*International Journal of Learning, Teaching and Educational Research*
*Vol. 21, No. 3, pp. 36-50, March 2022*
*https://doi.org/10.26803/ijlter.21.3.3*
*Received Nov 8, 2021; Revised Feb 24, 2022; Accepted Mar 12, 2022*

# Item Analysis of a Reading Test in a Sri Lankan Context using Classical Test Theory

**Fouzul Kareema Mohamed Ismail**
PhD candidate in TESL, Department of Curriculum and Instruction,
International Islamic University Malaysia
Kuala Lumpur, Malaysia
*South Eastern University of Sri Lanka, Sri Lanka

**Ainol Madziah Bt Zubairi**
International Islamic University Malaysia
Kuala Lumpur, Malaysia

**Abstract.** This paper is based on a research study on a reading test that evaluates the different cognitive processes prescribed by Khalifa and Weir (2009). The 25-item test was designed based on a test specification targeted at the B2 level of the Common European Framework of Reference for Languages (CEFR). The responses of 50 students were used to check the validity and reliability of the test. The validity of the test was ascertained through item analysis involving item difficulty indices, item discrimination indices, and distractor analysis. Each item was studied to provide detailed information leading to the improvement of test construction. To achieve test reliability, the Kuder-Richardson Formula 20 (KR-20) was applied. The results were achieved by simply using Microsoft Excel. Findings revealed that the test met the standards for content validity, indicating acceptable item difficulty indices, with 17 items at the moderate level between the ranges of 0.30 and 0.79. Except for three items, all others functioned well to differentiate between high- and low-ability students, and only five items had malfunction distractors. Meanwhile, the reliability value of the test scores was 0.82, which is deemed a good value, proving the consistency of the test results. It signifies that more than half, that is 88%, of the test items were well functioning and that the test proved to be valid and reliable. The present research can contribute to students, teachers, and test-makers having an insightful understanding of item analysis and test development.

**Keywords:** cognitive processing in reading; distractor; item difficulty; item discrimination

## 1. Introduction

Reading is a multivariate ability that necessitates the complicated combination and integration of a wide range of linguistic, non-linguistic, and cognitive skills, ranging from extremely basic low-level processing abilities to high-level processing abilities. Assessing or testing reading is also a complex phenomenon (Alderson, 2000).

A test can be defined as a method of measuring a person's ability, knowledge, or performance in a specific domain (Brown & Abeywickrama, 2010). In this study, a test was utilized as method to measure the reading performance level of selected students of the South Eastern University of Sri Lanka. Considering the significance of assessing reading, the test was developed using the theoretical backgrounds in test development and validation to provide the necessary information for test designers and teachers in achieving a valid and reliable test that assesses what it is supposed to assess. The test paper consisted of 25 items, including 18 multiple-choice questions (MCQs) with 4 options each and 7 questions of multiple matching. Both response types belong to the category of selected-response methods (Khalifa & Weir, 2009; Urquhart & Weir, 1998).

Recent developments in research have recognized reading as a mental process, identified as cognitive processing by Khalifa and Weir (2009). These cognitive processes are included in this study to measure the reading performance level of the students. To do this, eight socio-cognitive processes were identified to evaluate the reading performance. These are: word recognition (WR), lexical access (LA), syntactic parsing (SP), establishing propositional meaning (EPM), inferencing (I), building a mental model (BMM), creating a text-level structure (CTLS), and creating an inter-textual representation (CITR). The first four processes have been categorized as low-level process (lower order thinking [LOT]) skills, with the latter four categorized as high-level process (higher order thinking [HOT]) skills (Bax & Chan, 2016). This paper highlights the utility of the classical item analysis in providing detailed information about how items function in a test.

## 2. Literature Review

In ascertaining the reliability and validity of a test, knowledge of test development and validation is of great importance. In developing a test, test and item design is pivotal. Furthermore, the validity and reliability of a test require theoretical knowledge of item analysis. This study adopted an approach based on classical test theory (CTT).

### 2.1 Test Development and Validation

The validity and reliability of the test determine whether the test serves its purpose well. Validity is the extent to which a test measures what it is meant to measure (Messick, 1989). To check whether test items are valid, content validation of the test items is crucial before administering a test (Halek et al., 2017). As Creswell (2012) and Crocker and Algina (1986) mentioned, content validation can be performed by a group of experts in a specific content area.

