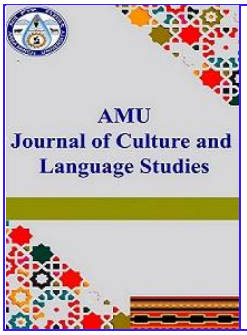


**AMU Journal of Culture and Language Studies**

**DOI:** https://doi.org/10.59122/1347376

**Volume: 2 Number: 1 , 2023 , Pages: 42~62**

**ISSN:2789-2581**

Received: 01 April 2023;
Accepted: 05 May 2023;
Published: 30 June 2023..

# Proposing Difficulty and Discrimination Level Indices to Validate Teacher-Made Multiple-Choice Test Items: A Remedy to Variations Observed in Scholars' Propositions

**Wendiyfraw Wanna**

[1]Arba Minch University

**ABSTRACT**

The purpose of this research article was to propose difficulty and discrimination level indices as a remedy to variations observed in those of scholars and check the efficacy of the proposed level of indices by comparing them with those of researchers. To achieve this, a multiple-choice vocabulary test was constructed and given to 4th year medical students at AMU. Then, a comparative item analysis was done with the researcher's proposed level of indices and with those of two researchers. The analysis revealed that the test had several items that needed to be discarded and replaced by better items as the items had poor discrimination index. This was due to the researcher's choosing the most frequent words from the Academic Word List. This was due to the test-takers' familiarity with the words. In addition, it was found that there were a few items that needed some revisions. The problems with these items were associated with having many non-functional distractors which attracted none or few of the test-takers. Besides, the comparison showed that the researcher's level of indices was better in some respects than those of the two researchers. It was recommended that the area needed further studies.

**Keywords**: multiple-choice, item analysis, difficulty index, discrimination index, distractor efficiency, stem, and distractors

## 1. INTRODUCTION

Scholars seemed to lack consistency in their use of item analysis parameters to validate their respective tests. In other words, every scholar used difficulty and discrimination indices of his/ her choice which were somewhat different from others. These variations were observed despite the educational level of the students who took the tests.

Scholars from general education and language education showed differences in their criteria to determine the difficulty and discrimination power of test items. Regarding item discrimination power, Ebel and Frisbie (1991) and Madsen (1983) appeared to have noticeable differences: Ebel and Frisbie (1991) proposed items having a discrimination index of < *0.19* were poor and thus they had to be rejected and Madsen (1983) claimed items having discrimination index of ≤*10* should be discarded because they were unacceptable. Similarly, Agrawal (1986) and Madsen (1983) exhibited differences. Agrawal (1986) believed that items with a discrimination index of *>0.20* were satisfactory but Madsen (1983) stressed that items with a discrimination index of ≥*15* were acceptable. Concerning the difficulty of test items, Agrawal (1986) and Madsen (1983) had differences. Agrawal (1986) thought that items with a difficulty index of *<0.20* were difficult and should be revised but Madsen (1983) stressed that items with a difficulty index of *<30* were too difficult and needed revision. Besides, Madsen's (1983) acceptable range for the difficulty index appeared to be too wide or crude. It needed further plausible ramifications concerning language test items.

Several researchers conducted item analysis studies at postgraduate and undergraduate levels with different cut-off points and designations. Boopathiraj and Chellamani (2013) were interested in analyzing the quality of teacher-made multiple-choice test items primarily prepared for postgraduate students. They used the criteria: *>90= too easy; < 20 = too difficult; 50 = moderately difficult* to assess the difficulty level of the items whereas they reported to have used: *>.20 = satisfactory* to work out the discrimination power of the items. Their classifications were not as detailed. Similarly, Marie & Sreekala (2015) performed item analysis to validate a test they prepared for undergraduate physical science students. They used the following criteria to measure the difficulty level of the items: *<0.20 = very difficult, 0.20-0.50 = good, 0.50-0.80 = best, >80 = very easy* and they used the next scale to measure discrimination power: *< 0.19 = poor, 0.20 – 0.29 = marginal, 0.30- 0.39 = good, > 0.40 = very good discrimination*. In the same vein, Mahjabeen et al.'s (2017) study was aimed at evaluating the qualities of multiple-

choice test items of a midterm test in the Department of Pathology in Islamabad Mental and Dental College. They applied the following criteria: *>70%=too easy, 30-70%=average, 50-60%= good, <30%=too difficult* to determine difficulty level and *≤ 0.2= poor,0.21-0.24=acceptable,0.25-0.35=good, ≥ 0.36=Excellent* to work out the discrimination power. At the same time, Odukoya et al. (2017) conducted a study that aimed to analyze the quality of multiple-choice test items administered to undergraduate students who took university-wide courses at private universities in Nigeria. They used the following criteria: *>70= too easy; < 20 = too difficult* to work out the difficulty of items but they used: *< 0.166= poor discriminator*. Their classifications were less clearly specified than other researchers. Later, Musa et al. (2018) conducted an item analysis study aimed at testing the qualities of multiple-choice items in the Physiology examination administered to medical students at Khartoum University. The researcher used the following criteria: *<30% = too difficult, 30% - 70% = good, >70% = easy* to work out the difficulty of the items and *<0 = worst, <0.2 = poor, 0.20-0.29 = marginal, 0.30-0.39 = good, ≥ 0.40 = very good* to identify the discriminatory power of the items. Recently, Sharma (2021) constructed 20 multiple-choice test items to test English speech sounds. The test-takers were undergraduate English major students in Nepal. The researcher used the following criteria to work out test item qualities: *<0.20 = most difficult; 0.20-0.39 = difficult; 0.40-0.59 = moderately difficult; 0.60-0.79 = moderately easy; 0.80-0.89 = easy; >0.90 = easiest* to determine the difficulty level of items but he used: *<-0.01 = worst; <0.20 = not discriminating; 0.20-0.29 = moderately discriminating; 0.30-0.39 = discriminating; ≥0.40 = very discriminating* to find out the discriminating power of the items.

