On the Construct Validity of High Stake Tests: The Case of Iranian National University Entrance Exam

Aliakbar Jafarpour Boroujeni*

Abstract

An important assumption in language testing is that test items or observable variables tap the underlying latent traits hypothesized in the theoretical model or constructs governing the design of the testing instrument. Accordingly, the present study sought to investigate the reliability and construct validity of Iran's national university entrance exam. The participants of the study were 100 students who were making themselves ready for the big test in two private institutes of Ghalamchi and Modaresane Sharif. To collect the relevant data, the participants were asked to fulfil a sample test of university entrance exam. The result indicated that the present test is moderately reliable (r=0.5), consulting kr-21. For the justification of the construct validity, this study utilizes exploratory factor analysis and confirmatory factor analysis to examine factors of the empirical data. Through the operation of the empirical data analysis, important results are illustrated by the exploratory factor analysis. It has been stated that several factors did not place under the same trait and therefore they should be removed from the test. The results from the present study call for a recognition of importance of improving the test design for EFL learners, teachers and material developers.

Key words: *construct validity, confirmatory factor analysis, reliability, university entrance exam, EFL learners*

1. Introduction

A well-defined test should have some special characteristics as the sign of being good enough to be trusted. For instance, to understand the functioning of a test, it is important that the test which is used consistently discriminates individuals at one time or over a period of time. In other words, reliability is the extent to which measurements are repeatable - when different persons perform the measurements, on different occasions, under different conditions, with supposedly alternative instruments which measure the same thing. In sum, reliability is the consistency of measurement (Bollen, 1989), or stability of measurement over a variety of conditions in which basically the same results should be obtained (Nunnally, 1978). Also, a well-defined test is the one which can measure the test-takers' talent in the subject as well. On the other hand, a test should be valid enough to be trusted. Validity of a test refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests (Messick, 1992). Traditional approaches to validity have been based on

* Shahrekord University, Iran
E-mail: aliakbar_jafarpour@yahoo.com

three distinct categories of validity evidence: content-related, criterion-related, and construct-related evidence of validity. However, Messick (1989) has argued that viewing these categories as separate lines of validity evidence is inadequate. In his unified framework of validity he has stressed the importance of viewing validity as a 'unitary concept' in which construct validity is an essential component that encompasses content and criterion-related validity. His framework suggests that the consideration of values and consequences of score use has an essential role in validity consideration.

As a result, test developers must define what the test intends to measure and make sure that it truly measures what it aims to measure before they make an inference about an examinee's competence from test scores. The matter of consistency between a construct theory and test performance is associated with construct validity, which is the main focus of this study. Construct validity refers to the extent to which test scores support the presence of the construct that underlies the test. In fact, in construct validation it is aimed to provide evidence that supports specific hypotheses about relationships between hypothesized abilities and test scores (Bachman, 1990).

Correlational approaches to construct validity are the most common path for testing this relationship. Construct validity is based on abstract and theoretical abilities and traits. The process of construct validation in correlational approaches begins with operational defining of the construct, based on theoretical reasoning. This part of the process is similar to the process of content validation. Therefore, construct validation is always dependent on theory (Cronbach & Meehl, 1955). After that the hypothesized relationships between test scores and concepts are empirically examined to estimate the extent to which variance in the test reflects variance in the underlying construct. This approach is based on a confirmatory mode. In the confirmatory mode, we begin with hypotheses about traits or abilities and how they are related to each other and attempt to either confirm or reject these hypotheses by examining the observed correlations. So construct validity is the unifying concept that incorporates criterion and content considerations.

It is essential to mention that estimating the validity needs a precursor, which is reliability. Reliability is one of the most important elements of test quality. It has to do with the consistency, or reproducibility, of an examinee's performance on the test. If a test is reliable, it is worth considering its validity.

