The Investigation of the TMU English Proficiency test: Reliability Related Issues

Dr. G.R. Kiany¹
M. Haghighi²

Abstract

The majority of Iranian universities which offer doctoral programs, have started to accept only those candidates who have a good command of English proficiency, which is usually measured either by a general English test administered by the same university, or by tests like MCHE, TOLIMO, or MOHMET which are administered by the related ministries. The results of these tests are used to make very important decisions on the fate of candidates; hence, these tests should be carefully investigated in terms of reliability, validity, and the other characteristics of the test. The English Examination of Tarbiat Modarres University (TMU) is a TOEFL-like

¹. Assistant Professor of Tarbiyat Modares University; email: RezaKiany@yahoo.com
². MA in TESL, Tarbiyat Modares University
English language proficiency test used as a prerequisite for the Ph.D. Entrance Examination at Tarbiat Modarres University. This year, 6000 male and female postgraduates from different universities and different fields took part in this exam. This study attempts to investigate its reliability, test difficulty, and speededness. To achieve the purpose of the study, the data obtained from more than 2000 male and female Ph.D. applicants were analysed through different types of reliability estimates. The reliability coefficients of the total test and its subparts were estimated through K-R21, K-R20, split half, and cronbach α. Also the Standard Error of Measurement (SEM) was computed. Then the reliability coefficients of different types of reliability were compared for significant differences. In addition, the test difficulty as a general characteristic of the test, which may influence the size of reliability estimates, was also investigated. Related to this issue, the difficulty values of the test items are calculated. Based on the results, the relation between test difficulty and different types of reliability were discussed. This study also investigated speededness as a source of measurement error. Here, we determined if section 2 and 3 of the TOEFL-like TMU English Examination were truly speeded according to established criteria. Two exploratory techniques employing regression analyses were used in an attempt to account for the possibility
of random or patterned responses at the end of each section. One technique provided an estimate of the degree to which all examinees truly reached 75 percent on the sections, and the second determined if 80 percent of the examinees truly completed the last set of items. The results revealed that although the TMU English Examination was to some extent reliable, the reliability coefficients did not meet the levels which would be required for a high-stake test. Also it was found that based on norm-referencing criteria, the TMU exam did not have an acceptable level of difficulty, that is, it was too difficult for the examinees. Further, the results showed that the test was slightly speeded as far as the ETS criteria are considered. Results of this study may be of benefit to test developers in general and language test designers in particular. More specifically, policy makers and test developers of TMU, the Ministry of Science, Research and Technology, the Ministry of Health and Medical Education and other such institutes may get some benefit from this study.

**Key words:** Reliability – TMU English Examination – Speededness

**INTRODUCTION**

A language proficiency assessment attempts to measure a person's ability to understand and produce the language. It may be a global test
covering overall skills, i.e. listening, reading, writing, and speaking; or it may focus on only one or two of these skills. A proficiency test which is used to place students in a language program, to exit them from the program, or to admit them to other educational or employment opportunities should possess the basic characteristics of a test. If we are to interpret the score on a given test as an indicator of an individual’s ability, that score must be at least reliable and valid. These qualities are thus essential to the interpretation and use of measures of language abilities, and they are the primary qualities to be considered in developing and using tests. While validity is the most important quality of test use, reliability is a necessary condition for validity, in the sense that test scores that are not reliable cannot provide a basis for valid interpretation and use (Bachman, 1990). One of the general goals of test construction is to develop an instrument that achieves the necessary degree of reliability and validity for the intended use. Reliability is a function of variation. Variation can be systematic, which is related to the students’ ability, or unsystematic which is related to measurement errors. The aim in testing is to produce tests which measure systematic rather than unsystematic variation, and the higher the proportion of systematic variation in the test scores, the more reliable the test is (Bachman, 1990). Although it is usually impossible to achieve a perfectly reliable test, test constructors can make their tests as reliable as possible by reducing the causes of unsystematic variation to a minimum (Alderson, Clapham, and Wall, 1995). There are many factors other than the ability being measured that cause unsystematic variation in the test scores, and that constitute sources of measurement errors. Clearly, identifying and distinguishing the various factors that are
potential sources of measurement errors on a given test is thus crucial to the investigation of reliability. One of these sources of measurement errors is the effect of time limits on test performance (speededness). Speededness in testing is the effect of time limits on the test-taker’s scores. An exam is speeded to the extent that those taking it score lower than they would have if they had unlimited time. Speededness is often measured by calculating the proportion of examinees who do not reach a certain percentage of test items (Schnipke & Scrams, 1997).