On the other hand, quite a few item analysis studies were done at lower grade levels. Karim et al. (2021) performed an item analysis of a teacher-made multiple-choice reading test administered to thirty-five junior high school students in West Nusa Tenggara Province. These researchers used the following criteria: *0.00-0.30 = difficult; 0.31-0.70 = moderate; 0.71-1.00 = easy* to determine the difficulty level of the items and they used: *-0.01= worst; 0.00-0.20 = poor; 0.20-0.29 = mediocre; 0.30-0.39 = good; >0.39 = excellent* to work out the discrimination power of the items. Quite lately, Marsevani (2022) has been concerned with analyzing the quality of multiple-choice items designed for elementary school young learners in Indonesia. He used the following classification:*< 30%= Difficult; 30 - 70% =Acceptable; > 70%= Easy* to work out the difficulty level whereas *<-0.01= defective item/wrong key; 0-0.19= poor discrimination power;*

*0.2 - 0.29= acceptable discrimination power; 0.3 - 0.39 =good discrimination; and > 0.4= excellent discrimination* to find out the discrimination index.

Apparently from the preceding descriptions, we noticed that scholars had shown variations in their identification of *easy*, *too easy*, or *very easy* items. Most scholars proposed items having > 70% were *easy* or *too easy* while many claimed >80% were *very easy* and still a group of researchers had the opinion that items with > 90 % were *too easy* or *easiest*. At the same time, the scholars had variations in identifying difficult items. Many of them stressed items with difficulty levels of < 20% were *difficult* or *too difficult* whereas equally quite a lot of them identified items with <30% were *difficult* or *too difficult*. Still others suggested items with < 40% were *difficult*. On the other hand, Mahjabeen et al.'s (2017) specification of the discrimination index was rather different from the rest of the scholars. Despite having little variations in determining levels of discrimination indices, scholars lacked unanimity to determine the difficulty level of test items.

In Ethiopia, quite a few studies tried to determine the quality of multiple-choice test items using different criteria. Hassen (2022) tried to run item analysis to validate national examination using very crude criteria. To check item difficulty, he used: *>0.85 = very difficult and <0.30 very difficult.* At the same time, he used: *0.30-0.39= reasonably acceptable and 0.40-0.85= very ideal* to identify the discriminatory power of the test items. Lalem M. et al., (2022), on their part, attempted to check the validity of multiple-choice items of qualification examination given to final-year intern medical students at Debre Tabor University. To determine the difficulty of a test item, they used: *<29= hard; 30-70= desirable; 71-79= moderately easy; >80= easy*. However, they used the following: *<0.00= negative; <0.19= poor; 0.20-0.29= acceptable; 0.30-0.39= good; ≥0.40= excellent* to check the discriminatory power of the test items.

On the other hand, Ethiopian teachers rarely checked the validity of their classroom tests using any of the criteria. If teachers were to use them, they would be in dilemma as to which criteria to use. If teachers were free to use different criteria for different groups of learners, how could they maintain the quality of their test items? The same items might be difficult and not discriminating against one group but easy and discriminating against others. Besides, how can they maintain fairness among their students because of applying different criteria for different groups? It is much more advisable to use consistent criteria to check test quality especially when students come from nearly the same socio-economic backgrounds as is the case in Ethiopia. Hence, the

current research was aimed at proposing workable criteria as a remedy to the variations observed among scholars so that teachers or researchers could use them to validate their tests consistently.

### 1.1 Objectives of the Study

The current study was aimed to:

1. propose workable difficulty and discrimination level of indices based on the scores that high and low-scoring students might score for a specific multiple-choice test item
2. perform item analysis of a teacher-made test to check the quality of the test items using the researcher's proposed criteria
3. check the efficacy of the proposed criteria by making comparisons to other researchers' criteria while performing item analysis of the vocabulary test.

## 2. LITERATURE REVIEW

### 2.1 Testing and its Importance

A test is a tool by which we collect data about test-takers' abilities, knowledge, or performance and make measurements (Brown, 2004). Hence, testing empowers us to make important decisions about the fate of the test-takers (Carroll, 1961). In this regard, tests serve different purposes in the educational sector. For example, tests are given to identify students' strengths and weaknesses and thereby they provide feedback to teaching (i.e. *diagnostic tests*) (Henning, 1987; Suppiah, 2020). This means that the teacher makes adjustments to instructional materials to remedy students' weaknesses. Tests are also given to sort out learners in various ability groups (i.e. *placement tests*). In this regard, educational institutions assign learners to different groupings based on the placement test scores. Furthermore, international organizations use tests to recruit learners with high proficiency levels (i.e. *proficiency tests*). In other words, they use tests as gate-keepers or decide whether the test-takers have the required level of proficiency to be able to pursue their studies in those educational institutions. If tests are so important, then much emphasis needs to be given to the quality of the tests. Therefore, *item analysis* has to be conducted to ensure the quality of the test items. As in the literature to date, item analysis is usually done exclusively on multiple-choice items.

**2.2 Multiple-choice Items**

Multiple-choice items are one of the objective test formats which consist of a stem and distractors. The stem is an incomplete sentence that is to be completed by one of the distractors or choices. One of the distractors is unequivocally the right answer that the test-takers are expected to choose. Multiple-choice items can easily and correctly be scored and save time (Jia, et al., 2020). Hence, multiple-choice items are so reliable and affordable that test constructors usually prefer them (Rauch & Hartig, 2010). Furthermore, such kinds of items provide high content validity by accommodating a large proportion of content in a particular test (Klufa, 2015).