Thus, in order to have a trustable and standard test, especially in a vast range, the validity of the test should be taken into account, of course, after calculating the test reliability. Among these tests, Iran's University Entrance Exam (UEE) has been one of the main concentration centers of Iranian researchers. After two decades of criticism (Farhadi, 1985; Jafarpour, 1997 & 2002; Yarmohammadi, 1986) a new test called Department of English Language Test was added to other four test departments (Science, Math and Physics, Humanities, and Art) of the National University Entrance Examination in 2002. The test aims to select students for English language literature, TEFL, and translation associate degree and undergraduate programs.

The results of Iran's national university entrance examination have a huge impact on the future career of the Iranian test-takers. Therefore, for a fair and reasonable competition all its parts must be theoretically and practically scientific. An assessment is called fair and valid if it is based on accurate data that ensure reliability and validity standards.

The present study is an investigation of the reliability and validity of General English Language in Iran's university entrance examination for English language.

2. Statement of the Problem

As it was mentioned above, UEE plays an important role in the future of test-takers in Iran, since their forthcoming career will be chosen based on their field studied in university. The situation is the same in English Department. As a result, the content validity of English language test of Iran's UEE has been tested by Razmjoo (2006). Although construct validity is one of the most important factors to take into consideration, and an essential component in Messick's (1989) framework and the most important criteria for the quality of a test, no research has been done to examine the construct validity of UEE's English language test. Even worse, the test developers and administrators neglect the establishment of the validity of the test before use (Razmjoo, 2006).

Razmjoo (2006) explains that the UEE's English language test does not have content validity. He also argues that the first three or four semesters of all English language majors - translation, teaching English as a foreign language (TEFL), and English literature, are assigned to teaching English language instead of teaching specialty courses such as applied linguistics, translation, or English literature, since the test does not really assess the applicants' language competence which turns out not to be satisfactory for their undergraduate studies.

3. Objectives of the Study

The present study seeks to continue efforts of the reliability and validation of the UEE's general English language test and to address some of the limitations in this field. The main objectives of the study include examining the reliability of the test, and also examining the construct validity of the test.

4. Research Questions

In an attempt to estimate the reliability of UEE's general English language test and to investigate its construct validity, the proposed study will try to collect information in order to answer the following questions:

RQ1: Is the general English section of INUEE reliable?

RQ2: Does the general English section of INUEE have construct validity?

5. Significance of the Study

The results of UEE are very influential for Iranian students. It, therefore, needs careful and extensive considerations and investigations. One cannot prepare such a test overnight, it needs constant evaluation and modification to ensure that it truly reflects the test-taker's aptitude.

The proposed study is significant and essential as it is part of such an effort. The results of this study could be beneficial to Iran's Education Evaluation Organization and its test developers.

6. Theoretical background

6.1. Test Reliability

Test reliability refers to the consistency of scores students would receive on alternate forms of the same test. Due to differences in the exact content being assessed on the alternate forms, environmental variables such as fatigue or lighting, or student error in responding, no two tests will consistently produce identical results. So, reliability involves the consistency, or reproducibility, of test scores. That is, the degree to which one can expect relatively constant deviation scores of individuals across testing situations on the same, or parallel, testing instruments. This property is not an inactive function of the test. Rather, reliability estimates change with different populations (i.e. population samples) and as a function of the error involved. In fact, even the same test administered to the same group of students a day later will result in two sets of scores that do not perfectly coincide. Obviously, when we administer two tests covering similar material, we prefer students' scores to be similar. The more comparable the scores are, the more reliable the test scores are (Brown, 1997).

These facts underscore the importance of consistently reporting reliability estimates for each administration of an instrument, as test samples, or subject populations, are rarely the same across situations and in different research settings. More important to understand is that reliability estimates are a function of the test scores yielded from an instrument, not the test itself (Thompson, 1999). Accordingly, reliability estimates should be considered based upon the various sources of measurement error that will be involved in test administration (Crocker & Algina, 1986).

It is important to be concerned with a test's reliability for two reasons. First, reliability provides a measure of the extent to which a testee's score reflects random measurement error. Measurement errors are caused by one of three factors (Thompson, 1999):

a. teste-specific factors, such as motivation, concentration, fatigue, boredom, momentary lapses of memory, carelessness in marking answers, and luck in guessing,
b. test-specific factors such as the specific set of questions selected for a test, ambiguous or tricky items, and poor directions, and

c. scoring-specific factors, such as different scoring guidelines, carelessness, and counting or computational errors.