In addition to sources of measurement errors, some of the general characteristics of the test such as test difficulty influence the size of the estimates of reliability. According to Ebel & Frisbie (1986), the difficulty of the test is a direct function of the difficulty of the test items. If we include only very difficult items, the test will be very difficult; likewise, if all the items are relatively easy, the test will also be easy. As Walsh & Betz (2001) suggest, overall level of difficulty should be appropriate for the group with whom the test is to be used and so that the test yields a relatively wide range of scores, preferably having a normal distribution.

Purpose of the Study

In this study, the English Examination of Tarbiat Modarres University (TMU) is under investigation. This test is similar to TOEFL in its structure, and it examines the candidates’ ability in structure, vocabulary, and reading comprehension. Admission in this TOEFL like examination is a prerequisite for those who are going to sit for Ph.D exam at Tarbiat Modarres University; that is, the applicants have to gain over 50, the cut-off score, to be allowed to
participate in Ph.D. Entrance Exam at Tarbiat Modarres University. Since the scores of English Examination of TMU are used as the bases for making important decisions that could affect the fate of subjects, the quality of the test should be investigated carefully. One of the test qualities is reliability coefficient. In other words, there was a need to estimate test score reliability so that we knew how much measurement errors might interfere with the interpretability of the scores. Because reliability is influenced by the group tested, the test content, and testing conditions, it was not possible to settle for a single method for estimating reliability (Ebel & Frisbie, 1986).

Fundamental to the proper evaluation of test quality is the identification of major sources of measurement errors. One of these sources of errors, affecting reliability, is the speededness. This study attempts to investigate speededness as a source of measurement error in the English Examination of TMU.

In addition to investigating reliability, the close examination of individual items that constitute the test is an integral part of test construction. Therefore, the analysis of the individual items is included in this study.

To achieve the purpose of the study, the following research questions were formulated:

1) To what extent are the TMU English Examination and its subparts reliable?

Q2) Based on the Norm – referencing criterion, do TMU English Examination items have acceptable level of difficulty?

Q3) Is TMU English Examination speeded based on the established criteria of speededness?

For these research questions, three hypotheses were proposed:
1) Neither the total nor the subparts of TMU English Examination are highly reliable.

2) Based on the Norm – Referencing criterion, TMU English Examination items do not have acceptable level of difficulty.

3) TMU English Examination is not speeded based on the established criteria of speededness

METHOD
Subjects
About 6000 Iranian male and female postgraduates or graduate students took part in TMU English Examination in Esfand, 1379 (February, 2001). They were from different universities in Iran with different academic background. Some of the applicants were going to sit for Ph.D exam of Tarbiat Modarres University. In order to be allowed to participate in Ph.D exam, they had to obtain a predetermined criterion, i.e 50% of total scores in TMU English Examination. The applicants’ higher education in Tarbiat Modarres University is dependent upon the result of TMU exam. Some others were Ph.D students who had obtained a score very close to cut-score 50 in previous exams of TMU. They had been admitted in Ph.D programs of Tarbiat Modarres University, provided that they could get a score over 50 before their comprehensive exam; otherwise they would be dismissed. From among these applicants, 1984 subjects who were students of humanities were selected for this study.

Instrumentation
A 100 item proficiency test, similar to TOEFL in format was administered. This TOEFL-like English examination of Tarbiat Modarres University included: structure and written expressions,
vocabulary, and reading comprehension. The sections consisted of 35, 35, and 30 items, respectively. It should be mentioned that this exam was administered by the university authorities, and the researcher did not have any role in developing or administering the test.

Procedure

A TOEFL-like proficiency test was administered to the candidates. The time allotted as the duration for the test was 100 minutes. After the administration procedure, the test was scored for the total and the subparts. The scoring scheme of this study was rights-only scoring; that is, the examinees were not penalized for wrong answers.

RESULTS AND DISCUSSION

Data Analysis for Hypothesis I:

The first research question was related to the reliability of TMU English Examination. In order to establish estimates of the reliability of the test, a number of approaches were taken. In this study, the internal consistency reliabilities were examined for the total test and its subparts through using Cronbach alpha, split-half, and KR-21 methods. Table 1 shows the results.

