On the Substantive and Predictive Validity Facets of the University Entrance Exam for English Majors

Kioumars Razavipur

Shahid Chamran University of Ahvaz, razavipur57@gmail.com

Received: 15/03/2014 *Accepted: 23/07/2014*

Abstract

This study examined the substantive and predictive validity facets of the University Entrance Examination for English Major (UEEEM) students. To that aim, 111 English major students were recruited to report their scores on each of the subtests of the test as well as their grade point average. Sequential multiple regressions and factor analysis were used in the analysis of the data. Results acknowledged a 2-factor structure underlying the test. Moreover, multiple regression analyses indicated that the presence of a large number of seemingly construct irrelevant items, Arabic, and theology, coupled with items with no unique contribution to variation to the response variable, the General English subtest, has compromised the predictive validity of the test. The paper concludes with the implications it carries for the test's stakeholders, particularly its developers and score users.

Keywords: Language Tests; Validity; Substantive; Predictive

1. Introduction

Validity is no doubt the most commonly used term in the literature of educational measurement and language testing. The reason for this high frequency is clear: The entire business of testing and assessment in both general and language education comes down to designing and administering valid instruments for the measurement of intended abilities or knowledge areas. Yet, finding completely valid instruments for measuring mental abilities has proved stubbornly elusive. Nor has yet emerged any consensus in the literature as to what the precise nature of validity is and how it should be established (see Bachman & Palmer, 2010; Borsboon, 2005; Fulcher, 2010; McNamara, 2006).

Up until the seminal paper by Messick (1989), the componential paradigm of validity dominated the literature and despite Messick and later scholars' calls for a unified paradigm of validity, the componential paradigm is yet much in vogue especially in other branches of applied linguistics. In the componential view, validity is of many types including, but not limited to, content, face, construct, and empirical validity types. However, in Messick's unified framework, all validity types are only different facets or aspects of a unified concept of validity. For

Messick, no one facet of validity is adequate to establish the appropriateness of the inferences that are made of test scores. Consistent with this unified conceptualization of validity, this paper examines two validity facets, the predictive and substantive, of the University Entrance Examination for English Major (UEEEM), the national language test that is administered annually to screen candidates for undergraduate language programs at state universities in Iran.

Predictive aspect of validity is concerned with whether scores obtained from a measure for which we seek validity evidence correlate significantly positive with those from another measure whose validity is already established (see Bachman, 2008; Campbell & Fiske, 1959). If the UEEEM fulfils on its promise to select the most promising candidates, it must predict the future performance of candidates who gain entry into language programs. In validity lingo, the target language domain (TLU) is English major students' academic performance during their undergraduate programs, which is operationally defined as their grade point average (GPA). As a test is valid to the degree it can predict test takers' performance in nontest situations, the extent to which the UEEEM is capable of predicting GPAs translates into its predictive validity. The substantive aspect of validity relates to what traditionally was defined as *construct validity*: The extent to which a test measures what it purports to measure.

As Farhadi and Keramati (2009) posit, the national tests designed by the Center of Educational Measurement (CEM) are seldom, if ever, subjected to independent validation studies. Given that vital decisions are annually made on the basis of scores obtained from the UEEEM and the enormity of consequences that they have for the large population of its stakeholders, the test is obviously of high-stakes. As Fulcher and Davidson (2009) maintain, a high-stakes test is like an important building, which cannot be suddenly destroyed and then rebuilt from the scratch. Instead, to meet the new needs that arise as a result of new circumstances, they are in need of constant renewal, restoring, and maintenance. Fulcher and Davidson name this process "test retrofit," which is integral to the success of a test for the purpose it is designed for in the first place as well as for the new needs that emerge throughout a test's life span. Such deliberate attempts at evolving a test over time have also been referred to as test maintenance (Kunan, 2000).

The retrofit or maintenance process necessitates continuous gathering of validity evidence and correspondent changes in the test congruent with such evidence. Seen in this light, validation is a never ending process because we never validate a test per se, but we validate the inferences and decisions that are made based on test scores (Messick, 1996). In line with this argument, it is of paramount importance to continually seek evidence about the validity or otherwise of the UEEEM to tailor it to the new needs and demands brought about by new generations

of test-takers, changing needs of beneficiary institutions, or technological breakthroughs.