Although multiple-choice items are easy to score, they pose difficulty to test constructors and time time-consuming. The difficulty is often associated with getting plausible distractors. Some of the problems while constructing multiple-choice include: test-constructors end up having more than two or more acceptable answers and some of the distractors may not attract the test-takers at all. To solve these problems and make the test items more efficient, the test has to be tried out on some learners and it should undergo item analysis.

In the process of test construction, test constructors engage in doing item analysis, especially for multiple-choice items. Primarily, item analysis is carried out to get a diagnostic assessment of what test-takers have learned or failed to learn (Boopathiraj & Chellamani, 2013**).** Furthermore, the purpose of carrying out item analysis is to maximize and maintain the qualities of a test (Sharma, 2021).

This can be achieved by keeping good items but revising or discarding poor items. Good items are believed to have average difficulty and good discriminatory power. However, poor items are believed to be too easy or too difficult. Furthermore, test items with poor or negative discriminatory power fall under the poor item category. Such items are usually discarded or replaced by others. Hence, doing an item analysis is expected of every worthy test constructor since maintaining good test quality should be the highest priority in the process of test construction.

In short, item analysis is a statistical technique used to assess how the test-takers have responded to multiple-choice test items, which eventually helps to determine the quality of the test as a

whole (Koçdar, et al., 2016). Item analysis focuses on three aspects of the multiple-choice test items: difficulty index, discrimination index, and distractor efficiency.

### 2.3 Difficulty Index

Difficulty index refers to how difficult or easy a multiple-choice item is for a particular group of test-takers. The difficulty index ranges from 0%-100% or 0.00-1.00. This means that the higher the percentage the easier the test item is. Conversely, the lower the percentage the more difficult the test item is. Despite variations in cut-off points regarding difficulty levels, most scholars agree that items with a difficulty index of >90% are thought to be too easy while those with <20% are believed to be too difficult (Boopathiraj & Chellamani, 2013**).** Agarwal (1986) suggests that an item with >50% difficulty index has an optimum difficulty. However, language experts seem to have somewhat a different suggestion. For example, Madsen (1983) states that an item with 30%-90% has an acceptable difficulty index. Similarly, Brown (2004) asserts that an item should have a 15%-85% difficulty index to be acceptable. Hence, we may notice variations among scholars in the range of acceptable difficulty indices.

### 2.4 Discrimination Index

The discrimination index refers to what extent a multiple-choice item discriminates between the performance of high-scoring and low-scoring students (Boopathiraj & Chellamani, 2013). As a rule of thumb, high-scoring students should get an item right while low-scoring students should get the item wrong. Most scholars think that the discrimination index ranges from 0.0-1.0 (Boopathiraj & Chellamani, 2013; Karim et al., 2021; Sharma, 2021; Shanmugam, et al., 2020). These scholars might have used 2(HC-LC)/N to do the computation. However, the researcher realized that the range could not exceed 0.50 when $\mathrm{D.I} = \frac{\mathrm{Hc-Lc}}{\mathrm{N}}$ was used according to the researcher's observations. As a rule, it is believed that the higher the discrimination index the more discriminating the item will be between high-scoring and low-scoring students. With such an item, high scorers are more likely to get the item right while low scorers get the item wrong. Agarwal (1986) suggests that items having a discrimination index between 0.20 to 0.29 are believed to have moderate discriminating power. However, Madsen (1983) states that an item should have ≥ 0.15 discrimination index to be acceptable. Brown (2004), on his part, asserts that an item with a 0.50 discrimination index has a moderate discrimination power. As he demonstrated on Page 59, Brown based his specifications on $\frac{HC-LC}{1/2N}$ . His difference from the

other scholars was that he took only one of the groups' numbers (10) to make the computation. Otherwise, the 0.50 discrimination index should have been the highest discrimination power according to my observations (Appendix 3). Thus, we may realize that scholars have variations in their proposals of the acceptable range of discrimination index.

### 2.5 Distractor Efficiency

The options in multiple-choice items include two elements: the key and distractors. The 'key' is the only correct option whereas the rest are distractors or wrong answers. In this regard, 'distractor efficiency' refers to the ability of a distractor to attract more low-scoring students than high-scoring ones. Thus, a distractor can be 'non-functional' or 'functional'. Tarrant et al. (2009), Vyas and Supe (2008), and Patil and Patil (2015) contend that distractors selected by ≤5% of the test-takers are referred to as 'non-functional' whereas those selected by ≥ 5% of the test-takers are said to be 'functional'. Distractor efficiency may range from 0-100%. A test item may have a Distractor Efficiency (DE) of 33.3%, 66.6%, or 100% when it has 3 distractors (Mahjabeen, et al., 2017). The distractor efficiency of a test is determined by the number of FD (functional distractors) divided by the total number of distractors the test contains and multiplied by 100.

$$\text{DE}= \frac{number\ of\ FD}{total\ number\ of\ distractors} \times 100$$

As a rule of thumb, a distractor should be revised or discarded on two conditions: (1) if it distracts few or no test-takers or (2) if it attracts the majority of the test-takers from both groups (Odukoya, et al. 2017).

## 3. METHODOLOGY

### 3.1 Design of the Study

The design of the current study was a descriptive case study as it tried to describe a particular group of university students' performance concerning specific multiple-choice test items of a teacher-made test. Furthermore, the study tried to figure out how the scholars might have worked out their propositions of levels of difficulty and discrimination indices and proposed new cut-off points to compensate for the variations observed.