These errors are random in that their effect on a student's test score is unpredictable – sometimes they help students answer items correctly while other times they cause students to answer incorrectly. In an unreliable test, students' scores consist largely of measurement error. Also, an unreliable test offers no advantage over randomly assigning test scores to students. Therefore, it is desirable to use tests with good measures of reliability, so as to ensure that the test scores reflect more than just random error.

The second reason to be concerned with reliability is that it is a forerunner to test validity. That is, if test scores cannot be assigned consistently, it is impossible to conclude that the scores accurately measure the domain of interest. Ultimately, validity is the psychometric property about which educators are most concerned. However, formally assessing the validity of a specific use of a test can be a laborious and time-consuming process. Therefore, reliability analysis is often viewed as a first-step in the test validation process. If the test is unreliable, one needn't spend the time investigating whether it is valid – it will not be. If the test has adequate reliability, however, then a validation study would be worthwhile.

6.2. Test Validity

In language testing, validating a test means being able to establish a reasonable link between a test-taker's performance and his/her actual language ability. So, the question in validating a test, as Lado states, is: "Does the test measure what it is intended to measure?" (1965, p.30). As reliability ensures the consistency of a test, its being reliable is a precondition for its validity. It is so important since how can we learn anything about a person's language ability if the test does not even yield consistent results (Alderson, Clapham, & Wall, 2005, p.187)? In fact, talking of a test's validity is quite misleading because what is validated is not the test itself. Rather, it is a matter of validating the inferences we draw and "the interpretations and uses we make of test scores" (Bachman, 1990, p.236). Validity, then, can be seen as a concept allowing us to concern test scores with meaning. This unitary notion of validity has traditionally been subdivided according to the kind of evidence on which the interpretations are based. Usually, one will come across the terms 'construct validity', 'content validity', 'criterion-oriented validity', 'concurrent validity', 'face validity' and 'consequential validity'. It should, however, be understood "that these types are in reality different methods of assessing validity" and "that it is best to validate a test in as many ways as possible" (Alderson, et al., 2005, p.171).

Furthermore, it is believed that in interpreting test-scores, even the most valid and reliable test can only reveal what the testee is able to do, but not what he cannot do. Even the best test cannot rule out the possibility of the test-taker's suboptimal performance due to factors unrelated to the test (Bachman, 1990, p.146). Therefore,

if a testee is unable to fulfill a certain task in a testing situation, it does not necessarily mean that she/he is unable to fulfill this task in real life.

Achieving test validity is an essential concern in test development, mainly when a test is used for high-stakes purposes. However, as Messick commented "many test-makers acknowledge a responsibility for providing general validity evidence of the instrumental value of the test, but very few actually do it" (1992, p. 18). More recently, Weir (2005) reported that while most examinations claim different aspects of validity, they often lack validation studies of actual tests that demonstrate evidence to support inferences from test scores.

Messick presented a unified and expanded theory of validity, which included the evidential and consequential bases of test interpretation and use (1995). Table 1 shows how this theory works. Notice that the evidential basis for validity includes both test score interpretation and test score use. The evidential basis for interpreting tests involves the empirical study of construct validity, which is defined by Messick as the theoretical context of implied relationships to other constructs. The evidential basis for using tests involves the empirical investigation of both construct validity and relevance/utility, which are defined as the theoretical contexts of implied applicability and usefulness.

Table 1 Facets of test validity based on Messik

	Test Interpretation	**Test use**
Evidential basis	Construct validity	Construct validity + relevance and utility
Consequential basis	Value implications	Social consequences

The consequential basis of validity involves both test score interpretation and test score use. The consequential basis for interpreting tests requires making judgments of the value implications, which are defined as the contexts of implied relationships to good/bad, desirable/undesirable, etc. score interpretations. The consequential basis for using tests involves making judgments of social consequences, which are defined as the value contexts of implied consequences of test use and the tangible effects of actually applying that test. For example, the value implications and social consequences issues have special importance in Japan, where the values underlying tests like the university entrance exams and the social consequences of their use are so omnipresent in educators' minds (Messick, 1988, 1989; Green, 1998; Linn, 1998; Lune, Parke, & Stone, 1998; Moss, 1998; Reckase, 1998; Taleporos, 1998; and Yen, 1998.)