To narrow the abovementioned lacuna in validation studies regarding UEEEM, the current study constitutes a modest attempt to investigate its substantive and predictive validity facets. The structure of this paper is as follows: A brief report on previous studies which have addressed the various validity aspects of the UEEEM. A conscious decision has been made to limit the review to validation studies related to UEEEM or other similar national tests because a full treatment of the validity theory and all the numerous validation studies that have been conducted on different tests goes beyond the intent and scope of the present paper. Afterwards, details regarding the participants of the study, instruments, and results are presented. The paper ends with a discussion of the findings and suggestions for both the improvement of the test as well as for future research.

2. Literature Review

Studies addressing various validity concerns of international and national high-stakes language tests abound. However, not many studies have been carried out on the UEEEM. The limited number of studies that have been conducted can be categorized into one of the following strands: test washback, test fairness, and test validity. For space and relevance reasons, only studies in the latter two strands are reviewed.

Compared to the validity issue, even far fewer studies have addressed the fairness of the UEEEM, and none has approached the issue taking into account the perceptions of the test-takers. Using a differential item functioning (DIF) approach, Barati, Ketabi, and Ahmadi (2006) investigated the general English module of the University Entrance Examination, administered to high school graduates of math, sciences, and humanities, who sit the test for admission to tertiary institutions. They sought to find out whether any of the test items had bias in favor or against test-takers of different high school backgrounds, namely, math, science, and humanities. They found thirty three DIF items which worked differently across different fields of study. In a similar study but with a different test, Ali Rezaee and Shabani (2010) investigated the differential item functioning of University of Tehran English Proficiency Test across genders. They found that more than one-third of the test items demonstrated significant gender differences.

Studying test method facets, Ahmadi (2011) compared the performance of students majoring in English translation on two tests of translation: one multiple-choice and the other open-ended. No significant correlation was found between the two tests; nor did the students perform better on the multiple-choice test of translation than on the open-ended test, as the researcher expected. Whereas we need

to approach the findings from that study with caution on the grounds that the way open-ended tests of translation were scored remains subjective, the results of the study have serious implications for the UEEEM whose format is predominantly multiple-choice.

Sahrayi and Momghani (2013) studied the reliability and validity of the MSRT test, previously known as MCHE. In their study, MCHE turned out to be of an acceptable overall construct validity and reliability. Nevertheless, the various components of the test when examined individually did not enjoy the same level of validity.

Sheikholeslami (1999, as cited in Sahrayi & Momghani, 2013) investigated the concurrent validity of the general section of the M.A. TEFL examination. The criterion measure he selected was a version of TOEFL. He found high correlations between the grammar sections in the M.A. test and that in the TOEFL test. However, the correlations between the reading and vocabulary sections in the former test with those in the latter were low.

Pishghadam (2003) investigated the predictive validity of the test with two groups of students who differed in their scores on UEEEM. The results showed that participants' scores on the UEEEM had a positive relationship with the participants' GPAs at the end of the first semester. However, the CEM has undertaken numerous changes to the content and structure of the test the reasoning behind which is currently not known. Generally speaking, the test has moved away from the discrete point testing and has gradually evolved into a more integrative one, accommodating more cloze test and pragmatic knowledge items. The type of input, including reading comprehension passages, that are given to test-takers have also changed considerably.

To my knowledge, no research concerning the substantive validity of the UEEEM has been undertaken to date. Nor are there studies addressing the equal predictive validity of the test for different quotas of test takers across participants of various ability levels. In this spirit, the questions we will address in this paper are as follows:

1. What is the underlying factor structure of the UEEEM?
2. Which module or cluster of modules of the UEEEM predicts future performance, defined as participants' GPAs?

3. Method

A total of 111 English major junior students, both translation and literature majors, participated in this study. The participants were studying at Shahid Chamran Univeristy of Ahvaz and were selected on a convenience basis. They were required

to answer a questionnaire which asked them to report their scores on the five modules of the UEEEM as well as their GPAs. It took the participants, at most, 5 min to complete the questionnaire.

It is of note that the data upon which we have based this study consist mainly of objective data, which are not prone to the types of fluctuations associated with attitude scales. We simply asked the participants to report their scores on each of the five modules of the UEEEM as well as their GPAs. The stakes of the UEEEM are so high for test takers that none had problems remembering their scores on the test. Such data are not subject to variations over time or forms as items on most attitudinal measures are. Therefore, the nature of the data relieved the researcher from a concern with reliability or validity issues. In other words, the instrument used clearly elicited what it was intended to elicit, thus enjoying construct validity.

