An Item Analysis of an English Achievement Test
Taken by EFL College Students in Taiwan

Yu-mien Shih

Abstract

This study attempts to examine a mid-term English achievement test held in a university in Taiwan. Test items are analyzed, based on assessment devices such as Item Facility, Item Discrimination and Distractor Analysis, to evaluate if the test was appropriately administered in terms of level of difficulty, power of discrimination and function of distractors. Research issues include (1) whether items in four categories are easy or difficult for test takers; (2) whether items can discriminate the performance of top scorers and low scorers; and (3) whether distractors function well or badly. The result of a questionnaire indicates that about 60 percent of test takers think vocabulary is the most difficult category for them, which corresponds with the result of the study. Research findings have shown that vocabulary tests are the most difficult part of the whole test and exhibit the most powerful discrimination. A higher percentage of items in listening, reading, and grammar tests show little or no discrimination power. More nonsense distractors appear in listening tests. The outcome of the analysis suggests that the test needs to be revised. Suggestions are made to improve the quality of the achievement test in order to effectively measure students’ performance of learning.

Key Terms: Item Facility  Item Discrimination  Distractor Analysis  Achievement Test
INTRODUCTION

Achievement tests are directly related to language courses. The purpose of achievement tests is to establish how successful students or the courses themselves have been in achieving objectives (Hughes, A. 1989). Achievement tests assess what students have achieved from their courses within a period of time and provide some information with regard to their current level of progress or whether they are ready for the subsequent stages of learning.

Achievement tests are administered most frequently in language programs than any other kinds of tests and influence greatly students’ motivation and attitude of learning. Thus, the test should be fair enough to measure students’ learning, degree of progress and language ability. Achievement tests are intended to inform students of their actual performance in a period of time. The importance of the evaluation of achievement tests can’t be overemphasized.

This study attempts to evaluate a unified mid-term achievement test administered jointly by Chinese and foreign English teachers for college freshmen English program in a university in Taiwan.

Purpose

This study aims to analyze items on an achievement test based on the assessment devices such as Item Facility, Item Discrimination and Distractor Analysis. Item analysis was conducted to examine the degree to which each item is effective in terms of the level of difficulty, the power of discrimination and the function of distractors.

Item facility indices show how easy or difficult each item is. Item discrimination indices help decide if items have any discriminating power between top and lower scorers. A distractor analysis determines if distractors are functioning as they should do.

Conducting item analysis is to identify the appropriateness of items with regard to the level of difficulty, the power of discrimination and the function of distractors. Besides, item analysis helps examine particular objectives of learning which students haven’t achieved properly and seek to review and re-teach.

Significance of the Study

Teachers, students as well as test developers all benefit from this study.

First, teachers can understand how much students have learned from the materials assessed and check the discrepancies between their expectations and students’ actual performance. In addition to students’ performance, the study may also reveal how well the teachers have put across the materials and how well
Teachers receive feedback from the test as a basis of modifying test items and adjusting their course objectives as well as teaching content. Secondly, students may understand their strengths and weaknesses with respect to their language learning. The item analysis provides students information as to how well they perform in some areas and how badly they do in others, prompting them to make more efforts in the parts they haven’t mastered. Thirdly, test developers can benefit from this study by understanding the weak points of the test as a whole and seek to make revisions in the future.

Research Questions

Three research questions are proposed as follows:

(1) How many easy and difficult items are there in four categories based on the item facility indices?
(2) Can items discriminate significantly between top scorers and lower scorers in accordance with item discrimination indices?
(3) How many problematic items and inappropriate distractors are there in this test?

This is a preliminary analysis of items in an achievement test. The paper didn’t mention reliability, validity and usability. However, they are important components in a test. If the same test is given to the same student or on two different occasions, the reliable test should yield similar results (Brown, 2004). Validity means the extent to which a test measures what it is supposed to measure. A valid reading test should actually measure reading ability. Usability refers to the extent to which a test is practical in actual implement. Test takers and teachers should bear in mind that these notions are indispensable in evaluating a test.

LITERATURE REVIEW

A test is a method of measuring a person's ability, knowledge or performance in a given domain (Brown, 2004). Language tests have different kinds of purposes. Bailey (1998) holds that there are 8 kinds of language assessment: aptitude tests, language dominance tests, proficiency tests, admission tests, placement tests, diagnostic tests, progress tests and achievement tests.

Aptitude tests are designed to assess a learner’s potential capacity for learning languages. Aptitude tests are intended to test students’ ability to learn any language. Language dominance tests involve assessing potentially bilingual students in both languages they have been
exposed to, in order to determine which is the dominant language for purposes of instruction. As for proficiency tests, they aim to assess the overall language use in a variety of situations involving four skills and all levels of language. GEPT (General English Proficiency Tests) in Taiwan are proficiency tests. Admission tests are used to provide information about whether a candidate is likely to succeed in a particular program. TOEFL (Test of English as a Foreign Language) is an admission test for universities in America and other countries. Placement tests are intended to evaluate students’ skills relative to the levels of a particular program they are about to enter. Placement tests are commonly used to assign students to particular levels of a program. Diagnostic tests are used to identify students’ particular strengths and weaknesses. Diagnostic tests are highly related to the syllabuses of the specific courses. Teachers can adjust the instruction to be more appropriate for the students in the class on the basis of diagnostic tests. Progress tests are used as part of an ongoing assessment procedure during the course of instruction. Thus, progress tests are also closely related to the course content. Achievement tests, the same as diagnostic tests and progress tests, are based on the objectives of the course. An achievement test assesses what has been “achieved” or learned from what was taught in a particular course or a series of courses (Cohen, 1980). The above tests may overlap. For example, achievement tests may be used as a diagnostic purpose.