Clearly then, while construct validity is still an important concept, our responsibilities as language testers appear to have expanded considerably with Messick's (1995) call for test developers to pay attention to the evidential and consequential bases for the use and interpretation of test scores.

Messick's (1995) unified view of validity predicated that validity is a multi-dimension concept, which can only be established by integrating considerations of content, criteria, and consequences into a comprehensive framework for empirically testing rational hypotheses about score meaning and utility. It is widely recognized that the validation process should start from the very beginning of test development. It is maintained that, in addition to a posteriori validity evidence- which traditionally focused on scoring validity, criterion-related validity and consequential validity - a priori validity evidence - such as test design decisions and the evidence that supports these decisions - also makes a significant contribution to the establishment of validity (Schilling, 2004). Similarly, Weir (2005) by stating that "the more fully we are able to describe the construct we are attempting to measure at the a priori stage, the more meaningful might be the statistical procedures contributing to construct validation that can subsequently be applied to the results of the test" (p. 18), highlights the importance of a priori validity evidence, because the statistical analysis at a posteriori stage does not generate conceptual labels by themselves, and therefore, to make the scores meaningful, the test developers can never escape from the need to define what is being measured at the beginning of test development.

7. Methodology

7.1. Participants

The participants were selected for the present study were one hundred male and female Iranian high school graduate students, with the same L1 background (Persian) 50 males and 50 females. The age of the participants ranged between 19 to 22 which were randomly selected out of the whole population who were getting prepared in the two institutes of Ghalamchi and Modaresaan-e-Shaarif for the university entrance exam. These participants all had similar purposes for continuing education and enrolling in national universities.

7.2. Instruments

According to the syllabus by the Iranian ministry of education , students at different levels of high school were supposed to have acquired with different test formats, such as multiple choice, cloze tests, open ended and open questions, transformation and substitution drills for different sections that are vocabulary, reading, and grammar. The instrument analyzed in this study is specifically grammar and vocabulary, reading comprehension, and cloze sub-tests. Twenty-five four-option multiple choice items divided into three tasks and varying in the number of items tested and their associated task-based theme.

Data for this study was collected in one session of paper and pencil testing. It should be noted that university entrance exam consisted of two sections. One is devoted to specialized questions which are different in terms of participants' major, while the other part is devoted to the participants' shared courses. Our main interest

in this study is on the second part. Participants took part in this section of exam without being informed about it that is, for taking a more valid result. Data then was collected and analyzed for the case of statistical investigation.

7.3. Procedure

The study described in this report investigated the suitability of test items in university entrance exam. Data collection and analysis included the following procedure. It was carried out in one session of paper and pencil without making participants informed about the exam date. Paper tests were passed out and participants were informed about the penalty score that they receive if they make their choices by chance. Then examinees answered all the twenty- five questions , in an accurate time similar to that of konkoor; that is real entrance exam of national university, at the end their exam papers was collected for the further investigation. After reviewing the participants answer sheet, it was revealed that test –takers had missed some sub-tests or even the total paper. Also a few students had answered the paper in a strange way. For example one participant only check A answer for the grammar section, and the second sub-test just B and the third just C, and continued the same way to the end of the test. Such participants were considered as the missed data and were therefore removed from the further analysis; finally the score from the 100 of the participants were included in the main data analysis.

8. Results

Descriptive statistics of each sub-section of the test that are grammar / vocabulary, cloze test and reading comprehension is presented distinctively in Table 2. The table displays descriptive statistics in terms of mean, percentage, standard deviation and the obtained score from. Among 25 questions that was devoted to the general English test, 10 items was devoted to the grammar sections which estimates participants knowledge of grammar, the other section was devoted to the close section that is intended to assess the understanding of the features of written texts as well as grammatical knowledge and pragmatic knowledge about vocabulary in certain context. In other words, cloze section assesses higher –order processing abilities. In the present study the cloze test items appeared to be accounted by two factors: form and meaning. So it can be said that cloze section measures only grammar forms and meaning rather than overall language proficiency. The other ten items evaluated participants' reading comprehension.