For analysis, to examine the underlying factor structure of the test, exploratory factor analysis was performed. Moreover, sequential multiple regression analyses were run to examine the predictive validity of the test. In so doing, the participants' GPAs were taken as the response variable, and the UEEEM five modules served as the explanatory variables. All statistical analyses were run using SPSS 18.

4. Results

The first research question of the current study is about the underlying factor structure of the UEEEM. Exploratory factor analysis (EFA) with Varimax rotation was performed to examine the latent factors which underlie performance on the test. Factor analysis, like other parametric tests, presents a number of assumptions, which should be satisfied before performing the test. In particular, the normality of distribution, the factorability of the data, and the adequacy of the sample are among the main assumptions of the test. To examine the normality of the distribution, the researcher inspected the kurtosis and skewness of each item—all proved to be within the acceptable range (see Bachman, 2008).

Table 1 shows the results of KMO and Bartlett's test, which confirm the adequacy of the sample as well as the factorability of the data:

Table 1. *KMO and Bartlett's Test*

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.572
Bartlett's test of sphericity	Approximate chi-square	165.543
	df	10
	Sig.	.000

Table 2 gives the EFA outcome, which points to a neat two-factor structure. It could be seen that the two English modules of the test have very high loadings on component one, termed English proficiency (EP) factor, and the other three modules have loaded on the second component, termed the Native Language and Culture (NLC) factor:

Table 2. *Rotated Component Matrix*

	Components	
	1	2
Persian	.194	.801
Theology	-.009	.750
Arabic	.061	.654
General English	.919	.127
Special English	.928	.057

Also checked were the scree plot and the eigenvalues, the results of which were consistent with the two factor structure illustrated in Table 2. It is of note that the Persian literature loaded on the first component as well, but the value of loading is not large enough to be considered significant, as in most texts on factor analysis the minimum value of loading is set at .3 or more (see Hatch & Lazartan, 1991). Therefore, it is concluded that two separate factors comprise the underlying structure of the UEEEM, namely, the EPF and ENF.

The next research question is about the test modules that explain the observed variance in the participants' GPAs as the response variable to be predicted by the specified set of explanatory variable: Special English, General English, Persian, Arabic, and Theology. To explore the relationship between the response variable and predictor variables, multiple regression analysis was used.

As it is possible for two variables to show a statistical correlation without having a linear relationship, the first assumption to be satisfied in performing all statistical analyses resting on correlation is the linearity of the relationship between variables. To do so, the Loess line across the scatter plots for each predictor variable and the response variable was examined: None of the predictors appeared to have a nonlinear relationship with the response variable. Another important assumption of multiple regression analysis is that the data should be normally distributed. Though to date, there is no fail-safe way to determine that the data are normally distributed, the data were inspected both numerically and visually. The numerical tests, skewness of kurtosis, were reported earlier. As for the visual examination, histograms were checked. Neither Stem and Leaf plots nor Q-Q plots indicated any serious violations of the normality assumption. Finally, the assumption of

multicolinearity requires that explanatory variables should not be too highly correlated. To test this assumption, both VIF and numerical matrix of correlations were examined. Though the matrix of correlation pointed to the high correlation of special and general English modules, the VIF columns turns out to be less than 5 (Larsen-Hall, 2010), which indicates that the two predictor variables do not violate the requirements of the multiconlinearity assumption.

What follows are the results from sequential multiple regressions conducted, using scores reported on the participants' GPAs as well as their scores on each of the test's modules. Table 3 is illustrative of the descriptive statistics regarding the participants' scores. The easiest test module, according to Table 3, is the General English subtest ($\bar{x}$ = 75.96) followed by Persian language and literature ($\bar{x}$ = 59.61). In close succession is the module related to Islamic culture and ideology ($\bar{x}$ = 59.25), followed by Special English ($\bar{x}$ = 52.61). The most challenging part is the Arabic language module with a mean of 46.8, which is also the subtest with the largest dispersion (*SD* = 23.95):

Table 3. *Descriptive Statistics of the Participants' Scores*

	Mean	Std. Deviation	*N*
GPA	16.46	1.66	111
Special English	52.61	17.23	111
General English	75.95	16.88	111
Farsi	59.61	17.47	111
Arabic	46.80	23.95	111
Theology	59.25	15.81	111