Table 1: Reliability Estimates of the Test

<table>
<thead>
<tr>
<th></th>
<th>KR-21</th>
<th>Alpha</th>
<th>Split-half</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>0.75</td>
<td>0.78</td>
<td>0.69</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>0.74</td>
<td>0.77</td>
<td>0.68</td>
</tr>
<tr>
<td>Reading</td>
<td>0.71</td>
<td>0.75</td>
<td>0.65</td>
</tr>
<tr>
<td>Total</td>
<td>0.86</td>
<td>0.88</td>
<td>0.78</td>
</tr>
</tbody>
</table>
According to Alderson et al. (1995), a well-constructed objective test of 100 multiple-choice items should have a reliability index of 0.95. Also Guerrero (2000), Scholfield (1995) and Davies (1990) believe that a test should have at least a consistency of 0.90. As it was mentioned earlier, admission in the TOEFL-like examination of Tarbiat Modares University (TMU) is a prerequisite for those who are going to sit for Ph.D. exam. Since the scores of this test are used to make important decisions on the fate of the subjects, it should have high reliability.

The reliability of TMU English Examination can be called into question on several grounds. Using 0.90 as recommended by Davies (1990) as an acceptable level of internal consistency on a high-stakes test, none of the sub-tests of TMU English Examination demonstrated acceptable reliability. The different reliability coefficients for section I ranged from 0.69 to 0.78, for section II from 0.68 to 0.74, and for section III from 0.65 to 0.75. It is clear that they are far below what would be required for a high-stakes test. Considering that three different types of reliability show low reliability coefficients for three sections, we can claim that the sections do not provide acceptable reliability for the intended use of the test.

The reliability of the total test is nearer to the 0.90 criterion. Alderson et al. (1995) recommends 0.95 as high reliability. Hence, we can claim that though the test has been found to be to some extent reliable, the reliability coefficients do not meet the levels which would be required for a high-stakes test.

Data Analysis for Hypothesis II:

The null hypothesis for the second research question was that based
on the norm-referencing criterion, TMU English Examination does not have an acceptable level of difficulty, as far as items are concerned.

Accordingly, in this study the difficulty and discrimination values of the test items were calculated first through classical item analysis and then through IRT methods. It should be mentioned that the term misfit is used for statistical tests to fit to a measurement model; however, in this study we used the term misfit to refer to unacceptable items from the point of view of difficulty and discrimination, based on norm-referencing criterion.

**Phase 1: Classical Item Analysis**

In this phase of item analysis, the performance of two misfit statistics was examined: P-value (item difficulty), and biserial correlation (item discrimination).

**P-value:** for each item, the item difficulty was calculated. Items having P-values of less than 0.37 or greater than 0.63 were considered to be misfitting for this analysis (Osterlind, 1998; Reynolds, 1994; Tuckman, 1978). It should be mentioned that this range of difficulty is based on norm-referencing criterion and is not necessarily absolute. The results are represented in Table 2.

**Table 2: Statistics of item difficulty**

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>0.37-63</th>
<th>0.07</th>
<th>0.07-17</th>
<th>0.17-27</th>
<th>0.27-37</th>
<th>0.63-73</th>
<th>0.73-83</th>
<th>0.83-93</th>
<th>0.93-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>35</td>
<td>14</td>
<td>0</td>
<td>4</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>35</td>
<td>13</td>
<td>0</td>
<td>3</td>
<td>7</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Reading</td>
<td>30</td>
<td>12</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>39</td>
<td>0</td>
<td>9</td>
<td>21</td>
<td>26</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Acceptable | Difficult | Easy
Considering the range of 0.37 to 0.63, as criterion for acceptable level of difficulty for norm-referenced tests, only 39% of the items had acceptable level of difficulty in this test. Also, it was found that 56% of times had item difficulty below 0.37 and only 5% of items were above 0.63; that is, 56% of the items were difficult and 5% were easy.

Thus, considering these results and taking the range of 0.37 to 0.63 as the norm-referencing criterion for acceptable level of difficulty (Farhady et al. 1995, Reynolds et al. 1994, Osterlind 1998), we can claim that TMU English Examination was a difficult test for the population under investigation.