### 2.2 Classical Test Theory and Item Difficulty Analysis

CTT is based on a test score theory that introduces three concepts: the test score (also known as the observed score), the true score, and the error score (Eleje et al., 2018; Hambleton & Jones, 1993; Magno, 2009; Yusup, 2012). The main advantage of CTT, according to Hambleton and Jones (1993), is its "relatively weak assumptions (i.e., they are easy to meet in real test data)" (p. 40), which makes it simple to employ in a variety of testing scenarios. Item difficulty indices, item discrimination analysis, and distractor analysis are the primary features of CTT, which are key elements in item analysis as well. These three types of analyses can provide evidence for validity arguments, as supported by Tamil (2015), Manalu (2019), and Shanmugam et al. (2020).

Item analysis is the process of evaluating test questions involving a systematic procedure that provides specific information about the items constructed (Pratiwi et al., 2021). Barnard (1999) defined the item difficulty index as the proportion of the group who answer the questions correctly. This is determined by the total number of correct responses for a particular item divided by the total number of students and multiplied by 100 to get the percentage (Tamil, 2015).

$$\text{Difficulty index (D)} = \frac{\text{students with correct answers}}{\text{total number of students}} \text{ x } 100$$

However, Samad (2004, p. 103) did not convert the result into a percentage as in the present research. Item difficulty can range from 0 to 1 and often involves "decimal points". This can be calculated by using the number of students who answered an item correctly and dividing it by the number of students who attempted to answer the item. Some researchers have identified items with values less than 0.30 as difficult items, and those greater than 0.70 as easy items (Bichi & Embong, 2018; Shanmugam et al., 2020; Zubairi & Kassim, 2006). Nonetheless, this study used Tamil's (2015) recommendations. He categorized items scoring 0.0–0.29 as difficult, 0.30–0.79 as moderate, and 0.80–1.00 as easy items. Very easy and very difficult items may need to be revised to achieve a valid test (Bichi & Embong, 2018; Samad, 2004; Shanmugam et al., 2020; Tamil, 2015) if the test is achievement-based. A good test must have a variety of items, ranging from easy, moderate, to difficult items (Wright & Stone, 1979).

### 2.3 Item Discrimination

Item discrimination is a method used to determine how well an item distinguishes between pupils of high and poor ability. Item discrimination values range from -1 to 1 (Samad, 2004). The number of correct answers from students in the upper and lower ability groups, respectively, is used to measure item discrimination.

$$\text{Discrimination index (R)} = \frac{(H - L)}{27\% \text{ of total}}$$

H = number of correct answers from the top 27% of students
L = number of correct answers from the bottom 27% of students (Tamil, 2015).

The aim is to divide the group into three, the upper 27%, middle 46%, and lower 27%. Some textbooks use 25% as the cut-score instead of 27%. However, based on Kelley (1939), using upper and lower groups consisting of 27% from the extremes of the criterion score distribution is optimal for the study of test items (Tamil, 2015). Samad (2004), Tamil (2015), Bichi and Embong (2018), and Shanmugam et al. (2020) classified four types of item discrimination interpretations, as shown in Table 1, which is almost similar to Ebel and Frisbie's (1991) classification.

**Table 1: Item discrimination reading**

| Range | Verbal description |
|---|---|
| 0.40 & above | Very good item |
| 0.30–0.39 | Good item |
| 0.20–0.29 | Fair item |
| 0.09–0.19 | Poor item |

Items that fall under the poor item category should be revisited and eliminated from the test if there are no optimal justifications to have them.

## 2.4 Reliability of the Test

The most important element of CTT is test reliability, which is generally accepted as a requirement for a test to be recognized as adequate quality for practical usage (McNamara, 1996). The uniformity of measurement is referred to as reliability. Theoretically, it is identified as the ratio of observed-score variance caused by true-score variance (Crocker & Algina, 1986; Ebel & Frisbie, 1991). Reliability indicates the consistency of test scores or other evaluation outcomes from one measurement to the next. It has been established that the reliability of a test determines whether the test can be trusted following the criteria set out. When dealing with the same group at a different time or opportunity, a test is considered reliable if it consistently produces the same result (Samad, 2004).