**Achievement tests are criterion-referenced**

Achievement tests and diagnostic tests are criterion-referenced tests, different from norm-referenced tests, in which students’ scores are compared with the achievement of other students. In criterion-referenced tests, students’ performance is measured with respect to the degree of their learning or mastery of the prespecified content domain. The test is administered in accordance with teaching objectives, so it is highly related with the teaching of the course. Achievement tests are assessed to provide students with information as to language skills or elements they need to improve. Teachers need to give appropriate feedback or what Oller (1979, p.52) called “instructional value” to students after the test.

**METHODS**

**Participants**

Subjects include 32 (24 male and 8 female) native Chinese freshmen in the beginning level of English class in a university. They were sorted into the beginning level of English class by a freshmen English placement
test at the start of the semester. They are non-English major students from the departments of Applied Mathematics, Material Engineering and Environmental Engineering. Generally speaking, they don't use English very well. They are required to spend two hours in class and several hours in a self-study center practicing English, using English language software. They took this class as a required course. The goal of the English course is to develop their communicative ability in four skills. In class, they are usually asked to do some tasks to enhance their ability of listening, speaking, reading and writing. For example, they listen to CD, speak to introduce themselves, or read texts and so on. In a word, the purpose of the English class is to facilitate students’ communicative ability of English.

Material

The aim of the English course is to enhance students’ English ability of listening, speaking, reading, and writing. Students are required to be able to use basic greetings, introducing personal information, talking about favorites, describing countries or cities etc. They need to master grammatical structures like subjective and objective pronouns, possessive adjective, yes/no questions and answers, past tense, present continuous tense and present perfect tense, and so on. Students are encouraged to read four simplified readers in a semester. The content of a simplified reader is included in the test. The textbook is a new edition “World Link,” published by Thomson Heinle Company in 2005. The test was constructed according to the objectives of the course mentioned above. There are four categories: listening, reading, grammar and vocabulary. The item formats, item types, and item numbers in each category are displayed in Section 3.3.1.1 below.

Data Collection and Analysis

Data was sampled from 32 examinees’ response to items on the mid-term achievement test for the beginning level of college freshmen English class.

This test is a paper-and-pencil test consisting of 75 test items including 4 parts: listening, grammar, reading and vocabulary. The types of task include short answers, dialogue, listening cloze, question and answer, scanning, editing, ordering and matching. The types of test format involve multiple-choice, true or false, fill-ins, and sentence completion. Multiple-choice and true/false-choice formats constitute the largest part of all kinds of formats. Concerning scoring, each category takes 25 points.

The categories, task types, format types, total number of items and proportion of score in each category of the test are displayed in Table 1.
A questionnaire was designed to elicit the opinion of test takers on the test. Twenty-three students (9 female 14 male) are in response to the questions raised in the questionnaire and hand them back. Results show that about 60 percent of test takers think vocabulary is the most difficult category for them, which corresponds with what the score indicates. About 52% of students aren't satisfied with the score they obtain. Nearly 57% of test takers think they didn't perform well on the test because of their poor ability of English.

Descriptive statistic instruments used such as frequency distribution, mean, median, mode, measure of dispersion, range and standard deviation are described as follows.

Scores are easiest to work with if they are organized in a frequency distribution as shown in Table 2. The scores distribute from the top, 75, to the bottom, 22.

Table 1: The Categories, Tasks, Item Numbers and Formats of the Test

<table>
<thead>
<tr>
<th>Category</th>
<th>Score</th>
<th>Task</th>
<th>Total item number</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening</td>
<td>25</td>
<td>Short answer</td>
<td>10</td>
<td>Multiple-choice</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dialogue</td>
<td>5</td>
<td>Multiple-choice</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cloze</td>
<td>10</td>
<td>Fill-in</td>
</tr>
<tr>
<td>Grammar</td>
<td>25</td>
<td>Question-Answer</td>
<td>10</td>
<td>Multiple-choice</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Question-Answer</td>
<td>5</td>
<td>Fill-in</td>
</tr>
<tr>
<td>Reading</td>
<td>25</td>
<td>Scanning</td>
<td>10</td>
<td>True/false-choice</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ordering</td>
<td>5</td>
<td>Sentence-completion</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Editing</td>
<td>5</td>
<td>Sentence-completion</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>25</td>
<td>Matching</td>
<td>10</td>
<td>Fill-in</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Matching</td>
<td>5</td>
<td>Multiple-choice</td>
</tr>
</tbody>
</table>

Total items: 75 Total score: 100

Table 2: Frequency Distribution of the Scores for 32 Chinese Students on an English Midterm Test

<table>
<thead>
<tr>
<th>Mark</th>
<th>Frequency</th>
<th>Mark</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>1</td>
<td>53</td>
<td>1</td>
</tr>
<tr>
<td>73</td>
<td>1</td>
<td>51</td>
<td>1</td>
</tr>
<tr>
<td>69</td>
<td>1</td>
<td>49</td>
<td>2</td>
</tr>
<tr>
<td>68</td>
<td>1</td>
<td>48</td>
<td>2</td>
</tr>
<tr>
<td>66</td>
<td>2</td>
<td>44</td>
<td>1</td>
</tr>
<tr>
<td>65</td>
<td>1</td>
<td>42</td>
<td>1</td>
</tr>
<tr>
<td>63</td>
<td>2</td>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td>61</td>
<td>1</td>
<td>38</td>
<td>1</td>
</tr>
<tr>
<td>60</td>
<td>1</td>
<td>35</td>
<td>1</td>
</tr>
<tr>
<td>58</td>
<td>1</td>
<td>34</td>
<td>2</td>
</tr>
<tr>
<td>56</td>
<td>2</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>55</td>
<td>1</td>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>54</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total 1667 32 (students number)
Mean

There are 32 students participating in this test. The total score of all students is 1667. Mean 52.1 is gained from using total scores (1667) divided by total student number (N=32).

