A Fairness Issue: Test Method Facet and the Validity of Grammar Sub-tests of High-Stakes Admissions Tests

Roya Khoii, Nazli Shamsi
Islamic Azad University, North Tehran Branch

Abstract

The scores on tests of English as a foreign language are widely used for admissions purposes. Given their high stakes nature, such tests need to enjoy a high level of construct validity. This study investigated the effects of different test formats on the measurement of grammatical knowledge by comparing the construct validities of two different formats of error-identification grammar tests (one with four options for each item and one with a no-error option as the fifth option) and a multiple choice grammar test. After administering all the three tests to 131 Iranian EFL learners, it was found that while the TOEFL error-identification test enjoyed the highest level of construct validity, the no-error option in the SAT error-identification test reduced its construct validity to a considerable degree, indicating that including no error options in admissions test does not allow an accurate evaluation of examinees' grammatical knowledge. Moreover, it lent support to the idea that the test method facet could be a strong source of bias in language testing affecting the fairness of the decisions in relation to admitting students to specific academic programs.

1. Introduction

One of the fundamental uses of testing in an educational program is to provide information for making decisions with regard to individual admissions. Generally speaking, educational admissions are based on measurement contributing data to institutional decisions about whether and on what basis to admit students for study in an institution, college, university or program. In this regard, admissions-test scores serve as a kind of "common metric" for expressing students' ability or preparedness on a common scale. When the number of applicants exceeds the openings available, the admissions-tests results are used to identify, select and admit the ones with the potential for success.

Every year many students stake their futures on university admissions tests; however, only a few of them are admitted. In such situations the measurement device which is a test should be constructed with utmost care to represent data of high credibility. Therefore, every test which is administered deserves some kind of evaluation because not all tests are well-developed, nor are all testing procedures wise and beneficial. In order for a decision to be fair, tests must enjoy a high level of construct validity.

Scholars in the field of language testing have continuously tried one means or another to find a reliable, valid and practical measure of different aspects of second or foreign language. Although there are some arguments against using multiple choice (MC) techniques, the demands of an educational context with large numbers of test-takers, and the need for fast marking might make the use of these tests inevitable. The TOEFL and the SAT are two standardized multiple-choice tests which are used as admissions tests, and many universities, institutions and organizations in different countries, including Iran, use the scores on these tests or their modified versions for admission into MA and PhD programs.

2. High-Stakes Tests

According to Loschert [1], high-stakes tests are assessments in which "Students, teachers, administrators, and entire school systems must account for student performance" (p. 1). Tests that are used to make high-stakes decisions are frequently standardized assessments, such as the TOEFL or SAT. Students' scores on these tests may be used to determine promotion to the next grade level, which curricular track students will follow in school, whether or not they will graduate, or which university they will be admitted to. Since high-stakes tests are meant to raise standards for student learning, test takers may be challenged to meet higher levels of academic achievement than before. On the other hand, the scores on such tests play an extremely significant role in determining the future life of students, their families, and society in a broader perspective.

As beneficial as it may seem to use standardized high stakes tests for university admissions purposes, some complications arise concerning the validity and reliability of such tests for applicants to higher education abroad. Educators must consider what is actually being assessed by any given test: Is the test measuring test takers' academic knowledge and skills, or is it primarily a test of their language skills? When applicants take standardized tests, the results might be affected by different factors such as their knowledge, gender, and culture, as well as the test

method facet, therefore weakening the test's validity for them. Now, if the applicants fail to demonstrate their knowledge due to the presence of bias regarding any of the above factors, the test itself will work in favor of some students and to the disadvantage of some others. As a result, it will not meet the criteria of fairness in testing, and the test results will not be a valid reflection of what the students know and can do [2].