### 3.2 Population and Sampling

The population of the study included university students who were taking courses at Arba Minch University specifically in the College of Medicine and Health Sciences. A simple random sampling method was used to select participants who belonged to one group. Of 80 fourth-year medical students, only 38 were selected to sit for a vocabulary test. As the test was primarily constructed to test the vocabulary knowledge of students at Intermediate B1 Level (Internationally Accepted Level), it would be suitable for medical students who were believed to have high proficiency levels. With the consent of the college dean and department head, a suitable time was arranged to administer the test to the sample students.

### 3.3 Data Collection Tools

The data for the current study were collected through two methods: a vocabulary test and newly proposed levels of difficulty and discrimination indices. The words for constructing the vocabulary test were drawn from AWL (Academic Word List) compiled by Averil Coxhead (1998), as cited in Shimmt (1997). Averil organized the words according to range, frequency, and uniformity. Serial numbers were used to indicate the frequency: 1 indicated the most frequent but 10 was the least frequent. Hence, the researcher selected only those words indicated as most frequent. These words constituted the keys (i.e. correct answers). The distractors, however, were provided by the researcher. The vocabulary test contained 30 multiple-choice items. The items were characterized by having an incomplete stem and four distractors. An attempt was made to make the distractors more plausible for the test-takers. The other tool was the newly proposed levels of difficulty and discrimination indices which were proposed based on the researcher's prior observations during his preparation to give a testing course to his postgraduate students majoring in TEFL.

### 3.4 Researcher's Observations

The researcher tried to figure out how the scholars worked out their proposals of multiple-choice test items' difficulty and discrimination indices. To do so, he tried to imagine what configurations may be possible to do the computations. That is to say, he needed to imagine the possible number of high-scoring and low-scoring test-takers who could get a particular multiple-choice item right. To do this, he had to decide upon the total number of test-takers. Suppose the

total number of test-takers was 39. Then, he divided 39 by 3, which yielded 13. To do the computations, the researcher had to list the test scores of all 39 test-takers in descending order. Next, he worked on the first 13 high scorers followed by the third 13 low scorers' test papers, and got the possible configurations as in Appendix 1. The configurations were put in pairs; for example, (13,12). The first one indicated all high scorers got the item right whereas the second one showed that all low scorers except one got the item right. Of course, these configurations could also apply in situations where the test-takers number was 39, 36, 33, 30, 27, 24, 21, or 18.

Having identified the possible configurations, the researcher proceeded to work out difficulty and discrimination indices with the assumption that the total number of test-takers was 39, 36, 33, 30, 27, 24, 21, or 18. To find out the difficulty index $Diff. = \frac{Hc+Lc}{N}$ was used where HC= high correct; LC= low correct; and N= total number of students in the high group and low group (Appendix 2). In addition, $\text{Disc.} = \frac{\mathbf{Hc-Lc}}{\mathbf{N}}$ was applied to work out the discrimination index (Appendix 3).

Based on these data, the researcher attempted to propose his difficulty and discrimination indices as shown in Table 1 and 2, respectively.

Table 1: Researcher's Difficulty Index

| Range | **Description** |
|---|---|
| ≥0.90 | too easy |
| 0.80-0.89 | Easy |
| 0.50-0.79 | moderate |
| 0.33-0.49 | difficult |
| <0.33 | too difficult |

Table 2: Researcher's Discrimination Index

| **Scale** | **Description** |
|---|---|
| <0.00 | Defective/ negative |
| 0.00 | Not discriminating |
| <0.10 | Poor discriminator |
| 0.11-0.19 | Less discriminating |
| 0.20-0.30 | Acceptable discriminator |
| 0.31-0.40 | Good discriminator |
| 0.41-0.50 | Very good discriminator |

### 3.5 Data Processing and Analyzing

First of all, 38 students' test papers were scored and then were put in descending order. The top 13 papers were identified as 'high scoring', the last 13 were taken as 'low scoring' and the middle group of 12 papers were put aside. Second, the number of test-takers who got each item right was counted for both groups separately. Third, FUNCTIONAL and NONFUNCTIONAL distractors were sorted out. 'Functional distractors' were those distractors selected by ≥5% of the test-takers whereas 'Non-functional distractors' were those distractors selected by <5% of the test-takers. Fourth, $Dif = \frac{Hc+Lc}{N}$ and $\text{Dis.} = \frac{\mathbf{Hc-Lc}}{\mathbf{N}}$ were used to work out difficulty and discrimination indices, respectively. Eventually, the researcher's indices and those of the other two researchers were used to make the comparison.

## 4. PRESENTATION OF THE FINDINGS

### 4.1 Checking for Normality

**Tests of Normality**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| vocabulary test scores | .161 | 38 | .014 | .939 | 38 | .039 |

a. Lilliefors Significance Correction

It was quite necessary to check for normality of the test scores before attempting to work out the internal consistency otherwise this could be a serious methodological flaw if it was not done (Ghasemi & Zahediasl, 2012). As in the data above, both Kolmogrov-Smirnov and Shapiro-Wilk tests showed that the vocabulary test scores were normally distributed as the P-values (0.14 and 0.39) were less than 0.05. Hence, it was possible to run the parametric test (i.e. Pearson's correlations) to check for internal consistency. To perform the correlation, split-half data was prepared by splitting the scores into even and odd items. Then, Pearson's correlation was run and the p-value was 0.51which suggested a medium strength of relationship between the test items.