Biserial correlation: for an item discrimination index a biserial correlation was calculated for each item. For this analysis, items with biserial correlations of less than 0.30 were considered to be misfitting (Fulcher, 1997; Wall et al, 1994; Heward, 1989). The results showed that 37% of the items were below 0.30 criterion. That is, they were either nonfunctioning or malfunctioning. Five items were malfunctioning because they were items with negative item discrimination. The rest of the aforementioned 37% were nonfunctioning. The results also showed that only 13% of the items had item discrimination above 0.5. The rest of the items (50%) had item discrimination from 0.30 to 0.50. So it can be concluded that most of the items were not suitable functioning items; that is, these items did not have a good discrimination power.

Phase II: Item Analysis Through IRT Methods

IRT offers many important advantages over Classical True Score Theory (CTS) in terms of item parameter estimates that are sample independent, which are very useful properties for practical
applications such as tailored testing, test equating, and identifying item bias, and in terms of its increased power in estimating the precision of measurement.

Accordingly, in this study the test items were analyzed through IRT methods, too. The program used in this analysis was Rasch Model. The results are shown in table 3.

Table 3: Statistics of b parameter (item difficulty) in IRT analysis

<table>
<thead>
<tr>
<th></th>
<th>+1 to -1 (acceptable)</th>
<th>Above +1 (difficult)</th>
<th>Below -1 (easy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>18</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>13</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>Reading</td>
<td>10</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>41</td>
<td>59</td>
<td>0</td>
</tr>
</tbody>
</table>

To construct a test that screens out the examinees properly, we should select items that have large a-parameter (item discrimination) and b-parameter (item difficulty) near 0.0 (Bachman, 1990; Osterlind, 1998; and Anastazi, 1982). In this study, we considered one standard deviation above and below the mean ability (\( \mu = 0.0 \)) as the criterion for acceptable item difficulty in IRT analyzes. Considering the range of item difficulty from -1 to +1 (-1 \( \leq b \leq +1 \)) for acceptable level of difficulty, we found that 41 percent of the items had acceptable level of difficulty, and 59 percent of the items were not acceptable. It is worth mentioning here that the values of all 59 percent of items with unacceptable item difficulty were above +1; that is all of them were difficult items. Analyses through IRT methods showed that there were no easy items in this test. Therefore, it can be concluded that the test
was difficult for the population (similar conclusion when CTS was used).

As for the discrimination value of the items or a-parameter, experts recommend that the value of a-parameter below 0.5 is unacceptable (University of Illinois IRT Laboratory, 2001). As it was mentioned earlier, the larger the a-parameter the better discriminant the item. The results of IRT analyzes showed that 15 percent of the items had unacceptable item discrimination. Comparing the results of classical and IRT analyses, we can claim that TMU English Examination was a difficult test for the population. The fact that about 60% of the items were difficult support this claim.

The results also showed that 27 percent of the items (27 items) were functioning items; that is, they had acceptable item difficulty and item discrimination through both classical and IRT analyses. Further, it was found that 15% of the items were non-functioning ones; that is they did not have acceptable level of difficulty and discrimination in both classical and IRT analyses. The rest 58 percent of the items had at least one misfit statistic in classical or IRT analyses. In pretesting, the non-functioning items are eliminated from the test. However, there was no pretesting for TMU English Exam. Since this test has a great effect on the examinees' future, the scores of such a test should be interpreted cautiously. At least, in such cases, we can omit the non-functioning items in scoring and interpret the results of the new scores. This process helps decision-makers to be more accurate in their decisions.
Data Analysis for Hypothesis III:

The null hypothesis for the third research question was that TMU English Examination is not speeded based on the established criteria of speededness.

A test is considered to be essentially unspeeded by Educational testing Service (ETS) criteria if virtually all examinees reach 75 percent of the items and at least 80 percent of examinees reach the last item (Hatch & Swinford, 1981; Swinford, 1971).

Accordingly, the ETS criteria should be analyzed in order to be confident in the results. For analyzing the first criterion (Did virtually all examinees complete the first 75 percent of the items?), both simple and multiple regression analyzes were employed. First, simple linear regression was performed using the scores on the last 25 percent of the items as the criterion. In the second stage, a multiple regression was applied using scores on approximately the 15 percent of item immediately preceding the last 25 percent of the items. In both cases, the predictor variable (x) was the scores on the set of items representing approximately the first 60 percent of the items. The raw score for the predictor is very unlikely to be affected by speededness since the first 60 percent of the items are not located near the end of the test.