3. Fairness and standardized tests

Fairness has been defined as the absence of bias or arbitrariness in measurement. Although test fairness has been a fundamental concept in the evaluation and in the forefront of discussions in the field of language assessment since the late 1990s, many scholars have argued that this concept is over-reaching and ambiguous, and its pursuit in language testing is both unattainable and unnecessary [3]. Fairness is one of the key issues that concerns people about any testing procedure. This concept applies particularly to the use of tests for selection, but it is rarely a concern for any decision based on the test results. In fact, fairness is a social rather than a psychometric concept and its definition depends on what one considers to be fair [4].

Although standardized tests have been developed and used as a useful tool to provide information for making decisions with regard to individual admissions, they are not the most appropriate tool for measuring high stakes, and their function as the only means to assess a school or a student's academic performance is under question. Many of the standardized tests are not properly developed, which may lead to high failure rates, increase the number of students dropping out of school, impose loss of self-esteem, and hinder educational advancement.

4. Bias and test fairness

In order to make sure that the test is fair, a few factors or pillars of testing should be considered. Bias is one factor which can be a threat to the fairness of a test [3]. A biased test is one in which there are systematic differences in the meaning of test scores associated with group membership, which means people from two groups who have the same observed score do not have the same standing on the trait of interest.

The possible sources of bias which can reduce fairness to a considerable degree include poor test administration, inappropriate test choice, and test taker's expectation. This is almost in accordance with how Society for Industrial and Organizational Psychology or SIOP [4] views fairness in terms of the equitable treatment of all examinees, which includes testing conditions, access to practice materials, performance feedback, retest opportunities, and other features of test administration.

McNamara and Roever [5] are concerned with how fairness reviews and codes of ethics link testers to stakeholders and to the wider language testing community. Fairness reviews are used by large testing organizations such as the Educational Testing Service (ETS) to attempt to eliminate bias before it occurs and ensure test content that will not be seen as controversial. The International Language Testing Association's Code of Ethics (2000) and Draft Code of Practice (2005) are designed to raise ethical awareness and to inform practice.

Fairness concerns are an integral part of the development and scoring of all tests. ETS [6] lists the many activities that ensure fairness as following:

- Involvement of minority educators and representative committees in every phase of the development and scoring processes
- Multiple fairness evaluations by trained reviewers
- Routine analyses of test questions to establish that questions do not unfairly contribute to group differences
- Rigorous training for all persons involved in the development or scoring of test questions to ensure that all examinees have an equal opportunity to demonstrate their skills and abilities.
- Appropriate accommodations (e.g., alternate formats, extra time) for examinees who may need them.

Jackson [7] argues that in an ideal world not only would the testing process be fair, but the results of testing would lead to a fair outcome. Therefore, he illustrates the main components of any consideration toward fairness as fairness of outcome, and fairness of process. Fairness of outcome is measured by the reliability and validity of the testing process; that is, the more reliable and valid the test, the less the measurement error. Therefore, it is not possible to guarantee a fair outcome without perfect validity. In addition to the quality of outcome, the process of achieving that outcome should be fair too. In other words, fairness is not enough, and the admissions processes must also be perceived as fair. However, a fair process does not necessarily guarantee a fair outcome. A process can be defined to be fairer when there is less systematic error (bias).

5. Validity

Messick [8] defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness and actions based on

test scores or other modes of assessment." His view of validity refers to the degree to which we are justified in making an inference to a construct from a test score, rather than a property of a test. That is, the behavioral inferences that one can extrapolate from test scores is of immediate focus. In order to validate an inference, not only the validation of score meaning is required but also the validation of value implications and action outcomes for particular applied purposes and of the social consequences of their use are important [8]. In order to be valid, the inferences made from scores need to be "appropriate, meaningful, and useful [9]."

One important issue with regard to test validity is whether the context can affect the test validity. Sometimes a test which is valid for a purpose in one setting might be invalid or have different validity in a different setting, or it might be invalid for another purpose in the same setting. Although certain kinds of tests may work well in some situations, they cannot be used in other situations unless their validity in the new settings is checked. Since validity is the only essential justification for test interpretation and use, professional judgments are required about the tests' validity in each measurement enterprise [8]. Therefore, the validity of all L2 tests needs to be investigated before using them for admissions purposes with different nationality EFL learners.