### 4.2 Items with Unacceptable Difficulty and Discrimination Indices

Table 3: Items to be Discarded and Replaced

| Item no | HG | LG | HG+LG | Diff. I. | Interpretation | HG-LG | Dis. I. | Interpretation |
|---|---|---|---|---|---|---|---|---|
| 2 | 13 | 13 | 26 | 1.00 | too easy | 0 | 0.00 | no discri. |
| 3 | 12 | 12 | 24 | 0.92 | too easy | 0 | 0.00 | no discri. |
| 12 | 13 | 13 | 26 | 1.00 | too easy | 0 | 0.00 | no discri. |
| 13 | 13 | 13 | 26 | 1.00 | too easy | 0 | 0.00 | no discri. |
| 21 | 12 | 12 | 24 | 0.92 | too easy | 0 | 0.00 | no discri. |
| 23 | 13 | 13 | 26 | 1.00 | too easy | 0 | 0.00 | no discri. |
| 24 | 13 | 13 | 26 | 1.00 | too easy | 0 | 0.00 | no discri. |
| 11 | 13 | 12 | 25 | 0.96 | too easy | 1 | 0.04 | poor discr. |
| 20 | 13 | 11 | 24 | 0.92 | too easy | 2 | 0.08 | poor discr. |
| 28 | 13 | 11 | 24 | 0.92 | too easy | 2 | 0.08 | poor discr. |
| 26 | 3 | 1 | 4 | 0.15 | too difficult | 2 | 0.08 | poor discr. |
| 18 | 4 | 5 | 9 | 0.35 | difficult | -1 | -0.04 | defective |

HG= high group; LG = low group

The item analysis revealed that several items needed to be discarded and replaced by better ones. Of these, 7 items (Items 2, 3, 12, 13, 21, 23 and 24) were found to be too easy. Many scholars agree that such items are not worth testing because the test-takers have already mastered the vocabulary tested. Consequently, these items were reported to have no discrimination power as 3 of the distractors attracted none of the test-takers (Table 3). Hence, these items had to be discarded and replaced by vocabulary items that the test-takers had not yet mastered. At the same time, the analysis suggested that 3 items (Items 11, 20, and 28) were too easy but had poor discrimination power. This means that two of the distractors could not attract any of the test-takers while 1 of the distractors attracted only 1 of the test-takers. Still, these items should be discarded and replaced by other items testing vocabulary items that the students might not have mastered (Table 3). However, Item 26 was found to be too difficult and had poor discrimination power. Similarly, this particular item was not worth testing. Besides being too difficult, the distractors failed to attract the test-takers (Table 3). Therefore, this item should be discarded and replaced by another item testing vocabulary that the test-takers might not have mastered. The last item that should be discarded and replaced by a better item was Item 18. Unlike the rest of the test items, this item was found to be difficult and had negative discrimination which was an indication of abnormality. That is to say, more students from low scorers answered it correctly than high scorers. Therefore, this item should be discarded and replaced by another item.

### 4.3 Items with Unacceptable Discrimination Indices

Table 4: Items to be Revised and Improved

| Item No | HG | LG | HG+LG | Diff. I. | Interpretation | HG-LG | Dis. I. | Interpretation |
|---|---|---|---|---|---|---|---|---|
| 1 | 12 | 11 | 23 | 0.88 | Easy | 1 | 0.04 | poor discr. |
| 10 | 11 | 10 | 21 | 0.81 | Easy | 1 | 0.04 | poor discr. |
| 6 | 10 | 9 | 19 | 0.73 | Moderate | 1 | 0.04 | poor discr. |

As in Table 4, the item analysis showed the presence of some items that needed only revision. These items were three in number. The first two items (Items 1 and 10) were found to be easy but had poor discrimination power. This might suggest that the right answer stood out clearly in comparison to the distractors. Again, the distractors were not strong enough to attract the test-takers. Specifically, two of the distractors were selected by none of the test-takers whereas 1 distractor was selected by 1 test-taker. Similarly, Item 6 was found to be moderate or acceptable in difficulty level but it had poor discrimination power. So, the problem with these items was the selection of distractors. Hence, these items required replacing these distractors with different ones.

### 4.4 Distractor Efficiency

The vocabulary test had 30 items and each item contained 3 distractors and 1 key. Hence, the number of distractors was 90 (30x3). The total number of functional distractors amounted to 45. Mahjabeen, et al. (2017) contended that DE (distractor efficiency) should be calculated by dividing FD (functional distractor) by the total number of distractors and multiplying the dividend by 100. Hence, the overall DE of the vocabulary test was found to be 50% = 45/90 x100. This means that the distractor efficiency was moderate as half of the distractors (50%) were selected by <5% of the test-takers (Appendix 4). Hence, the test contained a lot of non-functional distractors which needed to be revised or discarded.

As in Appendix 4, 23.3% of the test items had a DE of 100%. This means that all of the distractors were able to attract the test-takers. So, they are functional distractors. At the same time, 20% of the test items had a DE of 66.6%; meaning, two of the distractors could attract the test-takers. One of the distractors should be discarded and replaced by a good one. Similarly, 40% of the test items had a DE of 33.3%. That means only one of the distractors was able to attract the test-takers. Two of them should be discarded and replaced by better distractors.

Eventually, 5.2% of the test items had a DE of 0%. That means none of the distractors attracted the test-takers. All of them are non-functional distractors and they should be discarded and replaced by good ones.

### 4.5 Discussion

As explained in the preceding sections, there appear to be variations in scholars' difficulty and discrimination indices. Besides, the scholars seem to be inclined to work with indices of their preferences. Hence, the discussion section tries to compare how the current researcher's proposed indices differ from those of two of these scholars.