G: grammar from 10 scores

C: Cloze Test from 5 scores

R: Reading Comprehension from 10 scores

Table 2. Mean, Standard Deviation, Percentage

Section	Score	Std-deviation	Mean	Percentage
Grammar	10	2.68	4.09	40
Cloze Test	5	2.53	4.37	20
R	10	0.10	2.88	40
Total	25	7.63	3.78	100

As a table 4.1 implies reading comprehension and grammar cover more parts of the test. , meaning that these two parts were somehow paid more attention in test designing process. The descriptive statistics for each of the 25 items were calculated and are presented in Table 4.1. As the means for dichotomous items show the difficulty level of each sub-test, it can be inferred that item difficulty ranged between 4.37(cloze test; the most difficult section) and 2.88 (Reading Comprehension; the easiest), and the standard deviations from2.68 to 0.10.

After investigating the grammar sub-test carefully, we came to the conclusion that test items in the grammar sub-tests were coded into form and meaning items of university entrance exam of 1391 was divided into two formats of form and meaning based on the theoretical model, by which grammatical knowledge is hypothesized to consist of grammatical form and meaning. Therefore, the items were coded and divided into two scales of form (FR) and meaning (MG).

To reach an agreement about certain items, the researchers consulted with an English expert who was familiar with grammatical knowledge categories

Table 3 shows an original taxonomy of 10 test items.

Table 3 Grammatical Taxonomy

Scale	**Number**	**Items**
Form	5	1, 5, 7,8,10
Meaning	5	2,3,4,6,9

The mean score was obtained as 30 and the scores ranged between 15 and 24. The KR-21 for reliability estimates of the test was .50 suggesting that the test was reliable with reference to its internal consistency.

8.1. The result of the first research question

The first question was:

Q1: *Is the general English section Iran's university entrance exam of 1391 reliable?*

In order to investigate the aforementioned research question, KR-21 formula was used to see if the general English test of university entrance exam of 1391 was reliable. Reliability estimates were then calculated based on KR-21 to examine the degree of relatedness among 25 items in the entire test. This method as the most practical and convenient ones of estimating test score reliability is preferred for this study over the other methods of estimating the reliability. The result of the analyses in Table 4 showed that the test was moderately reliable at 0.5 level. Although the overall reliability estimates of each of entire items in the test were not as high as we expect, they exhibited internal consistency within the items of the test.

Table 4 Reliability of the Test Base on KR -2**1**

KR-21= [R= 0.5

More specifically, KR-21 was estimated not only for the total of NUEE test but also for each of its sub-test, that was due to the fact that each of the sub-tests in NUEE was claimed to measure a set of traits/ skills. The results for reliability estimates of the NUEE are presented in Table 5.

Table 5 Reliability Estimates: NUEE

Sub-test	No. of Items	Items	KR-21
1	10	1-10	.652
2	5	11-15	.231
3	10	16-25	.640

KR-21 coefficient of 0.50 showed that the NUEE of 1391 has a moderate internal consistency; however, one of the sub- tests (cloze test) showed a low reliability.

8. 2. The Result of the Second Research Question

Q2: *Does the general English section of university entrance examination of 1391 have construct validity?*