Median

A second measure of central tendency is the term median. It's the numerical point in the distribution at which half of the obtained scores lie above and half below. The median is 53.5 (54+53/2).

Mode

The third measure of central tendency is the term mode. Mode is the most frequently occurring score which is located as the peak of the curve. Sometimes, distributions have more than one mode. There are 7 modes in the distribution.

Measures of Dispersion

There are two ways of measuring dispersion. Range can measure the distance between the highest and lowest scores and standard deviation the spread of scores.

Range

One simple way of measuring the spread of marks is based on the difference between the highest and lowest scorers. The scores are arranged in the order from highest mark to lowest mark. The highest score on a 75-item test is 75 and the lowest 22. So the range is 53 (75-22).

Standard Deviation

The standard deviation (s.d.) is another way of showing the spread of scores. It measures the degree to which the group of scorers deviate from the mean; that is, it shows how all the scores are spread out and gives a fuller description of test scorers than the range. A standard deviation of 13.15 in the test shows a larger spread of scores.

There are three assessment devices used in this study: Item Facility, Item Discrimination and Distractor Analysis. The definition of them is described as follows:

1. Item facility is an index which represents the proportion of students who got the item right and shows how easy or difficult an item was for test-takers.

2. Item discrimination indices distinguish between how top scorers and low scorers perform on each item.

3. Distractor analysis is a procedure related to the multiple-choice format. It is conducted to see how distractors are functioning.

Item analysis provides important information with regard to the quality of a written test administered to examinees. Teachers can revise, remove or retain test items based on it.
**Item Facility**

Item facility (abbreviated I.F.) refers to an index of how easy an item is for the test takers. I.F. is a number typically printed as a decimal, ranging from 0.0 to 1.0. It represents the proportion of people who got the item right from all test takers (Bailey, 1998).

I.F. is calculated by the number of scorers who got the answer right divided by total number of test takers. This means for item 5 in Listening Task 7 (total correct answers) is divided by 32 (total number of test takers).

If the facility value is higher than 0.85, the item is too easy. If the facility value is lower than 0.3, then the item is probably too difficult and should be redrafted. The criteria of I.F. (Bailey, 1998) are stated as follows:

Criteria of I.F. Index

1. 0.85 and up – Items are too easy and need to make revisions.
2. Below 0.3 – Items are too difficult and need to make revisions.
3. 0.3 to 0.39 – Reasonably acceptable
4. 0.4 to 0.85 – Very good items

**Item Discrimination**

Item discrimination (I.D.) refers to a more detailed analysis of the test items than item facility. It shows how top scorers and lower scorers perform on each item. Item discrimination investigates whether an item with a low I.F. is actually difficult, or if there are any other factors, which cause the low percentage of correct responses for that item.

I.D. is calculated by \[
\frac{\text{Top (correct)} - \text{Low (correct)}}{\frac{1}{2} N}
\]

I.D. indices range from +1 to -1, with positive 1 showing a perfect discrimination between top and low scorers, with minus 1 showing wrong discrimination, and I.D. indices zero show no discrimination. The lowest acceptable values are usually at 0.25 or 0.35 (Oller, 1979). The criteria of I.D. are listed as follows:

Criteria of I.D. Index.

1. 0.4 to 1 – Good discrimination
2. 0.2 to 0.35 – Acceptable
3. 0 to 0.1 or below zero – No discrimination

Four categories include listening, vocabulary, grammar and reading. The indices of item facility and item discrimination of four categories are displayed as in Table 3:
Table 3: Item Facility and Item Discrimination of Four Categories