Table 4 gives the results of the multiple regression analyses. As noted earlier, the participants' GPAs constitute the variable the variance of which is expected to be explained by the five test modules as the explanatory variables. In sequential multiple regression, each variable that is entered into the model accounts for all the areas of overlap it has with the response variable, which is made up of both the unique variation it has in common with the response variable and the overlap it has with the response variable, which is shared by another explanatory variable. As a result, the order of entering explanatory variables should be based on theory, logic, or previous research. The order of entering variables in this study was motivated by both logic and previous research studies:

Table 4. *Multiple Regression Model Summary*

					Change Statistics				
Model	*R*	*R* Square	Adjusted R Square	Std. Error of the Estimate	*R* Square Change	*F* Change	*df*1	*df*2	*Sig. F* Change
1	.25	.06	.057	1.61	.066	7.67	1	109	.007
2	.25	.06	.04	1.62	.001	.003	1	108	.957
3	.39	.15	.12	1.55	.087	10.1	1	107	.001
4	.42	.17	.14	1.53	.024	3.1	1	106	.081
5	.45	.20	.17	1.51	.032	4.2	1	105	.042

a. Predictors: (Constant), Special English
b. Predictors: (Constant), Special English, General English
c. Predictors: (Constant), Special English, General English, Farsi
d. Predictors: (Constant), Special English, General English, Farsi, Arabic
e. Predictors: (Constant), Special English, General English, Farsi, Arabic, Theology

Informed by theoretical rational and common sense in regard to the proximity of each subtest to the construct of English language proficiency, the Special English section was entered first followed by General English. The next variable to enter into the model was Persian subtest on the grounds that one's L1 always affects his or her learning of other languages (see Cook, 2010). The penultimate and the final variables to enter were Arabic and theology, respectively because it is plausible to think that a language aptitude variable underlies the learning of all languages, including Arabic (Stansfield & Winke, 2008). Finally, theology was the last variable to enter as it was deemed to have the least to do with language learning and assessment.

The number in the *R* Square column is what is of interest in Table 4. It is clear that all the explanatory variables taken together account for only about .46 percent of the variance in the response variable, which, given the number of predictors, is not a high portion of the variance. Having said that, the number which matters most is the *R* Square Change, which indicates the magnitude of contribution each of the explanatory variables makes to the model. It is observed that the Persian test module followed by the Special English module explains the largest portion of variance in the response variable, respectively (.087, .066). Obviously, the modules of General English and Arabic add no considerable explanation to the model. Nonetheless, the relatively considerable contribution that Theology makes to the model is counterintuitive given that it is the variable assumed to introduce solely

construct irrelevant variance in test scores. The possible explanations for this odd outcome are discussed further in the next section.

The three multiple regression analyses reported above evidently demonstrate that the UEEEM fails to explain a large portion of variance in the participants' GPAs. It was also revealed that the two subtests of Special English and Persian contribute the largest amount of explanation to variations in the response variable.

5. Discussion

This study addressed the substantive and predictive validity facets of the the UEEEM. Factorial analysis of variance revealed that two mutually exclusive latent traits underlie the UEEEM. The two English subtests loaded highly on the EP factor and the three other subtests loaded on the NLC factor. This factor structure is evidence that the second factor compromises the meaningfulness of the inferences that are made of the test scores, mainly because the three subtests loading on the NLC factor introduce only the construct irrelevant variance to the test (see Barati, Ravand, & Ghasemi, 2013; Messick, 1996).

Regarding its predictive validity, the results from the multiple regression analyses revealed that, overall, the test does not enjoy a satisfactory level of predictive validity as its five subtests, on aggregate, could predict maximally 46 percent of variance in the response variable. The findings should be approached with caution as GPA, as the criterion measure, is itself subject to variation from a wide range of sources. In particular, the fact that it comprises of non-English courses as well as English ones compromises its validity as a representative criterion measure. Moreover, numerous personal and cultural influences are involved in classroom assessments which, in turn, affect the GPA one attains.