6. Validity and Fairness

Validity and fairness are intertwined [10]. If an assessment is unfair and misrepresents the standing of some individuals or some groups of individuals on the construct being measured, it is not valid to be interpreted or used. In a similar way, if an assessment is not valid in the sense that it tends to generate misleading conclusions or inappropriate decisions for some individuals or groups it will be unfair. The difference between validity and fairness is that the former focuses on the accuracy and appropriateness of score-based interpretations and decisions about all of the individuals, and the later focuses on group differences and on differences in the accuracy and appropriateness of interpretations and decisions across groups.

When fairness is linked to validity, the scope of fairness investigations, the kinds of evidence needed, and the means to organize and integrate fairness evidence depend on how validity is conceptualized and structured. Considering the fact that it is almost impossible for a test to be perfectly fair for the intended use, it is necessary to find a systematic way to identify areas where research and practice are most needed to focus resources on the key areas.

The relationship between validity and fairness depends on how broadly these two concepts are defined [10]. If validity is defined narrowly and fairness broadly, then validity could be conceived of as a component of fairness; but if fairness is defined in terms of a specific technical issue, like predictive bias and validity broadly, then fairness could be conceived of as a component of validity.

Kane [10] inclines to define both validity and fairness quite broadly, and so none of these concepts is completely included within the other. Rather, he sees them as closely related ways of looking at the same basic question. Applying this view to fairness in assessment, there would be two types of fairness: procedural fairness which requires that all test takers be treated in essentially the same way, take the same test or equivalent tests, under the same conditions or equivalent conditions, and their performances be evaluated using the same rules and procedures. That is, testing programs should be free from bias in the materials and procedures used. Procedural fairness is an essential requirement for both fairness and validity in testing. The other type of fairness, substantive fairness, in testing requires that the score interpretation and any test-based decision rule be reasonable and appropriate.

The development of conceptual frameworks for fairness in language testing has greatly expanded the scope of fairness; however, there has not been enough empirical research works in these frameworks. Xi [11] considers fairness as an aspect of validity, and surveys three approaches; the first approach sees fairness as an independent test quality apart from validity; in the second approach, fairness is an all-encompassing test quality which consists of different facets including validity; and in the third approach fairness is directly linked to validity. Xi rejects all these approaches and proposes her own approach which draws on current argument-based methods of test validation to systematize fairness investigations. She believes that fairness should be conceptualized as 'comparable validity for identifiable and relevant groups'. However, Kunnan [12] has doubts about the general usefulness of this approach as applied to language assessment and says he does not see how current validation frameworks provide the means to investigate all areas of test fairness. He also asks, if current validation frameworks can cover fairness matters too, why would a fairness argument need to be positioned within a validation argument? He adds that nesting or embedding fairness within validity of test score interpretations and decisions is a disservice to both concepts as the focus of research agendas could be diffused and confused. Finally, he argues that reducing fairness matters into a series of rebuttals of validation studies diminishes the role given to fairness, as it can only react to validation arguments rather than set its own agenda." [12]

7. Test format and fairness

The two major factors interacting in the process of testing are trait and method. Trait is the knowledge that is being measured, and method refers to the specific procedures or techniques for assessing the trait. Bachman [13] notes, "Characteristics of the test method can be seen as analogues to the features that characterize the context of situation." Since a given trait can be assessed through different methods, these methods can have different effects on that trait and the test-takers' scores.

The item format may limit or prevent certain construct elements from being included in the test, or otherwise interfere with it, causing distortions in the scores with the possible result that they no longer reflect the construct very well [14]. In other words, the format may make candidates think in certain, possibly undesirable ways. A characteristic of a good test is one in which the method has little effect on the trait. Therefore, it is necessary to develop a framework for delineating the specific features or facets of test method in order to understand variation in language test performance.