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1(listen)</td>
<td>1.0</td>
<td>0.0*</td>
<td>26(Grammar)</td>
<td>0.13*</td>
<td>-0.2*</td>
<td>51</td>
<td>0.78</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>1.0</td>
<td>0.0*</td>
<td>27</td>
<td>0.75</td>
<td>0.3</td>
<td>52</td>
<td>0.88</td>
<td>0.1</td>
</tr>
<tr>
<td>3</td>
<td>0.75</td>
<td>0.4</td>
<td>28</td>
<td>0.84</td>
<td>0.0*</td>
<td>53</td>
<td>1.0</td>
<td>0.0*</td>
</tr>
<tr>
<td>4</td>
<td>0.78</td>
<td>0.2</td>
<td>29</td>
<td>0.19</td>
<td>0.1</td>
<td>54</td>
<td>0.56</td>
<td>0.3</td>
</tr>
<tr>
<td>5</td>
<td>0.21*</td>
<td>0.0*</td>
<td>30</td>
<td>0.81</td>
<td>0.2</td>
<td>55</td>
<td>0.69</td>
<td>0.5</td>
</tr>
<tr>
<td>6</td>
<td>0.84</td>
<td>0.0*</td>
<td>31</td>
<td>0.5</td>
<td>-0.2*</td>
<td>56</td>
<td>0.84</td>
<td>0.3</td>
</tr>
<tr>
<td>7</td>
<td>0.53</td>
<td>0.5</td>
<td>32</td>
<td>0.59</td>
<td>0.5</td>
<td>57</td>
<td>1.0</td>
<td>0.0*</td>
</tr>
<tr>
<td>8</td>
<td>0.97</td>
<td>0.0*</td>
<td>33</td>
<td>0.63</td>
<td>0.1</td>
<td>58</td>
<td>0.59</td>
<td>0.5</td>
</tr>
<tr>
<td>9</td>
<td>0.47</td>
<td>0.33</td>
<td>34</td>
<td>0.63</td>
<td>0.4</td>
<td>59</td>
<td>0.41</td>
<td>0.3</td>
</tr>
<tr>
<td>10</td>
<td>0.88</td>
<td>0.0*</td>
<td>35</td>
<td>0.75</td>
<td>0.5</td>
<td>60</td>
<td>0.69</td>
<td>0.4</td>
</tr>
<tr>
<td>11</td>
<td>0.94</td>
<td>0.1</td>
<td>36</td>
<td>0.88</td>
<td>0.0*</td>
<td>61(Vocabulary)</td>
<td>0.59</td>
<td>0.8</td>
</tr>
<tr>
<td>12</td>
<td>0.81</td>
<td>0.2</td>
<td>37</td>
<td>0.19*</td>
<td>0.5</td>
<td>62</td>
<td>0.56</td>
<td>0.6</td>
</tr>
<tr>
<td>13</td>
<td>0.47</td>
<td>0.4</td>
<td>38</td>
<td>0.78</td>
<td>0.78</td>
<td>63</td>
<td>0.59</td>
<td>0.6</td>
</tr>
<tr>
<td>14</td>
<td>0.81</td>
<td>0.2</td>
<td>39</td>
<td>0.56</td>
<td>0.55</td>
<td>64</td>
<td>0.03*</td>
<td>0.1</td>
</tr>
<tr>
<td>15</td>
<td>0.75</td>
<td>0.1</td>
<td>40</td>
<td>0.78</td>
<td>0.4</td>
<td>65</td>
<td>0.31</td>
<td>0.3</td>
</tr>
<tr>
<td>16</td>
<td>1.0</td>
<td>0.1</td>
<td>41(Reading)</td>
<td>0.69</td>
<td>0.5</td>
<td>66</td>
<td>0.22*</td>
<td>0.5</td>
</tr>
<tr>
<td>17</td>
<td>1.0</td>
<td>0.0*</td>
<td>42</td>
<td>0.75</td>
<td>0.5</td>
<td>67</td>
<td>0.16*</td>
<td>0.4</td>
</tr>
<tr>
<td>18</td>
<td>0.22*</td>
<td>0.4</td>
<td>43</td>
<td>0.97</td>
<td>0.1</td>
<td>68</td>
<td>0.47</td>
<td>0.7</td>
</tr>
<tr>
<td>19</td>
<td>0.41</td>
<td>0.4</td>
<td>44</td>
<td>0.81</td>
<td>0.4</td>
<td>69</td>
<td>0.47</td>
<td>0.8</td>
</tr>
<tr>
<td>20</td>
<td>0.59</td>
<td>0.6</td>
<td>45</td>
<td>0.03*</td>
<td>-0.1*</td>
<td>70</td>
<td>0.13*</td>
<td>0.3</td>
</tr>
<tr>
<td>21</td>
<td>0.94</td>
<td>0.0*</td>
<td>46</td>
<td>0.91</td>
<td>0.0*</td>
<td>71</td>
<td>0.63</td>
<td>0.7</td>
</tr>
<tr>
<td>22</td>
<td>0.03*</td>
<td>0.1</td>
<td>47</td>
<td>0.91</td>
<td>0.2</td>
<td>72</td>
<td>0.25*</td>
<td>0.1</td>
</tr>
<tr>
<td>23</td>
<td>0.53</td>
<td>0.3</td>
<td>48</td>
<td>0.69</td>
<td>0.0*</td>
<td>73</td>
<td>0.31</td>
<td>0.3</td>
</tr>
<tr>
<td>24</td>
<td>0.28*</td>
<td>0.4</td>
<td>49</td>
<td>0.78</td>
<td>0.5</td>
<td>74</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>25</td>
<td>0.25*</td>
<td>0.5</td>
<td>50</td>
<td>0.78</td>
<td>0.6</td>
<td>75</td>
<td>0.19*</td>
<td>-0.4*</td>
</tr>
</tbody>
</table>

I.F. = Item Facility ---- Underlined number: too easy/ Numbers with asterisk mark: too difficult  
I.D. = Item Discrimination ---- Items with asterisk mark: no discrimination

Table 3 shows that the I.F. value of 15 items is higher than 0.85, which designates they are too easy. A total of 45 items within the standard range (0.3-0.85) indicates that they are acceptable ones. The I.F. value of 15 items is below 0.3 which reveals they're too difficult. As for item discrimination, there are 24 items showing no discrimination. 3 items show wrong discrimination and the other 48 items have good discrimination power.
There are three levels of items: easy, moderate and difficult. The distribution of items in four categories in terms of level of difficulty is indicated in Table 4:

**Table 4: The Distribution of Items in Terms of Level of Difficulty in Categories**

<table>
<thead>
<tr>
<th>(Total Items)</th>
<th>Easy</th>
<th>Moderate</th>
<th>Difficult</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening (25)</td>
<td>8</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>Grammar (15)</td>
<td>1</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>Reading (20)</td>
<td>6</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>Vocabulary (15)</td>
<td>0</td>
<td>9</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 4 shows that more difficult items (40%) but no easy items appear in vocabulary tests. Overall, items are easier in reading tests. Most items in reading tests are moderate.