#### 4.5.1 Comparing Researcher's Difficulty Index with that of Marsevani

Table 5: A Comparison between Researcher's and Marsevani's Difficulty Indices

| Item no | HG | LG | HG+LG | Diff. I. | Researcher's | Marsevani's |
|---|---|---|---|---|---|---|
| 2 | 13 | 13 | 26 | 1.00 | too easy | easy |
| 3 | 12 | 12 | 24 | 0.92 | too easy | easy |
| 11 | 13 | 12 | 25 | 0.96 | too easy | easy |
| 12 | 13 | 13 | 26 | 1.00 | too easy | easy |
| 13 | 13 | 13 | 26 | 1.00 | too easy | easy |
| 20 | 13 | 11 | 24 | 0.92 | too easy | easy |
| 21 | 12 | 12 | 24 | 0.92 | too easy | easy |
| 23 | 13 | 13 | 26 | 1.00 | too easy | easy |
| 24 | 13 | 13 | 26 | 1.00 | too easy | easy |
| 26 | 3 | 1 | 4 | 0.15 | too difficult | difficult |
| 28 | 13 | 11 | 24 | 0.92 | too easy | easy |

As noted in the introductory section, Marsevani (2022) used much sparser difficulty indices in comparison to that of the researcher: *P < 30% Difficult; P = 30 - 70% Acceptable; P > 70% Easy* to work out the difficulty index. In other words, Marsevani's 'easy' items started from 0.70 and encompassed what the researcher designated as 'too easy'. This may suggest that the researcher made more distinctions to isolate items not worthy of testing. Regarding Item 18, Marsevani assigns 'difficult' which the researcher identified as 'too difficult'.

### 4.5.2 Comparing Researcher's Discrimination Index with that of Sharma

Table 6: A Comparison between Researcher's Discrimination Index and that of Sharma

| Item no | HG | LG | HG-LG | Dis.I. | Researcher's | Sharma's |
|---|---|---|---|---|---|---|
| 4 | 13 | 10 | 3 | 0.12 | less discri. | not discriminating |
| 5 | 11 | 7 | 4 | 0.15 | less discri. | not discriminating |
| 7 | 11 | 7 | 4 | 0.15 | less discri. | not discriminating |
| 8 | 11 | 6 | 5 | 0.19 | less discri. | not discriminating |
| 14 | 12 | 8 | 4 | 0.15 | less discri. | not discriminating |
| 16 | 12 | 9 | 3 | 0.12 | less discri. | not discriminating |
| 17 | 10 | 7 | 3 | 0.12 | less discri. | not discriminating |
| 19 | 11 | 7 | 4 | 0.15 | less discri. | not discriminating |
| 22 | 13 | 10 | 3 | 0.12 | less discri. | not discriminating |
| 27 | 7 | 2 | 5 | 0.19 | less discri. | not discriminating |
| 29 | 13 | 8 | 5 | 0.19 | less discri. | not discriminating |
| 30 | 11 | 7 | 4 | 0.15 | less discri. | not discriminating |

In the introductory part, it was shown that Sharma (2021) used a bit different discrimination index when compared to that of the researcher: *negative = worst; <0.20 = not discriminating; 0.20-0.29 = moderately discriminating; 0.30-0.39 = discriminating; ≥0.40 = very discriminating* to find out the discriminating power of the items. Sharma designated all indices below 0.20 as 'not discriminating'. This may imply that he considered a difference of 3-5 between high and low-scoring students was not sufficient to discriminate between test-takers. However, the researcher questions this claim. If we take item 8 above, out of 13 test-takers, 11 got the item right from high-scorers and only 6 out of 13 got the item right from low-scorers. So, how can we say that this item is not discriminating? However, the researcher believes that this item and the others having differences of 3-5 should be labeled as 'less discriminating' as in Table 2 above.

## 6. CONCLUSION AND RECOMMENDATIONS

As noted in the introductory section, the researcher noticed disparities in the indices used by different researchers. As a result, he tried to propose his indices as a remedy to the variations observed. Then, the researcher constructed a test to measure university students' vocabulary knowledge. After scoring, he conducted item analysis and found that some of the test items needed to be discarded while a few of them needed some revision. Furthermore, he found that almost half of the distractors were not efficient. Next, the researcher made a comparison between his indices and those of other researchers to show differences. The analysis revealed some differences were identified indicating the efficacy of the researcher's indices. Finally, the researcher recommends that further studies need to be done in the area as this study focuses on students with high proficiency in the Ethiopian context.

## 7. REFERENCES

Agarwal. Y.P. (1986). *Statistical methods, concepts, applications and computations*. New Delhi: Sterling Publication.

Boopathiraj, C., & Chellamani, K. (2013) Analysis of test items on difficulty level and discrimination index in the test for research in education. *International Journal of Social Science & Interdisciplinary Research*, *2*(2):189-193.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. White Plains, NY: Pearson Education.

Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In Allen, H. B. and Campbell, R. N. (Eds.) (1965), *Teaching English as a second language: A book of readings* (pp. 313–330). New York: McGraw Hill.

Ebel RL, Frisbie DA. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, New Jersey: Prentice-Hall.

Ghasemi, A., & Zahediasl, S. (2012) Normality tests for statistical analysis: A guide for non-statisticians, *Int J Endocrinol Metab*, *10*(2), 486-9. https://doi.org/10.5812/ijem.3505

Hassen, H. (2022) How Ethiopian Standardized National Examinations Achieve Their Goal? 2014/15 University Entrance Examination Exam in Focus, *African Journal of Social Sciences and Humanities Research, 5(3): 1-14.*

Henning, G. (1987) *A guide to language testing: Development, evaluation, and research*.

Asia: Heinle & Heinle Publishers.

Jia, B., He, D., & Zhu, Z. (2020) Quality and Feature of Multiple-choice Questions in Education. *Problems of Education in the 21st century*, 78(4):576-594. https://doi.org/10.33225/pec/20.78.576

Karim, S. A., Sudiro, S. & Sakinah, S. (2021) Utilizing test items analysis to examine the level of difficulty and discriminating power in a teacher-made test. *Journal of English Education, Literature, and Culture, 6*(2), 256-269.