What was quite unexpected from the regression analysis was the role played by the Theology subtest in explaining some variation in the response variable. It was surprising because it is the most construct irrelevant of the test modules, hence the decision to enter it in the equation as the last predictor. Two different lines of explanation are discernable in this regard: First, it might simply be because of the considerable number of courses on the B.A. curricula which relate to theological issues, whose scores could not be eliminated from the GPAs in the analyses because it necessitated having access to detailed lists of scores for each participant. The other less likely possibility is that the academic climate of universities, especially in big cities, presents a major a challenge to students who abandon their traditional family life behind. To cope with the sweeping changes, commitment to religious values and beliefs may give some students an edge, providing them with better coping strategies. This, however, is sheer speculation.

Finally, the students' performance on the General English and Arabic explains no variance in the participants' GPAs, meaning that these two test components have no explanatory power in predicting future performance of test takers during their undergraduate programs. In other words, the two modules contribute no unique variance in the response variable. As practicality is a major concern in large scale, standardized assessment (Fulcher, 2010), the presence of a large number of items with no unique contribution to future performance is a waste of limited testing resources. This empirical finding is not rocket science to understand; rather, it can be explained to an audience of minimal statistical literacy. We may not even need statistical evidence to realize that the English language proficiency is measured more effectively if subtests are in close alignment with the construct we aim to measure. The question that remains is why policymakers insist on keeping these obviously irrelevant modules on a test designed for an entirely different purpose. For a plausible answer to this question, we need to go beyond the technicalities of language testing and take into account the larger sociopolitical and power issues that are involved in every testing regime.

Language testing scholars have long realized the crucial roles played by language tests in serving the interests and agendas of power (Fulcher, 2008; Fulcher & Davidson, 2007; Shohamy, 2001). The justification behind introducing subtests that from a technical perspective compromise the validity of the UEEEM should be sought in the sociocultural milieu of postrevolutionary Iran. In the aftermath of the Islamic Revolution, policymakers believed that for individuals to uphold revolutionary values and ideals, having expertise is not adequate; rather, they believed, and continue to think so, that commitment to Islamic ideology is prior to having expertise (see Farhadi & Keramati, 2009; Sakurai, 2007). Regardless of whether one agrees or disagrees with this line of thinking in general, what is technically at issue is to equate having information about an ideology with commitment to it. It is well-known that test takers study whatever they are tested on (see Alderson, 2004; Hamp-Lyons, 1998). Along this argument from the washback literature, performing well on a set of MC tests cannot be taken as evidence of commitment. Though this issue begs more empirical evidence, we doubt that test takers with higher scores on Arabic and Theology subtests enjoy higher levels of religiosity. Nor is it ethical conduct to promote hidden ideologies through tests to the detriment of the core construct that is being measured (Shohamy, 2001). In the long run, such practices count against the very revolutionary ideals policymakers seek to promote, as invalid tests fail to serve the very mission tests have evolved to accomplish, that is social justice.

There are lessons to take away from findings of this study for a broad range of stakeholders in language assessment. First, policymakers would serve national

interests better if they appreciate the vital importance of principled allocation of human resources, which is, in many cases, implemented through tests. National interests would be much better served through tests of adequate substantive and predictive validity. In fact, social justice can never be promoted by tests which are unable to discriminate among target ability levels. Keeping ideologies away in defining constructs under assessment would contribute to more prudent and productive allocation of human and financial resources.

Test designers should reconsider test methods (Bachman, 1990) they use in the design of the UEEEM as the importance of test methods has been voiced by many test scholars (Bachman, 1990). The findings of this study coupled with those from Ahmadi (2011) suggest that, at least, part of the blame for poor predictive validity of the UEEEM is to be placed on its format. As Ahmadi (2011) found, performance on a recognition test does not predict performance on an authentic translation task, which constitutes a big portion of target language use domain for the UEEEM test takers. Making use of more direct, authentic test tasks and items would enhance the congruity of test scores interpretation with the target language use situations outside the test. Another possible reason for the UEEEM's low predictive validity facet with the participants' GPAs is that it measures some language skill areas to the exclusion of others—listening is totally absent, and speaking is measured through a limited number of written conversational items, where test takers should fill in the blanks in a short written conversation. Generalizing from such indirect measures to real-life language use situations is difficult. On the other hand, we know that speaking and listening skills greatly influence students' GPAs in undergraduate programs. Accommodating such language skill areas in the UEEEM design would improve both substantive and predictive validity of the UEEEM.