There are three levels of discrimination: no, acceptable and good. The distribution of items in terms of power of discrimination is shown in Table 5:

**Table 5: The Distribution of the Discriminating Power of Items in Categories**

<table>
<thead>
<tr>
<th>Discrimination</th>
<th>No (0.0-0.1)</th>
<th>Acceptable (0.2-0.35)</th>
<th>Good (0.4 or over)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening</td>
<td>12</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Grammar</td>
<td>6</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Reading</td>
<td>8</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>3</td>
<td>4</td>
<td>8</td>
</tr>
</tbody>
</table>

Total item: 75

About 48 per cent of items can't discriminate well in listening tests. Only 20 per cent of items show no discrimination in vocabulary tests. Most items (80%) of vocabulary tests show powerful discrimination in this test.

**Distractor Analysis**

The purpose of distractor analysis is to investigate which distractors are functioning well as they should do and which are not. A distractor analysis is related only to the multiple-choice formats. Therefore, no analysis is done in reading tests.
The response frequency distribution in Table 6 gives a more detailed look at how the distractors are functioning. Three categories are included in this analysis.

<table>
<thead>
<tr>
<th>Item</th>
<th>Item scorers</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Item scorers</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Top Low</td>
<td>9*</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2*</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Top Low</td>
<td>9*</td>
<td>0</td>
<td>0</td>
<td>3*</td>
<td>2*</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Top Low</td>
<td>0</td>
<td>9*</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>8*</td>
<td>4*</td>
</tr>
<tr>
<td></td>
<td>Top Low</td>
<td>0</td>
<td>9*</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>7*</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Top Low</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>9*</td>
<td>5*</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Top Low</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>7*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Top Low</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Top Low</td>
<td>0</td>
<td>6</td>
<td>3*</td>
<td>3*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Top Low</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Top Low</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Top Low</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Top Low</td>
<td>7*</td>
<td>2*</td>
<td>1</td>
<td>2*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Top Low</td>
<td>2*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Top Low</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Top Low</td>
<td>0</td>
<td>2*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Top Low</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Top Low</td>
<td>0</td>
<td>9*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Top Low</td>
<td>1</td>
<td>5*</td>
<td>3*</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Top Low</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Top Low</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>8*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Top Low</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Top Low</td>
<td>2</td>
<td>7*</td>
<td>0</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Top Low</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Top Low</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>8*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Top Low</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Top Low</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>7*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Top Low</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Items with asterisk mark are the correct answers.
A distractor analysis uses response frequency distribution to differentiate those good distractors from nonsense ones. Table 6 shows not only options chosen by top scorers and low scorers for each item, but also why some items don't discriminate at all. Response frequency distribution combines information from distractor analysis and item discrimination analysis. Zero responses in Table 6 indicate that no one selected the distractors. Items 1, 2, 4, 8, 10 and 13 show that the same amount of top and lower scorers answer correctly and incorrectly. It reveals that these items have no power to discriminate top and lower scorers. If a distractor elicits few or no responses, then it may not be functioning well and should be replaced with a more attractive option (Henning, 1987 p.55). A total of 30 distractors don't function well due to the fact that no responses are elicited from top or low scorers respectively based on Table 6.

### Identifying Problematic Items

Problematic items are those which don't meet the following criteria and should be deleted and replaced. The criteria of normal items are stated as follows:

1. Item facility index is between 0.15-0.85.
2. Item discrimination index is 0.2 or over.

Based on the criteria set above, there are 39% problematic items in this test. They have no discrimination and tend to be too easy for students.

### Table 7: Problematic Items of Four Categories

<table>
<thead>
<tr>
<th>Problematic Items</th>
<th>Item No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category</td>
<td></td>
</tr>
<tr>
<td>Listening</td>
<td>1, 2, 5, 6, 8, 10, 11, 15, 16, 17, 21, 22</td>
</tr>
<tr>
<td>Grammar</td>
<td>26, 28, 31, 33, 36</td>
</tr>
<tr>
<td>Reading</td>
<td>43, 45, 46, 47, 48, 51, 52, 53, 57</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>64, 72, 75</td>
</tr>
</tbody>
</table>

Total: 29 items

A total of 39 percent of items are problematic items. The problematic items are those which don't function adequately. They serve no purpose and should be revised or replaced in order to yield a better test. For example, the problematic item, 75, in the vocabulary section:

5. Later she blamed her daughter for losing the keys to the house.

A. scolded  B. cured  C. praised  D. curried

There are four students in lower scorers who chose the correct answers but no one in the top scorers answered correctly. The reason may be students in lower scorers happen to make right guesses but those in
top scorers don't. Another problematic item, 26, is in the grammar section:

1. It is a book. _____ title is World Link. A. It's  B. Its  C. His

More students in lower groups answer correctly than those in top groups. The problem may lie in the fact that Chinese students are confused about the contracted form and possessive adjectives. These problematic items cannot really discriminate students’ ability of English.

RESULTS

Based on the above analysis, item facility indices show that there are more easy items in listening tests (32%) and reading tests (30%) than in grammar tests (7%) and vocabulary tests (0%). More difficult items occur in listening tests (20%) and vocabulary tests (40%) than in grammar tests (13%) and reading tests (5%).