For answering the second research question, we performed a series of explanatory factor analysis (EFAs), using SPSS version 10.1 for the PC to examine the patterns of correlation among the items within each component of

the test, namely Grammar and Vocabulary, Cloze Test and Reading Comprehension. Through the application of Explanatory factor analysis the identifaiability of the NUUE is examined, it indicates whether the items in each component were measuring the same underlying construct and if each component represented an independent construct. For each section, the data were analyzed and evaluated for factor analytic appropriateness. All appropriate decisions were based on the determinant of the correlation. It is a common place in research on different skills that identifiably of skills is addressed by applying factor analysis on the data (e.g., Rost, 1993, Alderson, 2000). The idea that items loaded on the same factor, assess the same skills. This study adopts the same techniques for analyzing data. That is if the items in the NUEE loaded on various factors (i.e. skills) the NUEE sub-tests construct could be identified as multi-visible. By contrast, in case the NUEE items loaded on different factor, hence this hypothesis would be rejected. Before conducting factor analysis, the test items were inspected for the correlation matrix and the correlation coefficient of .3 or above. The result reveled that there were some correlation coefficient of .3 or above. Furthermore, the test of factorability of data (i.e., the Kaiser Meyer Oklin) exceeded the recommended value of .6 (Kaiser, 1974) and the Barlett's test of Sphericity (Barlett, 1954) reached statistical significance. See appendix (4.2.). As Table 4.3 indicates each component of grammar, vocabulary, cloze test and reading was designed to measure three factors that are (morpho-syntactic form, lexical form and literal meaning). It shows that many of the vocabulary and grammar items except items 2, 3 and 10 Located on the same factors. This indicates that some of the vocabulary items were measuring the same trait as many of the grammar items. As Table 5 presents, in the course of analysis 3 items (G, 2 G, 3 and Voc, 10) produced extremely low factor loading (lower than 0.3).

Table 6 Correlation of the items in the grammar and vocabulary section

Grammar and voc section	N	P	R
Morpho- syntactic	5		
Lexical Form	3	0.001	0.248
Literal Meaning	2		
Total	10	0.001	0.248

Correlation is significant at 0.001 level

The cloze section is intended to assess the understanding of the features of written text as well as grammatical knowledge about vocabulary in certain context. In other words, cloze section assesses higher order processing abilities. In the present study, the cloze test items appeared to be accounted for by two factors: form and meaning. As Table 6 indicates cloze section measure only grammar forms and meaning rather than overall language proficiency.

Table 7 Correlations of the items in the cloze section

Cloze Section	N	P P	R
Cloze Form	3	0.018	.009
Cloze Meaning	2		
Total	5	0.018	.009

Correlation is significant at 0.05 level.

The reading section of our study is composed of two discrete passages with five items per passages. After performing factor analysis, we found the two interesting factors of: 1) reading for explicit information and 2) reading for inferential information maximize interpretability. Table 7 produces an interesting result. They indicate that all the items are text dependent. It shows that all factors are moderately correlated at the level of ($p \leq 0.05$). So, we can conclude that all items in the reading comprehension are correlated with a confidence level of 95%.

Table 8 Correlation of the items in the reading section

Reading Section	P	R
Reading for inferential information	0.018	0.334*
Reading for explicit information		
Total	0.018	0.334

Moderate correlation is significant at 0.5

9. Discussion and Conclusion

In the applied sciences, researchers have shown growing interest in exploring human characteristics and their underlying intelligence. For example, in the field of education, psychology, and sociology, researchers design questions to measure the attitudes or opinions on issues of the concern. With the wide availability of computing technology in the late 20th century, factor analysis becomes a common efficient tool to ascertain the underlying construct of the studied characteristics. Moreover, with the wide availability of computing technology in the late 20th century, factor analysis becomes a common efficient tool to ascertain the underlying construct of the studied characteristics. Based on the inquiry purpose, exploratory and confirmatory factor analyses are

distinctively named to fulfill their tasks. The former approach focuses on the acquisition of a factor structure accounting for the relationship with observed data, while the latter intends to test the hypothesized factor model.