There are definitely limitations to this research which can be avoided in future similar studies concerning test validity facets. The current study, like the one by Pishghadam (2003), took the students' GPAs as the target variable to be predicted. There is an inherent drawback in this procedure. As noted earlier, score comparability across classes, teachers, and universities is an issue when we take GPA as the response variable. Besides, even if the participants are sampled from the same university, the problem is not resolved because of the myriad of other factors that are involved in the students' performance in college—not to mention the highly subjective nature of the instructors' judgment of performance. Future studies using more reliable measures of student achievement as their criterion measure would help to substantiate or refute the claims made in the present study.

One of the perennial problems in empirical validity studies has always been the choice of a measure whose validity is already established. The criterion measure

(i.e., GPA) may, in fact, be the Achilles' hill of this study as it included the scores of both English and non-English courses. A major improvement that would be welcome in future validation studies of the UEEEM is to subtract the participants' scores on non-English courses from their GPAs so that the criterion measure against which the validity of the test is judged would be a more valid index of the students' achievements in college.

Finally, this study investigated the validity facets with the implicit assumption that L2 learners have independent competences in their L1 and L2. Therefore, the arguments made above hold promise only within such a framework. For instance, if the window through which one looks at bilingualism or multilingualism is that of a multicompetence theory (Brown, 2013), all the findings from this and almost all other validation studies will have to be looked with skepticism—if not with suspicion. For example, in a multicompetence framework of bilingualism, if a test appears to be comprised of two factors, each of which loading on competence in one of the two languages, it would be evidence of validity not to the contrary. Studies carried out within a multicompetence framework would widen our understanding of validity facets of the UEEEM and other national matriculation language tests in which multiple languages are tested in the same test.

References

Ahmadi, A. (2011). On the validity of a multiple-choice translation test as a substitute for an open-ended translation test in the Iranian university entrance examination. *Perspectives: Studies in Translatology. 19*(4), 307-314.

Alderson, J. C. (2004). Forward. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 19-36). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Ali Rezaee, A., & Shabani, E. (2010). Gender differential item functioning analysis of the University of Tehran English proficiency test. *Pazhuhesh-e Zabanha-ye Khareji, 56*, 89-108.

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

Bachman, L. F. (2008). *Statistical analysis for language assessment*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University Press

Barati, H., Ketabi, S., & Ahmadi, A. (2006). Differential item functioning in high-stakes tests: The effect of field of study. *IJAL, 9*(2), 27-49.

Brown, A. (2013). Multicompentence and second language assessment. *Language Assessment Quarterly*, *10*(2), 219-235.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Cook, V. (2010). *Translation in language teaching*. Oxford: Oxford University Press.

Fulcher, G. (2010). *Practical language testing*. London: Hodder Education

Fulcher, G. (2009). Test use and political philosophy. *Annual Review of Applied Linguistics, 29*, 3-20.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*: *An advanced resource book*. London: Routledge.

Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing*, *26*(1), 123-144.

Farhadi, H., & Keramati, H. (2009). Language assessment policy in Iran. *Annual Review of Applied Linguistics, 29,* 132-141.

Hamp-Lyons, L. (1998). Ethical test preparation practice: The case of TOEFL. *TESOL Quarterly, 32*(2), 329-337.

Hatch, E. M., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Boston, Mass: Heinle & Heinle

Kunan, A. J. (2000). Fairness and justice for all. In A. K. Kunan (ed.), *Fairness and validation in language assessment* (pp. 1-14). Cambridge: Cambridge University Press.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly*, *3*(1), 31-51

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell Publishing.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 243-256.

Pishghadam, R. (2003). *The predictive validity of the new version of the state universities entrance examination admitting candidates to English majors*. Unpublished master's thesis, Allame Tabataba'i University, Tehran, Iran.

Sakurai, K. (2007). University entrance examination and the making of an Islamic society in Iran: A study of the postrevolutionary Iranian approach to "Konkur." *Iranian Studies*, *37*(3), 385-406.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing, 18*(4), 373-391.

Stansfield, C., & Winke, P. (2008). Testing aptitude for second language learning. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of language and education: Vol. 7, Language testing and assessment* (pp. 273-284). New York: Springer.

صحرایی، ر. و ممقانی، ه. (۱۳۹۱). ارزیابی روایی و پایایی آزمون زبان انگلیسی وزارت علوم، تحقیقات و فناوری (MSRT). *اندازه گیری تربیتی*، ۱۰، ۲۰-۱.

.