In terms of item discrimination, more items in listening tests (48%), reading tests (40%) and grammar tests (40%) than in vocabulary tests (20%), show little or no discrimination power.

Concerning distractor analysis, there is a higher proportion of useless distractors in listening tests (35%) than in grammar tests (17%) and vocabulary tests (20%).

A total of 29 items (39%) are found to be problematic items. There are 12 problematic items in listening tests, 6 in grammar tests, 8 in reading tests and 3 in vocabulary tests.

On the whole, vocabulary tests are the best tests among the four categories in that they are the most difficult, display the most powerful discrimination and have the least problematic items. Items rejected as having inappropriate difficulty, no discrimination, and little variability should be revised.

FINDINGS AND DISCUSSION

Based on item analysis, item discrimination, distractor analysis, as well as the questionnaire collected, the following points were observed:

First, concerning four tests, vocabulary seems too difficult for students since there are no easy items in this category. Students reflect that some words have never been seen on the textbook.

Items of listening tasks have the least power in discriminating between top and lower scorers and are too easy. Achievement tests should include all parts of the course content. The test content lacks pronunciation tests like stressed syllables and contraction forms. It will be better if phone numbers or e-mail addresses are included in the listening test. Numerals are very good materials for listening comprehension tests. Also, the listening cloze in part three of the listening
section may test common knowledge without the comprehension of the listening passage. Thus, it isn't really testing the listening ability.

The instructions in the grammar section “multiple-choice” seem to be misleading and easily cause misunderstanding among the students. Some students assume that there are more than one correct options and choose several answers for an item. Actually, only one option for each item is correct. The appropriate translation of multiple-choice should be “選擇題.” Grammar tasks overlap with the editing task in part three of the reading section.

For reading tests, purely reading skills, such as skimming or identifying cohesive markers can be tested for actually testing reading ability. In addition, an ordering task in part two of the reading section, which is naturally a kind of writing test, should be put under a category of the writing test instead of that of the reading test.

Secondly, with respect to test formats, the inconsistency of the quantity of response options in multiple-choice formats appears in the listening task and the vocabulary task. There are three response options in the listening task but four options in the vocabulary task. Thirdly, response distribution frequency can be very useful data when test developers think about making revisions, as distractors can pinpoint the weaknesses of particular items. Besides, item analysis also reveals that there is variance among students of the same level English class. This reflects a question: “Does the placement test work well?” If the placement test does well in dividing students into different levels, then why on the same test, students’ scorers spread so widely as Table 2 indicates?

On the other hand, based on the analysis of items, it’s found that there are 39% problematic items in this test. The percentage seems too high. These problematic items need to be revised. The vocabulary test items seem a little difficult but they have better discrimination power between top and lower scorers. Listening and reading tests have higher proportions of easy items and aren’t powerful enough in discrimination. Grammar tests seem more reasonable with reference to the score gained by students.

**SUGGESTIONS**

The following suggestions are proposed. It is hoped that appropriate revisions are to be made based on the findings of the study.

Firstly, one of the origins for unreliability is unclear instruction (Hugh, 1998).

Test instructions should be precise and clear enough. A test is said to be valid if it measures accurately what it is intended to
measure. If testing reading comprehension, it should be sure to test only the reading skills without mixing with other skills, i.e. writing or listening. For example, part two of the reading section is an ordering task testing mainly writing performance. Part three of the listening section is a partial dictation, which integrates listening comprehension with reading and writing skills.

Secondly, the purpose of the course is to develop communicative ability of English. Testing language use instead of language usage should be encouraged. Test tasks designed should be as more authentic as possible to match the objectives of the course. Otherwise, successful performance on the test may not truly reflect successful achievement of course objectives (Hugh, 1998 p.11). In this regard, oral tests such as oral interview and self-introduction DVD are conducted separately to supplement the insufficiency of the achievement test. The multiple-choice items increase the chance of guessing the answers. Multiple-choice formats may not truly test students’ mastery of the course. Different types of test formats can replace multiple-choice formats. An achievement test should reflect the goal of the course. It should test not only receptive skills, such as listening and reading skills but also productive skills, like speaking and writing skills. There has been an oral interview test outside of class. But no writing tests are constructed in this test.

Thirdly, each item should have at least four response options and the effectiveness of distractors increases reliability (Lado, 1961 p.341). Too many inappropriate distractors exist in this test. They need to be revised or replaced.

The use of distractors greatly influences item facility indices and item discrimination indices, so the selection of distractors must be very careful. Also, items that students are able to answer without reading the text should be discarded.

**CONCLUSION**

This study has examined listening, grammar, reading and vocabulary tests in an achievement test. Item analysis reveals that the most difficult items are vocabulary tests and easier items appear in listening tests. Vocabulary items have better discriminating power to distinguish top scorers and low scorers. More nonsense distractors occur in listening tests. Besides, about 39% problematic items are inappropriate to evaluate students’ learning as an achievement test. The outcome of the item analysis suggests that the test needs to be revised.

Achievement tests play an important role in students’ language programs.
Achievement tests should meet students’ demand. Students who study hard should get higher scores than those who don't study hard. Too easy or difficult items can't test students’ actual performance. On the other hand, achievement tests should match the objectives of the program and detect what students have learned. In order for achievement tests to be fair to reflect these respects, items need to be appropriately constructed.

