volume 21. No.2 pp 42-48. ERIC, Singapore

[8] Feldt, L.S., & Brennan, R.L. (1989). Reliability. In Linn, R.L., editor, *Educational measurement* (3rd edn), New York: Macmillan, 10546

[9] Kato, K., Albus, D., Liu, K., Guven, K., & Thurlow, M. (2004). Relationships between a statewide language proficiency test and academic achievement assessments (LEP Projects Report 4). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

[10] Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement, 33*(1), 71-92.

[11] Lee, Y. W. (2006). Dependability of language scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language testing,* 23(2) 131-166.

[12] Linn, L. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.
[13] Messick, S. (1996). Validity and washback in language testing. *Language Testing* 13, 241–56.
[14] Messick, S. (1989). Validity. In Linn, R.L., editor, *Educational measurement* (3rd edn), New York: Macmillan, 13-103.
[15] Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 23(2), 13-43.
[16] Tonkyn, A. (1995). 'English language proficiency standards for overseas students: Who need what level?' *Journal of International Education*, 6, 3: 37–61
[17] VanMoer, A. (2006). Validity evidence in a university group oral test. *Language Testing* 2006, 23 (4) 411–440
[18] Yen, W. M., & Ferrara, S. (1997). The Maryland school performance assessment program: Performance assessment with psychometric quality suitable for high stakes use. *Educational and Psychological Measurement, 57*(1), 60-84.

Goudarz Alibakhshi is an assistant professor of applied linguistics at Yasouj University. He has published several papers in international journals. He has been teaching ESP, applied linguistics, language assessment and research methodology to undergraduate and postgraduate students at Iranian state universities since 12 years ago.

Hassan Ghand Ali is a lecturer at Payame Noor university of Masjede-Solayman, Iran. He has published and presented several papers at different international journals and conferences.