According to Willingham and Cole's fairness framework [15], the comparability of opportunity for examinees to demonstrate relevant proficiency, comparable assessment exercises and scores, and comparable treatment of examinees in test interpretation and use are the most important factors affecting fairness. The first quality is most relevant to the design stage in the assessment process; the second one is relevant to development and administration, and the third one to test use. They also propose that fairness issues be organized around four stages of the assessment process: design, development, administration and use. They argue that at different assessment stages, different validity issues are relevant, which determine the relevant fairness investigations. Therefore, in investigating test fairness, the design stage is one of the steps which should be taken into account. One important factor of test design is the decision about the format of a test. Not only can the test format affect test performance, but also the test validation.

Test performance seems to be greatly affected by test method. The test method effect is considered important because it is not known whether test performance is due to the test takers' knowledge or their ability to answer certain formats. Different testing methods produce different degrees of difficulties for the test taker. Therefore, each has a significant effect on students' scores in that construct. Moreover, different testing methods which aim at measuring the same trait can be significantly different from one another, and thus lead to different scores for the test taker. Testers' familiarity with one method can be one reason of this difference, which can be considered as a source of bias, if the other group are not familiar with that format.

Shohamy [16] reviews two major sources of bias or systematic errors as those associated with the test itself, such as method effects and those associated with the consequences and uses of language tests. Bias related to method examines how the method of testing affects the scores. As she illustrates, it may examine whether success on a reading comprehension test implies that the test score is more dependent on how it is being measured, e.g. by multiple choice, open ended questions, than on the trait of reading comprehension. Therefore, if the test takers had been tested by a different method, they would have obtained different scores. Method research proceeds mostly through empirical studies where the scores of matched groups of test takers are compared as a result of taking tests in two different methods. Studies on method effect, indicate that the test itself affects the scores that test takers obtain on tests. Thus, aspects related to item format, genre or testing tasks are found to affect test taker scores resulting in bias against certain test takers. Bias related to test use implies that a language test is being used for unfair purposes, such as forcing students to learn, teachers to teach, create fear and narrow the learning domain. Testers and test users need to consider ways of minimizing these unethical uses of language tests.

8. Testing grammar

In recent years, grammar has regained its importance in language teaching. Nowadays, it is believed that grammar is too important to be ignored, and without the knowledge of grammar, it is not possible to create many new sentences.

There are different formats for testing the knowledge of grammar; however, one of the most widely used types is the multiple-choice (MC) format. According to Gergely [14], "The declining fortunes of grammar appear to have affected the technique most closely associated with it: multiple-choice (MC)." He also argues that there has been little research in recent literature on MC format test, especially in connection with grammar testing. Nevertheless, the usefulness of MC items is limited. Gergely [14] proposes that MC items leave a part of the test untested and are, therefore, unsatisfactory for many testing purposes. More specifically, in relation to grammar testing, the MC format is seen unsuitable because the concept of grammar has broadened over the years. Another shortcoming of MC tests is the difficulty of writing good MC items. As mentioned before, despite all the arguments against multiple–choice grammar tests, these items are used in commercial tests such as the TOEFL and SAT. Beside the fact that the scores of these tests are objective, and they can be scored easily, these tests can help teachers and students to locate the areas of

difficulty since they test the knowledge of each component separately.

Gergely [14] clarifies some different types of MC items as follows:

- *Standard type*: A four-choice sentence-based item which presents an incomplete sentence stem followed by four options for completing the sentence.
- *Multitrak items*: Items which have the unacceptable choice as the correct answer, which is the issue of "testing negatively". The error-identification grammar tests which are used in the TOEFL and SAT admissions tests are included in this group. Unlike the standard type, such items do not require the students to complete the sentence; instead, they have to find the part containing an error.