Although the result of this analysis cannot be generalized, the analysis of construct validity of Iran's university entrance examination provides the following results:

As far as tests are useful tools to make decision about people's lives, language testing is becoming more and more challenging. For this reason, tests should provide an accurate picture of the test takers' ability to enable test users to make fair decisions. This is what makes testing complex. One of the functions of tests in Iran is for universities entrance examination. The tests were conducted monotonously for all participants to select students for universities and institutes of higher education. A large group of students from urban and rural areas compete for selection into universities and the only criterion are their scores in this test. The result of the test may have a permanent and deep effect on examinee's and their family's morale and future destiny. Bachman and Palmer (2000) define high-stake decisions as decisions that are likely to have a major impact on the lives of large numbers of individuals. During the course teachers receive feedback on the students` weaknesses and strengths and their extent of progress using language tests, the results of tests also provide teachers with the feedback that helps them identify the effectiveness of the approaches they have employed in their teaching. Since high stake tests are employed to make important decisions, researchers have done much research in this area. It is important to carry out a study on the analysis of the items of entrance examination in universities in Iran to detect item bias due to the effect of the examination on their life, language and culture. The main question of this study is to determine the reliability and construct validity of such a high-stake test. Bachman (2002) argued that a language test should be designed taking task characteristics into account as well as the construct definition of language ability in order to achieve the intended purpose of the test. Nevertheless, the testing of language knowledge—like language itself—should not occur in a vacuum, and the design of tests should reflect this fact. Bachman (2002) argued that a language test should be designed taking task characteristics into account as well as the construct definition of language ability in order to achieve the intended purpose of the test.

Although the current study can contribute to recent discussions concerning the importance of both construct definitions and test task characteristics in L2 performance, the importance and critical role of context in the process of test development and test score validation (Chalhoub-Deville, 2001, 2003), further empirical studies are required to elaborate how elements of context can be incorporated in a theoretical construct and the process of test design.

Human characteristics have always make researchers interested to explore the effect of them on underlying concepts of intelligence. For example, in the field of education, psychology, and sociology, researchers design questions to measure the attitudes or opinions on issues of the concern. The areas of language that received the greatest emphasis in the classroom were exactly the ones that received more attention and weight in UEE tests. Therefore we can conclude that the areas that receive little attention in UEE will be considered as secondary

practices in language classrooms. That is why grammar and vocabulary receive ample attention in language classes and listening, speaking, and pronunciation, receive scant attention. This does not, however, mean that teachers are not aware of the importance of these areas. In fact, teachers' responses verify that EFL teachers, especially pre-university teachers, believe that speaking is an important skill. However, there is a mismatch between what teachers do what they actually focus upon in their classes. Since tests are used to make decision which influence people's lives, therefore testers must be accurate about making fair test. In sum, the purpose of testing is to gather quantitative information about the degree of examinees command in particular area.

In sum, by accounting for all the present study's limitations, the results of this study can be cautiously used as a piece of supporting evidence in the validity argument presented for the Iran's' national university entrance exam.

As far as NUEE is in the multiple-choice format, teachers can use this format to see what points need to be explained more.

10. Pedagogical Implication of the Study

The results of this study can be beneficial for different parties involved in educational system. These implications are presented as the following:

The utmost implication of this study may also be of great values for syllabus designers, material developers, and educational managers, when preparing and designing English tests.

10.1. Implication for EFL Teachers

The results of this study have some implications for EFL teachers teaching. Firstly, the separation of testing from teaching and learning is somewhat impossible. As Heaton (1988) argues, testing and teaching are so interrelated that it is impossible to work in either field without being concerned with the other. Language testing is served by the research undertaken in such fields as language acquisition and language teaching (Buck & Tatsuoka, 1998a; 1998b). Language tests can be valuable sources of information about the effectiveness of learning and teaching. According to Spratt (2005), teachers play a significant role in determining the type and intensity of wash back effect, and they can be considered as one of the sources of promoting positive wash back.

Because many Iranian EFL teachers are not familiar with the adverse effects of teaching for the UEE, they try to adjust their methodology to the requirements of that test. Therefore, they need to become aware of the effect of UEE and try to minimize the negative wash back effects. As far as NUEE is in the multiple-choice format, teachers can use this format to see what points need to be explained more.

10.2. Implication for Syllabus Designers

Test developers should change the trend of this test and employ analytical approach in designing UEE. The present study attempted to determine the influence of UEE on EFL teachers` methodology and test development in high schools and pre-university centers. However, this research could only cover a small area of the subject matter. Further research can be conducted with more participants in other situations.