For teachers and test developers, developing good or poor tests is closely related to the processes of the curriculum. Good tests help teachers fulfill the objectives of the course. Analyzing the test, actually used in the classroom, provides information with regard to the quality of the test, students’ performance and teachers’ efforts. The result of item analysis can be a basis of the modification of teaching content and materials.

It is highly recommended that achievement tests should be examined after they are scored, based on an objective item analysis as well as teachers’ intuitions, experiences and knowledge in order to effectively measure students’ performance of learning.

**BIBLIOGRAPHY**


Heinle & Heinle Publishers.


Lin, Li-Chan, Tseng, Hui-Min & Wu, Shiao-Chi. (1999). Item analysis of the registered nurse licensure exam taken by nurse candidates from vocational nursing high schools in Taiwan. Vol. 9, No.1. (pp.24-31). National Science Council, ROC.


Appendix I

Questionnaire

Female _______ Male _______

In response to Unified Mid-term Exam of Freshmen English, please answer the following questions based on your own opinion:

1. _____ Do you think the mid-term exam is
   A. very difficult  B. a little difficult  C. OK  D. very easy
2. _____ Which sections do you think are difficult for you?
   A. Listening  B. Grammar  C. Vocabulary  D. Reading
3. _____ Do you think the time is enough for you to finish the test?
   A. very enough  B. enough  C. OK  D. not enough
4. _____ Do you think the speaker in the listening test speaks clearly?
   A. very clear  B. clear  C. OK  D. not clear
5. _____ Which kind of test format do you prefer?
   A. multiple-choice  B. fill-ins  C. true and false  D. sentence-writing
6. _____ Which kind of test format do you think is most difficult for you?
   A. multiple-choice  B. fill-ins  C. true and false
7. _____ Do you think test items are too many for you?
   A. too many  B. a little many  C. OK  D. too few
8. _____ Are you satisfied with your test score?
   A. very satisfied  B. a little satisfied  C. OK  D. dissatisfied
9. _____ Why do you think you don't do well in this test?
   A. poor background  B. not well-prepared  C. OK  D. dissatisfied

10. What do you want to express about this mid-term English exam?

____________________________________________________________________

____________________________________________________________________
Appendix II
Level One Unified Midterm Exam Fall Semester

SECTION ONE: LISTENING (25%)

Part One: Short Answer (10 pts) 聽完之後請選擇最佳的回答
Directions: Pick the best response to what you have heard.

______ 1. A. My name is Song Ling B. I'm from Chang Hua C. I'm a student.

______ 2. A. I'm very happy.  B. Nice to meet you, too.  C. How are you?

______ 3. A. My family name is Wang  B. My given name is Alice.
   C. My friends call me Penny.


______ 6. A. Yes, she's Brazilian. B. Yes, she's American. C. Yes, she's Italian.

______ 7. A. Yes, she's from London.  B. Yes, she's from Paris.
   C. Yes, she's from Washington.

______ 8. A. It's Beijing.  B. It's very modern.  C. It's in the north of Taiwan.

______ 9. A. I'm from Taipei.  B. It's very modern. C. It's in the north of Taiwan.

______ 10. A. I'm Spanish. B. I'm with my friend Paul. C. I'm in San Francisco.

Part Two: Dialog (5 pts.) 聽完對話之後請回答下列問題
Directions: Listen to the dialog and answer the questions on your paper.

______ 1. Why does the man want Susana to call him Michey?
   A. He looks like a mouse. B. It is his nickname.  C. It is his given name.

______ 2. What is Mickey's nationality?
   A. Filipino  B. Japanese  C. Korean

______ 3. How long has Susan been in the United States?
   A. half a year  B. a year  C. a year and a half
An Item Analysis of an English Achievement Test Taken by EFL College Students in Taiwan

_____ 4. How is New York City compared to Mickey and Susana's hometowns?
A. It is bigger.  B. It is more crowded C. It is more expensive.

_____ 5. Where will the two classmates go after class?
A. to the library  B. to a restaurant  c. to a nightclub

Part Three:  Fill-ins (10 points) 填空
Listen to the paragraph and write the words that are missing.
Let me introduce myself. My (1) ___________ is Amanda. My friends all call me Mandy. It's my (2) _______________ . I'm from Brazil. I live in the (3) ___________. Brasilia. Brasilia is very (4) ________________ . It is also a very (5) ________________ city. It's not too (6) ________________, either. There are lots of (7) ____________ and the food is very delicious. Now I live in Taiwan. Taiwan is very small and very (8) _________________. Everything is very (9) ________________ and convenient, though. I like the people and there are many beautiful (10) ________________ to visit. Taiwan is a nice place to live.

SECTION TWO:  GRAMMAR (25%)
Part One:  Multiple Choice (10 points) 複選題
_____ 1. It is a book. ____ title is World Link.  A. It's  B. Its  C. His
_____ 2. They are college students. _____ school is Feng Chia University.
   A. Theirs   B. They're   C. Their
_____ 3. A: What's his name?  B: ______ name is Joe. A. He's  B. It's C. His
_____ 4. A: Is her name Sara?  B: A. Yes, she does. B. Yes, it is. C. Yes, she is.
_____ 5. A: Are you in Ms. Lin's class?
   B: A. No, I am. B. Yes, I do. C. No, I am not.
_____ 6. You can call me ________ my nickname. A. at  B. for  C. by
_____ 7. Is Rosa ___________? A. Brazilian  B. Brasilia  C. Brazil
_____ 8. Is he from _________? A. Argentine  B. Argentina  C. Argentinian
_____ 9. Hong Kong is _________. A. noise  B. noisiness  C. noisy
_____10. New York is very _______. A. interested  B. Interesting  C. interest
Part Two: Fill in the blank (10 points) 填空
1. A: Who's she? B: ____________ name is Joanne. She's an athlete.
2. A: Are you from Brazil? B: No, I'm from Mexico.
   A: So you are _____________.
3. I love Tom Cruise. ________________ my favorite actor.
4. A: ________________'s with you?
   B: I'm here with Yun-sun. He's from Korea.