Clearly, error-identification items are useful for testing the grammar points which have few logical options. There is also another format of error-identification items in which students are given correct sentences together with the incorrect ones and are required to choose the fifth option, which will be the No-Error option if the sentence does not contain any error. However, in practice this method does not work too well, since many students tend to regard every sentence as having an error. Another argument against this method is its emphasis on more negative aspects of language learning. It is believed that simply recognizing an error is not sufficient, and the students ought to be encouraged to concentrate on recognizing and producing the correct form. Moreover, exposing students to incorrect forms of language is undesirable. Nevertheless, since such items are related to the skills required for checking, editing, and proof-reading any report, article, paper or essay, they are used in many tests such as the TOEFL and SAT.

Although the usefulness of Multiple-Choice items has been proved to be limited, they are still common in many educational contexts because they meet the criteria of practicality and can provide a wide range of scores to produce a fewer number of candidates with the same score. In such contexts, it is desirable to investigate the effects of these tests on students' performance and to find out which format is a better indicator of the students' abilities on the construct the test is going to measure. Therefore, MC tests require close investigation of their validity.

There are two formats of error-identification tests used in Iranian university admissions tests. Some include an error identification section which is similar to the SAT error-identification tests, in which the students are exposed to both correct and incorrect sentences, and each item involves a no-error option as the fifth option. Some others include an error-identification section similar in format to TOEFL grammar tests without a no-error option.

The following represent typical items appearing on the sample TOEFL and SAT error-identification tests as used in Iranian university admissions tests:

Sample error-identification item on the TOEFL:

Halifax is the largest (A) city and chief port (B) of Nova Scotia and is (C) the eastern terminus of Canada's two great railway system (D).

Sample error-identification item on the SAT:

The crowd, which (A) clamored for the play to begin, were surprisingly rowdy (B) for a Broadway audience (C) (D). No error (E)

9. Method

This part summarizes the data related to the research questions, participants of this study, and the used instruments for data collection.

9.1. Research questions

This study aimed at answering the following questions:

- Do different test formats used for the measurement of grammatical knowledge enjoy the same level of construct validity?
- Do scores on different formats of grammar tests correlate with each other?
- Is there a significant difference between the scores on different formats of grammar tests?
- Is it fair to use standardized tests for admissions purposes?

9.2. Participants

131 students at two different universities in Iran participated in this study. All the participants were undergraduate male and female Iranian students majoring in English translation, literature, and teaching who had passed at least 100 credits. The rationale behind selecting senior students was to have more proficient individuals.

9.3. Instruments

Three tests were used in this study. The first was a standardized 30 item multiple-choice grammar test

(modeling the TOEFL grammar section) in which the students had to choose the best option in order to complete each sentence.

The second was an error-identification test including 30 items each containing an error. This test was a modified version of the TOEFL error-identification test, and its items had been selected from well-known TOFEL preparation books.

The third was an error-identification test including 30 items each containing a no-error option. This test was a modified version of the error- identification part of the SAT, and its items had all been selected from Barron's Verbal Workbook for the New SAT.

All the instruments tested the same structural points and shared the same table of specifications. The three tests were corrected and scored separately, which provided 3 sets of scores for each individual. One point was given to each correct answer and no point to incorrect ones. There was no penalty for incorrect answers.

10. Data analysis and results

After administering the three tests to all the 131 participants, descriptive statistics were calculated for each of them. Then the reliability quotients of the tests were computed using the Cronbach's alpha. If a test enjoys strong internal consistency, it should show only moderate correlation among items. For exploratory purposes 0.60 is accepted; for confirmatory purposes 0.70 is accepted; and 0.80 is considered good [17]. The reliability quotients of the tests were .70, .70, and .76 for the MC test, the TOEFL, and the SAT, respectively, which were satisfactory.