Koçdar, S., Karadaö, N., & Sahin, M.D. (2016). Analysis of the difficulty and discrimination indices of multiple-choice questions according to cognitive levels in an open and distance learning context. *The Turkish Online Journal of Educational Technology, 15*(4), 16-24.

Klufa, J. (2015). Multiple choice question tests–advantages and disadvantages. Mathematics and Computers in Sciences and Industry Journal, 3, 91-97.

Lemlem M., Tegbar Y., Fantu A. (2022) Quality of multiple-choice questions in medical internship qualification examination determined by item response theory at Debre Tabor University, Ethiopia, *BMC Medical Education, 22(635): 1-11*.

Madsen, H.S. (1983). *Techniques in testing*. Hong Kong: OUP.

Mahjabeen, W., Alam, S., Hassan, U., Zafar, T., Butt, R., Konain, S., & Rizvi, M. (2017). Difficulty index, discrimination index, and distractor efficiency in multiple-choice questions. *Annals of PIMS*, 310-315.

Marie, S. & Sreekala, E. (2015) Relevance of Item Analysis in Standardizing an Achievement Test in Teaching of Physical Science in B.Ed Syllabus, *i-manager's Journal of Educational Technology, 12* (*3): 30-36.*

Marsevani, M. (2022). Item analysis of multiple-choice questions: an assessment of young Learners. *English Review: Journal of English Education, 10*(2), 401-408. https://doi.org/10.25134/erjee.v10i2.6241.

Musa, A., Shanee, S., Elmardi, A., & Ahmed, A. (2018). Item difficulty and item discrimination as quality indicators of physiology MCQ examinations at the Faculty of Medicine Khartou University, *Khartoum Medical Journal, 11*(2), 1477-1486.

Odukoya, J., Adekeye, O., and Igbinoba, A. (2017). Item analysis of university-wide multiple choice objective examinations: the experience of a Nigerian private university. *Qual Quant 52: 983-997.*

Patil V. C, & Patil, H. V. (2015). Item analysis of medicine multiple choice questions (MCQs) for under graduate (3rd year MBBS) Students. *Research Journal of Pharmaceut Biol Chem Sci*, *6*, 1242-1251.

Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two dimensional IRT analysis. Psychological Test and Assessment Modelling, 52(4), 354–379.

Sharma, L. R. (2021) Analysis of Difficulty index, discrimination index and distracter efficiency of multiple-choice questions of speech sounds of English. *International Research Journal of MMC (IRJMMC), 2*(1), 15-28. https://doi.org/10.31126/irjimmc. 35126

Schmitt, N. (2000) *Vocabulary in language teaching*. USA: CUP.

Shanmugam, S.K.S., Wong, V. & Rajoo, M. (2020) Examining the Quality of English Test Items Using Psychometric and Linguistic Characteristics Among Grade Six Pupils. *Malaysian Journal of Learning and Instruction. 17(2): 63-101.*

Suppiah, S. K., Wong, V., & Rajoo, M. (2020). Examining the quality of English test items using psychometric and linguistic characteristics among grade six pupils. *Malaysian Journal of Learning and Instruction*, *17*(2), 63-101.

Tarrant, M., Ware, J., & Mohammed, A.M. (2009). An assessment of functioning and nonfunctioning distractors in multiple choice questions: a descriptive analysis. *BMC Medicine Education*, *9*(40), 2-8.

Vyas, R., & Supe, A. (2008). Multiple choice questions: A literature review on the optimal number of options. *National Medical Journal, India*, *21*, 130-133.

## 8. APPENDICES

Appendix 1: Test score probabilities of 26 high and low-scoring test-takers

| **39 test-takers** | **N=26** | | **39 test-takers** | **N=26** | |
|---|---|---|---|---|---|
| 13,12 | 25 | 96% | 12,0;11,1;10,2;9,3;8,4;7,5;6,6 | 12 | 46% |
| 13,11;12,12 | 24 | 92% | 11,0;10,1;9,2;8,3;7,4;6,5 | 11 | 42% |
| 13,10;12,11 | 23 | 88% | 10,0;9,1;8,2;7,3;6,4;5,5 | 10 | 38% |
| 13,9;12,10;11,11 | 22 | 85% | 9,0;8,1;7,2;6,3;5,4 | 9 | 35% |
| 13,8;12,9;11,10 | 21 | 81% | 8,0;7,1;6,2;5,3;4,4 | 8 | 31% |
| 13,7;12,8;11,9;10,10 | 20 | 77% | 7,0;6,1;5,2;4,3 | 7 | 27% |
| 13,6;12,7;11,8;10,9 | 19 | 73% | | | |

| **39 test-takers** | **N=26** | | **39 test-takers** | **N=26** | |
|---|---|---|---|---|---|
| 13,5;12,6;11,7;10,8;9,9 | 18 | 69% | | | |
| 13,4;12,5;11,6;10,7;9,8 | 17 | 65% | | | |
| 13,3;12,4;11,5;10,6;9,7;8,8 | 16 | 62% | | | |
| 13,2;12,3;11,4;10,5;9,6;8,7 | 15 | 58% | | | |
| 13,1;12,2;11,3;10,4;9,5;8,6;7,7 | 14 | 54% | | | |
| 13,0;12,1;11,2;10,3;9,4;8,5;7,6 | 13 | 50% | | | |