After predicted scores were obtained on the last set of items (Ŷ last) and immediately preceding set of items (Ŷ mid), residuals were computed for the regressions and then standardized. Table 4 presents the numbers and percentages of examinees with standardized residuals below z = -1.645 and above z = +1.645 (critical value of z = 1.645) for the middle and last set of items (Secolsky, 1989).
Table 4: Numbers and percentages of Examinees with 
standardized residuals Below $z = -1.645$ and above $z = +1.645$ for 
middle and last item sets

<table>
<thead>
<tr>
<th></th>
<th>N below</th>
<th>Percentage</th>
<th>n above</th>
<th>Percentage</th>
<th>n total</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Middle item set</td>
<td>91</td>
<td>4.6</td>
<td>107</td>
<td>5.3</td>
<td>198</td>
<td>9.9</td>
</tr>
<tr>
<td>Last item set</td>
<td>87</td>
<td>4.3</td>
<td>117</td>
<td>5.9</td>
<td>204</td>
<td>10.2</td>
</tr>
</tbody>
</table>

As can be seen from these data, the percentage of examinees with standardized residuals below $z = -1.645$ on both middle and last item sets is 8.4 percent. This indicates that relatively small percentage of the examinees scored lower than predicted on both the middle and last item sets. The data also shows that 11.2 percent of examinees had standardized residuals above $z = +1.645$. That is, 11.2 percent of apparently lower ability examinees scored higher than predicted on both the middle and last item sets. Along the line of secolsky’s work (1989), we used the cut-off of 10-15 percent of the examinees with standardized residuals above $z = +1.645$ and below $z = -1.645$ to claim that a test is truly speeded. From the results, it appeared that 10.2 percent of the examinees had standardized residuals above $z = +1.645$ and below $z = -1.645$ on the last set of items. Also, it was found that 20.1 percent of the examinees had standardized residuals above $z = +1.645$ and below $z = -1.645$ on both the last and middle item sets. According to secolsky’s work (1989), if an examinee scored lower or higher than predicted on both the last and middle item sets, it was possible that the examinee may have either 'not reached' the last 25 percent of items or guessed randomly at these items. In this study, it appeared possible that not all examinees truly reached the 75 percent point on the test. Thus according to the first speededness criterion, it
can be claimed that the test has been slightly speeded for the population.

For analyzing the second speededness criterion (Did 80 Percent of examinees reach the last set of items?), we used some traditional measures (Secolsky, 1989). Of the 1984 examinees included in this study, 180 or 9.07 percent of the examinees responded with the same response to al least the last 10 items. If one added to this figure the proportion of examinees that did not complete the test (260 or 13.1 percent of the examinees), the percentage would rise to 22.17 percent. Since this figures still do not include the proportion of examinees that may have responded in a random responding fashion, the proportions truly reaching the last set of items may not have reached 80 percent.

From the results, it appears that fewer than 80 percent of the examinees truly completed the test. Hence this test appears to be speeded according to this criterion. The results of the analysis of the third hypothesis suggest that TMU English Examination is slightly speeded according to ETS speededness criteria.

CONCLUSION

While the emphasis of this study has been to examine the evidence of reliability in TMU English Examination, it is also necessary to consider the usefulness of these results. It is commonly viewed that a test which has poor reliability, can not be said to be valid, and therefore should not be used to make important decisions. High-stakes test development and use is a complex psychometric undertaking that challenges test users at every turn. As TMU English Examination is a high-stakes test whose results affect the educational career of the applicants, it should gain high quality needed for a good test. It is
important that institutions such as universities that hold such tests know the estimate of error in their testing procedure to improve screening programs and to be cautious in interpreting the results. As the findings of this study show and with regard to the results of the other research done on TMU English Examination by English department, it becomes obvious that the test has certain shortcomings, with regard to the population for whom the test is used. So it would be recommended that the university design a long-lasting program including pretesting, trialing, and item banking to alleviate the probable shortcomings of the test. As in TMU English Exam it might not be possible to do these processes, the decision-makers should be cautious about the interpretation of the test results. It should be mentioned that the test developers just design the test, and they do not have any role in determining the cut-off score or in decision making process. Thus It can be proposed that the policy makers define the passing criterion carefully and make decisions based on that criterion. It is worth mentioning that the purpose of this study was not to undermine the efforts or commitments of those individuals involved in the development of TMU English Examination, but to learn from their efforts and recommend to the policy makers as well as the test developers to be careful of the possible shortcomings of the test and more importantly of the probable faults in the decision making process.

References
Publication Co.


Multilingual Matters LTD.