10.1. Construct validity

In order to provide an answer to the first question of the study, the scores on the three tests were subjected to a factor analysis. This analysis was done to determine whether all these measures shared some common variance and, thus, could be said to tap the same underlying construct. To ensure higher precision, a principal axis factoring (PAF), as opposed to a principal components factoring (PCF), was employed to extract the initial factors. There are many ways to determine how many factors to extract However, as suggested by Sharma [18] and Zwick and Velicer [19], the eigenvalue-greater-than-one was selected as the extraction rule. This rule suggests that those factors whose eigenvalues (sum of squared loadings) are less than unity be excluded from the analysis. Table 1 shows that only one factor with eigenvalue more than one (2.13) was extracted, and all the tests loaded on the same underlying factor, that is, factor 1. Factor 1 also explained 60.18% of the total variance; in other words, more than half of the variance produced by the measures entered into the analysis was due to Factor 1, which can be best interpreted as accounting for students' grammatical knowledge.

Table 1. Total variance explained by factor analysis

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.13	71.08	71.08	1.80	60.18	60.187
2	.594	19.79	90.87			
3	.274	9.12	100.			

Note: The extraction method is Principal Axis Factoring.

Table 2 indicates that all the measures enjoyed high loadings on Factor 1. The highest belonged to the TOEFL error-identification test (.97) and the lowest to the SAT error-identification test (.59).

Table 2. Results of factor analysis

	Factor 1
MC test	.703
TOEFL	.978
SAT	.592

Although all the tests measured the same construct to a large extent, it seems that the no-error option of the SAT error-identification test significantly reduced its construct validity.

10.2. Correlational Analysis

In order to answer the second question of the study, a series of correlations between the scores on the three tests were calculated (Table 3).

Table 3. Correlation coefficients among the scores of the three tests

		TOEFL	SAT	MC test
TOEFL	Pearson Correlation	1	.581**	.690**
	Sig. 2-tailed		.000	.000
	N	131	131	131
SAT	Pearson Correlation	.581**	1	.417**
	Sig. 2-tailed	.000		.000
	N	131	131	131

MC test	Pearson Correlation	.690**	.417**	1
	Sig. 2-tailed	.000	.000	
	N	131	131	131

Note: Correlation is significant at the 0.01 level (2-tailed).

Table 3 reveals that the correlation between the TOEFL and MC tests of grammar was 0.69, while the correlation between the SAT and the MC tests of grammar was 0.417, which, although significant, was not promising. The correlation coefficient between the TOEFL and SAT error identification tests equaled +.58, which was significant at the 0.001 level (2-tailed test, at 129 degrees of freedom), but it was not very high. Thus it was concluded that the students' scores on the TOEFL error-identification test had a moderate correlation with their scores on the SAT error-identification test.

10.3. Mean comparisons

In the pursuit to provide an answer to the third research question, a one-way analysis of variance (ANOVA) and a post-hoc Tukey's test were carried out to compare the students' mean scores on the MC grammar test and the TOEFL and SAT error-identification tests (Table 5).

Table 4. ANOVA results for mean differences among the three grammar tests

source of variance	Sum of squares	Df	Mean square	F	Sig.
Between groups	977.42	2	488.71	19.52	.000
Within groups	9761.42	390	25.029		

According to Table 4, the obtained F ratio equaled 19.52, which was significant at p<.000 level (the degree of freedom was 2.390 for all the tests) suggesting that the differences among the means were significant. The significance of the F ratio in the analysis of variance indicates that there is at least one significant difference between the means of at least one pair of the groups compared [20]. In order to find out which two means are significantly different from each other, post hoc or follow-up tests are required. The highlighted values in Table 5 delineate the differences between the students' means on different tests.