Appendix 2: Difficulty indices of 39-18 test-takers

| **Test-takers** | **39** | **36** | **33** | **30** | **27** | **24** | **21** | **18** |
|---|---|---|---|---|---|---|---|---|
| **UG+LG*** | **N=26** | **N=24** | **N=22** | **N=20** | **N=18** | **N=16** | **N=14** | **N=12** |
| too easy | 96% | 96% | 95% | 95% | 94% | 94% | 93% | 92% |
| | 92% | 92% | 91% | 90% | 89% | 88% | 86% | 83% |
| easy | 88% | 88% | 86% | 85% | 83% | 81% | 79% | 75% |
| | 85% | 83% | 82% | 80% | 78% | 75% | 71% | 67% |
| | 81% | 79% | 77% | 75% | 72% | 69% | 64% | 58% |
| moderate | 77% | 75% | 73% | 70% | 67% | 63% | 57% | 50% |
| | 73% | 71% | 68% | 65% | 61% | 56% | 50% | 42% |
| | 69% | 67% | 64% | 60% | 56% | 50% | 43% | 33% |
| | 65% | 63% | 59% | 55% | 50% | 44% | 36% | 25% |
| | 62% | 58% | 55% | 50% | 44% | 38% | 29% | |
| | 58% | 54% | 50% | 45% | 39% | 31% | 21% | |
| | 54% | 50% | 45% | 40% | 33% | 25% | | |
| | 50% | 46% | 41% | 35% | 28% | 19% | | |
| difficult | 46% | 42% | 36% | 30% | 22% | | | |
| | 42% | 38% | 32% | 25% | | | | |
| | 38% | 33% | 27% | | | | | |
| | 35% | 29% | | | | | | |
| too difficult | 31% | 25% | | | | | | |
| | 27% | | | | | | | |

Appendix 3: Discrimination Indices of 39-18 Test Takers

| **test-takers** | **39** | **36** | **33** | **30** | **27** | **24** | **21** |
|---|---|---|---|---|---|---|---|
| **UG+LG** | **26** | **24** | **22** | **20** | **18** | **16** | **14** |
| Not discriminating | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Poor discriminator | 0.04 | 0.04 | 0.05 | 0.05 | 0.06 | 0.06 | 0.07 |
| | 0.08 | 0.08 | 0.09 | 0.10 | 0.11 | 0.13 | 0.14 |
| Less discriminating | 0.12 | 0.13 | 0.14 | 0.15 | 0.17 | 0.19 | 0.21 |
| | 0.15 | 0.17 | 0.18 | 0.20 | 0.22 | 0.25 | 0.29 |
| | 0.19 | 0.21 | 0.23 | 0.25 | 0.28 | 0.31 | 0.36 |
| Acceptable discriminator | 0.23 | 0.25 | 0.27 | 0.30 | 0.33 | 0.38 | 0.43 |
| | 0.27 | 0.29 | 0.32 | 0.35 | 0.39 | 0.44 | 0.50 |
| Good discriminator | 0.31 | 0.33 | 0.36 | 0.40 | 0.44 | 0.50 | |
| | 0.35 | 0.38 | 0.41 | 0.45 | 0.50 | | |
| | 0.38 | 0.42 | 0.45 | 0.50 | | | |
| Very good discriminator | 0.42 | 0.46 | 0.50 | | | | |
| | 0.46 | 0.50 | | | | | |
| | 0.50 | | | | | | |

Appendix 4: Distractor Efficiency

| **Item No** | **A** | **B** | **C** | **D** | **DE** | **categories** |
|---|---|---|---|---|---|---|
| 5 | 27 | 5 | 3 | 2 | 100% | 23.3% |
| 7 | 28 | 3 | 5 | 2 | 100% | |
| 8 | 5 | 27 | 3 | 3 | 100% | |
| 15 | 4 | 7 | 25 | 2 | 100% | |
| 18 | 15 | 5 | 11 | 6 | 100% | |
| 19 | 23 | 3 | 9 | 3 | 100% | |
| 27 | 12 | 10 | 14 | 2 | 100% | |
| 3 | 34 | 0 | 2 | 2 | 66.60% | 20% |
| 6 | 9 | 0 | 2 | 27 | 66.60% | |
| 14 | 30 | 5 | 3 | 0 | 66.60% | |
| 16 | 2 | 32 | 0 | 3 | 66.60% | |
| 25 | 0 | 16 | 19 | 3 | 66.60% | |
| 30 | 2 | 29 | 1 | 6 | 66.60% | |
| 1 | 3 | 35 | 0 | 0 | 33.30% | 40% |
| 4 | 2 | 1 | 35 | 0 | 33.30% | |
| 9 | 1 | 25 | 12 | 0 | 33.30% | |
| 10 | 1 | 2 | 33 | 0 | 33.30% | |
| 11 | 0 | 2 | 36 | 0 | 33.30% | |
| 12 | 36 | 0 | 2 | 0 | 33.30% | |

| Item No | A | B | C | D | DE | categories |
|---|---|---|---|---|---|---|
| 17 | 9 | 27 | 1 | 1 | 33.30% | |
| 20 | 5 | 0 | 33 | 0 | 33.30% | |
| 21 | 36 | 2 | 0 | 0 | 33.30% | |
| 22 | 0 | 0 | 35 | 3 | 33.30% | |
| 26 | 1 | 6 | 0 | 30 | 33.30% | |
| 29 | 33 | 2 | 1 | 1 | 33.30% | |
| 2 | 0 | 0 | 38 | 0 | 0% | 5.2% |
| 13 | 1 | 37 | 0 | 0 | 0% | |
| 23 | 38 | 0 | 0 | 0 | 0% | |
| 24 | 0 | 38 | 0 | 0 | 0% | |
| 28 | 0 | 36 | 1 | 1 | 0% | |

**Note:** The cells in yellow indicate the keys whereas the numbers in white background show functional distractors and those in red with light background refer to non-functional distractors. 45/90x100= 50% are non-functional distractors.