References

Alderson, J.C., Clapham, C., Wall, D. (2005). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

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

Bachman, L. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. and Palmer, A. (2000). (3rd ed.). *Language Testing in Practice*. Oxford: OUP.

Bartlett, M. S. (1954). A note on the multiplying factors for various chi square approximations. *Journal of Royal Statistical Society*, 16 (Series B), p. 296-298.

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* New York: John Wiley & Sons.

Brown, J. D. (1997). Statistics corner: Questions and answers about language testing statistics: Reliability of surveys. *Shiken: JALT Testing & Evaluation SIG Newsletter, 1* 2, p. 18-71.

Buck, G., & Tatsuoka, K. (1998a). Application of rule-space methodology to listening test data. Language Testing, 15, 118-142.

Buck, G., & Tatsuoka, K. (1998b). Application of rule-space procedure to language testing: Examining attributes of a free response listening test. Language Testing, 15(2), 199-157.

Chalhoub-Deville, M. (2001). Task-based assessment: A link to second language instruction. In Bygate, M., Shehan, P., & Swain, M. *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing,* p. 210-228, Harlow, UK: Longman.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing,* 20, p. 369-383.

Crocker, L. M., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52 (4), p. 281-302.

Farhadi, H. (1985). A study of English language test in national university entrance examination. *Roshd Journal of Language Learning,* 4, p. 10-15. In Persian.

Green, D. R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement, 17*, 2. p. 16-19.

Jafarpour, A. (1997). A review of English language test of graduate entrance examination. Department of English Language. *Journal of Isfahan University (Humanities)*, 8 (1), p. 15-22. In Persian.

Jafarpour, A. (2002). *An analysis of the methods and issues of university entrance tests*. Seminar Proceedings (p. 63-78). Isfahan: Isfahan University of Technology Press. In Persian.

Heaton, J.B. (1988). Writing English Language Tests Harlow: Longman

Kaiser, H. F. (1974). An index of factorial simplicity. Psychometrica, 39, 31-36.

Lado, R. (1965). Language Testing: A Construction and Use of Foreign Language Tests: A Teacher's Book. N.-Y.: Longman

Linn, R. L. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. Educational Measurement: Issues and Practice, 17(2), p. 28- 30.

Lune, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement, 17* (2), p. 24-28.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' and performances as scientific inquiry into score meaning. *American Psychologist*, 50 (1995), p. 741-749.

Messick, S. A. (1992). Validity of test interpretation and use. *Encylopedia of Educational Research (6th edition).* New York: Macmillan.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.), p. 13- 103. New York: Macmillan.

Messick, S. (1988). The once and future uses of validity: Assessing the meaning and consequences of validity. In Wainer, H & Braun, H.I. (Eds.). *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum, p. 33-45.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement, 17*(2), p. 6-12.

Nunnally, J. C. (1978). Psychometric theory. (2nd ed.). New York: McGraw-Hill.

Razmjoo, A. (2006). A content analysis of university entrance examination for English majors in 1382. *Journal of Social Sciences and Humanities of Shiraz University*, 23(1), p. 67-75.

Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement, 17* (2), p. 13-16.

Schilling, S. G. (2004). Conceptualizing the validity argument: An alternative approach. *Measurement*, *2*, p. 178-182.

Spratt, M. (2005). Washback and the classroom: the implications for teaching and learning of studies of washback from exams. *Language Teaching Research,* 9, 1, p. 5–29.

Taleporos, E. (1998). Consequential validity: A practitioner's perspective. *Educational Measurement, 17* (2), p. 20-23.

Thompson, B. (1999). Understanding coefficient alpha, really. Paper presented at the Annual meeting of the Education Research Exchange. College Station, Texas, February 5, 1999.

Weir, C.J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Basingstoke: Palgrave MacMillan.

Yarmohammadi, L. (1986). A review of English language test of national university entrance examination year 1986. *Shiraz University Journal of Social Sciences and Humanities*, 2 (1), p. 80-88. In Persian.

Yen, W. M. (1998). Investigating the consequential aspects of validity: Who is responsible and what should they do? *Educational Measurement, 17* (2), p. 5-6.