Part Three: Correct the following grammatical errors (5 points) 改錯
Ex. You is a student → You are a student.
1. Do you in this class? ___________________________________________________________________
2. Hers name is Helen. ___________________________________________________________________
3. Chinatown is a crowd neighborhood. ___________________________________________________________________
4. Where is your teacher come from? ___________________________________________________________________
5. What your city like? ___________________________________________________________________

SECTION THREE: VOCABULARY (25%)
Part One: Complete each sentence with the most appropriate word in the box.
(20 points) 填入最適字
crowded essence avoid envy encourage capital
gradual sympathy remedy barber nickname bend
1. Seoul is the ____________ of Korea.
2. The Feng Chia Night market is ________________ with people every night.
3. Red got his ________________ because he has red hair.
4. Chicken soup is a good ________________ for a cold.
5. Jack has a great car and a beautiful girlfriend, I really ________________ him.
6. Teachers should ________________ children to be polite and do their homework.
7. The change didn't happen quickly, it was very ________________.
8. Because their neighborhood is very dangerous, they ________________ going out alone after dark.
9. A ________________ is a man whose job is cutting men's hair.
10. He wanted to express his deep ________________ to John's family after his accident.
An Item Analysis of an English Achievement Test Taken by EFL College Students in Taiwan

Part Two: Choose the best synonym (words with the same meanings) for each underlined word. (5 points)
選出同意字

1. I spend a lot on expensive jewelry and clothing every month.
   A. costly     B. curse    C. vain   D. noisy

2. He stood there breathing deeply and evenly.
   A. upsetting    B. kicking    C. exhaling  D. greeting

3. She confessed to her parent for her desire to being an actress.
   A. invited     B. argued    C. excused   D. admitted

4. That he wrote such an excellent book is astonishing. I can’t believe it!
   A. resisting    B. confusing  C. amazing  D. interesting

5. Later she blamed her daughter for losing the keys to the house.
   A. scolded    B. cured    C. praised   D. curied

SECTION FOUR: READING (25%)

Part One: Read the following postcard and answer the True and False questions that follow. 閱讀測驗 (10 points) Please CIRCLE O your answers.

Dear Meg, July 3, 2005

I'm on vacation in Miami, Florida. It is so beautiful here. Right now I'm in a seafood restaurant near the ocean. There are lots of people from Cuba, Puerto Rico, and Mexico in this part of the United States. Everywhere we go, we hear people speaking Spanish. It's really interesting.

On the front of this postcard is a picture of the beach we are staying at – Long Beach. It's pretty crowded but it's a great place. There are lots of things to do. We can go swimming, go fishing, play volleyball, or just relax in the sun. In the evening, we go out to eat and then watch the sunset on the beach. Sometimes we go dancing at a nightclub on the beach which is really fun. I really like Miami. The people are friendly and the food is great. Not only that, but things are not very expensive here, either. The buses and transportation are convenient, too. Next year, you'll have to come here with me!

See you soon!

Tanya

T    F   1. Tanya lives in Miami.
T    F   2. Many people speak Spanish in Miami.
T  F  3. Tanya doesn't like the beach.
T  F  4. There isn't much to do in Miami.
T  F  5. There are many people on the beach.
T  F  6. There are many activities to do on the beach.
T  F  7. Tanya likes to go dancing.
T  F  8. Tanya thinks things cost a lot in Miami.
T  F  9. It is difficult to travel around in Miami.
T  F  10. Tanya thinks Meg would like to visit Miami.

Part Two: Order the words in the sentences so that they make sense. 句子重組 (5 points) Ex. Name his Mike is → His name is Mike
too nice it's to you meet _________________________________
color is what your favorite _________________________________
from where you are _________________________________
the France capital Paris of is _________________________________
extensive an London city is _________________________________
An Item Analysis of an English Achievement Test Taken by EFL College Students in Taiwan

台灣大學生英語成就測驗之試題分析

施玉勉

摘 要

本文是針對台灣的一所大學舉行的期中英語成就測驗作試題分析，評估試題在難易程度，鑑別力及誘答能力三方面的表現，並提出一些意見及建議。本文是從試題難易度、試題鑑別度及誘答選項分析三個方面作試題分析。研究議題包括 (1) 分析每個類型的題目是否太難、適中或太容易，(2) 試題能否鑑別高分學生群和低分學生群在每個題目上表現的差異 (3) 誘答選項分析是否展現誤答的功能。誘答選項分析只針對選擇題型。本文也對學生做問卷調查，訪查對考試的感想和意見，結果顯示學生對考試的感想與實際分析的結果吻合，皆認為字彙最難。整體研究顯示字彙測驗最難，也最有鑑別能力，聽力測驗較容易，聽力的誘答選項表現較差。試題分析的結果證明試題需作一些修正。作者基於分析結果提出討論及建議，期待能改善成就測驗的品質，更有效地評估學生學習的成果。

關鍵詞：試題難易度 試題鑑別度 誘答選項分析 成就測驗