Table 5. Multiple comparisons

(I) test	(J) test	Mean Difference (I-J)	Standard error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
MC	TOEFL	1.97*	.62	.004	.52	3.43
	SAT	3.86*	.62	.000	2.40	5.31
TOEFL	MC	-1.97*	.62	.004	-3.43	-.52
	SAT	1.88*	.62	.007	.43	3.33
SAT	MC	-3.86*	.62	.000	-5.31	-2.4
	TOEFL	-1.88*	.62	.007	-3.33	-.43

Table 6 summarizes the results of the Tukey's procedure to show which differences were significant.

Table 6. Summary of Tukey's results

Test	N	Subset for alpha = 0.01		
		1	2	3
SAT	131	11.23		
TOEFL	131		13.12	
MC	131			15.09
Sig.		1.00	1.00	1.00

The results reveal that the differences between the means for all the tests were significant when alpha equaled 0.01. Therefore, it was concluded that there was a significant difference between the students' scores on the different formats of grammar tests. In other words, the MC grammar test proved to be the

easiest, while the SAT error identification test was found to be the most difficult one.

Based on the responses to the first three questions of the study, it was also decided that the answer to the fourth question of the study was no. The justification will be provided below.

11. Discussion

The present study was conducted to explore the effect of test method facet on the validity of the grammar sub-tests of high stakes tests with a view toward test fairness.

The results of the factorial analysis revealed that the different formats of grammar tests produced different levels of construct validity, with the TOEFL format enjoying the highest and the SAT format the lowest level of construct validity. This indicated that the no-error option of the SAT error identification test might have played a role in its low level of validity, thus testifying to the presence of bias in this kind of testing.

Surprisingly, the correlational study of the students' scores on the three versions of the grammar test revealed that, although all of the tests measured the same construct and had been made on the basis of the same table of specifications, there existed only moderate correlations between them, with the highest between the TOEFL and MC tests (0.69), and the lowest between the SAT and the MC tests of grammar (0.417). This also indicated that the test format should have played a significant role in producing moderate correlations among the pairs of scores on the different tests.

The multiple comparison of the means of the three tests revealed that MC items are easier for the students to answer, while multitrak items with no-error options are much more difficult. However, the error identification test conforming to the TOEFL in format proved to be not only highly construct valid but midway between the other two tests in terms of difficulty. A test which is too difficult or too easy to answer does not provide us with reliable information about competency levels. Here, again it seemed that the test format had played a significant role in producing different outcomes and yielding different levels of difficulty.

In response to the fourth question of the study, it is to be noted that, as mentioned before, test validity and fairness are intertwined, and any changes in one produces changes in the other. The different levels of construct validity of the three formats, the moderate coefficient of correlations between the students' performances on them, and the grammar tests' different levels of difficulty indicated that it might not be fair to use specific test formats in standardized high stakes tests used for admission purposes.

12. Conclusion

In sum, the findings of this research indicated that different test formats produce different outcomes, confirming the results from numerous studies which have demonstrated that the methods used to measure language ability influence performance on language tests [13].

In many countries, including Iran, access to higher education is based on a supply and demand model, and social stratification corresponds to the number of years of formal education, which is directly associated with the prestige of the highest-reached educational institution. Considering the effects of these tests on individuals' lives, future, and careers, the large amount of money and energy spent in their construction and administration, and their possible undesirable backwash effects, this study highlighted the need for compensating for the effect of different methods and options which are used in testing a given construct, especially in high stakes tests such as admissions tests.

The statistical characteristics, particularly the high construct validity, of the error-identification test used in this study revealed that this format could be the most appropriate for testing the knowledge of grammar, especially when high stakes decisions are to be made. Since finding plausible distracters for standard multiple-choice items is not an easy job, especially concerning certain grammar points which yield few logical options, error-identification grammar tests can be a good replacement for multiple-choice grammar tests.

The results also point out the need for a closer examination of multitrak items with no-error options as used in the SAT. Since the no-error option of such items reduces the overall construct validity of the test to a considerable degree, the rationale behind employing this option should be re-examined. The debate over the value of the SAT as an effective tool in college admission practices continues since many people believe that it favors students from English-speaking homes [20]. While students may be fluent in English themselves, they may come from households where English is not the dominant language. Those students may have problems when it comes to identifying subtle differences in word connotations and problems in sentence structure, which is tested through error-identification items which contain a no-error option. Schaeffer [20] also claims that these tests do not necessarily reflect real differences among people, since item content differs from one test to another; even tests that claim to measure the same thing often produce very different results.

Since different standardized tests such as the TOEFL and SAT are widely used as admissions tests all over the world, and since every year many students stake their future on these tests, it is of utmost importance to make sure that their use is

justified in all countries and for all types of testees. Besides, it is necessary to reexamine their qualities in terms of test bias and test fairness in order to use the scores they produce as a criteria for making the best decisions in admissions contexts.

13. References

[1] Loschert, K., Raising the Ante for Students, Teachers, and Schools. Alexandria, VA: Association for Supervision and Curriculum Development. Retrieved November 4, 2002, from http:// www.ascd.org/frameinfobrief.html

[2] Coltrane, B., English Language Learners and High-Stakes Tests: An Overview of the Issues, *Eric Digest*, EDO-FL-02-07 • November, 2002.

[3] Davis, A., "Test Fairness: A Response", [Electronical version]*Language Testing, 27: 2*, 2010, 171-176 retrieved March 4, 2010.

[4] Society for Industrial and Organizational Psychology, Inc (SIOP) 2012.

[5] McNamara, T. & Roever, C., *Language Testing: The Social Dimension*, Oxford: Blackwell, 2006.

[6] Educational Testing Service, *Test Fairness and Validity*, 2012, Retrieved 12 August, 2012 from http://www.ets.org

[7] Jackson, C., *Understanding Psychological Testing*, British Psychology society, 1996.

[8] Messick, S., Validity, in R. L. Linn, *Educational Measurement*, American Council on Education, Oryx Press, New York, 1993.

[9] Gregory, R.J., *Psychological Testing: History, Principles and Applications*, Allyn and Bacon, Boston, 1992.

[10] Kane, M., "Validity and Fairness". *Language Testing, 27*(2), 2010, 177-182.

[11] Xi, Xiaoming, "How Do We Go about Investigating Test Fairness?" [Electronical version], *Language Testing, 27*: 2, 2010, 147-170, retrieved March 4, 2010.

[12] Kunnan, A. J., "Test Fairness and Toulmin's Argument Structure", [Electronical version] *Language Testing, 27*: 2, 2010, 183-189, retrieved March 4, 2010.

[13] Bachman, L. F., *Fundamental Considerations in Language Testing*, Oxford University Press, Oxford, 1990.

[14] Gergely. D., Investigating the Performance of Alternative Types of Grammar Items [Electronical version], *Language Testing, 24*: 1, 2007, 65-97.

[15] Willingham, W. W. & Cole, N., *Gender and Fair Assessment*. Mahwah, NJ: Lawrence Erlbaum, 1997.

[16] Shohamy, E., "Testing Methods, Testing Consequences: Are They Ethical? Are They Fair?" *Language Testing, 14*:3, 1997, 340-349.

[17] Garson, D., *Factor Analysis*, From Statnotes: Topics in Multivariate Analysis, Retrieved June, 12, http://www2.chass.ncsu.edu/garson/PA765/factor.htm, 2010, retrieved 21 December 2011.

[18] Sharma, S., *Applied Multivariate Techniques*, John Wiley & Sons Inc., USA, 1996.

[19] Zwick, W. R. & Velicer, W. F., "Comparison of Five Rules for Determining the Number of Components to Retain", *Psychological Bulletin, 99*: 3, 1986, 432-442.

[20] Schaeffer, R., *What's Wrong with Standardized Tests?* Fairtest, the national centre for fair and open testing [online version], retrieved 10 May from http:// www.fairness. Org