Items that have only right and wrong answers are known as dichotomous. An MCQ has a right answer and two, three, or more options as wrong answers (Ebel & Frisbie, 1991). The right answer is known as the "key" and the wrong ones as the "distractors" (Kastner & Stangl, 2011, p. 265). The internal consistency of measures using dichotomous choices is checked with the index (i.e., correct versus incorrect) provided by Kuder and Richardson (1937) known as the Kuder-Richardson Formula 20 (KR-20). A right question receives a score of 1, whereas a wrong question receives a score of 0. The index values thus range from 0 to 1.

$$\boxed{KR\text{-}20 = \frac{K}{K-1}\left[1 - \frac{\sum PQ}{\sigma 2X}\right]}$$

KR-20 = Kuder-Richardson Formula 20
K = number of questions
Σ = indication to sum
P = probability of correct answer
Q = probability of the wrong answer
σ = variance of the total scores of all the people taking the test (adapted from Tamil, 2015).

Since the above formula has been widely used, we employed it to estimate the reliability of test scores of test items (Zimmerman, 1972). We aimed to analyze each test item to determine the level of difficulty, examine the discrimination index for revision, do a distractor analysis to discover the distractor that is malfunctioning, revise the items and malfunctioning distractors, and check the reliability of the test.

## 3. Methods

This research attempts to develop and validate a test by providing a table of specifications followed by information based on the item difficulty index, item discrimination index, distractor analysis of each item, and reliability analysis. The main data were collected in this study by administering a test to a selected sample of students.

### 3.1 Data Collection Instrument Development

A table of specifications was constructed to ensure the test items are content valid (Table 2). Content validation by experts is a must to produce an effective measurement tool (Turner & Carlson, 2003). The table of specifications is the output of the content validation by expert judgement. Fulcher and Davidson (2007) believed that "specs are actually a common-sense notion in test development … Specs are often called blueprints, and this is an apt analogy. Blueprints are used to build structures, and from blueprints many equivalent structures can be erected" (p. 52). Moreover, the table of specifications was used to outline the reading test before the test took place to see the skills, methods, and items that are tested within the time limit (Brown & Abeywickrama, 2010).

**Table 2: Table of specifications**

| Item no. | Reading cognitive process | Type of reading | Explicit/ implicit | Source | Test method |
|---|---|---|---|---|---|
| 1 | EPM | (CR/G)* | Explicit | Newspaper article - | MCQ 4-option |
| 2 | EPM | (CR/G) | Implicit | | MCQ |
| 3 | EPM | (CR/G) | Implicit | | MCQ |
| 4 | EPM | (CR/G) | Explicit | | MCQ |
| 5 | EPM | (CR/G) | Implicit | | MCQ |
| 6 | EPM | (CR/G) | Explicit | | MCQ |
| 7 | BMM | (CR/G) | Implicit | | MCQ |
| 8 | EPM | (CR/G) | Explicit | | MCQ |
| 9 | LA | (ER/L)** | Explicit | | MCQ |
| 10 | LA | (ER/L) | Explicit | | MCQ |
| 11 | WR | (ER/G) | Explicit | | MCQ |
| 12 | EPM | (ER/G) | Explicit | | MCQ |
| 13 | BMM | (ER/G) | Explicit | | MCQ |
| 14 | BMM | (CR/G) | Implicit | Travel blog | MM*** |
| 15 | BMM | (CR/G) | Implicit | | MM |
| 16 | EPM | (CR/G) | Explicit | | MM |
| 17 | I | (CR/G) | Implicit | | MM |
| 18 | I | (CR/G) | Implicit | | MM |
| 19 | EPM | (CR/G) | Implicit | | MM |
| 20 | EPM | (CR/G) | Explicit | | MM |
| 21 | EPM | (CR/G) | Explicit | | MCQ |

| | | | | |
|---|---|---|---|---|
| 22 | EPM | (CR/G) | Implicit | MCQ |
| 23 | EPM | (CR/G) | Implicit | MCQ |
| 24 | I | (CR/G) | Implicit | MCQ |
| 25 | BMM | (CR/G) | Explicit | MCQ |

*Careful reading/global
**Expeditious reading/local
***Multiple matching

Table 2 provides the table of specifications of the test assessed in this study. The table shows that the reading test comprises the eight cognitive processes of Khalifa and Weir (2009), the item number, cognitive processing, information about the item, whether it is explicit or implicit, the source, and the test method used.

The test construction was guided by the Common European Framework of Reference for Languages (CEFR), which is commonly used in language tests and policies throughout Europe (Deygers et al., 2018). The CEFR has six levels of scales, namely A1, A2, B1, B2, C1, and C2, from lower levels to higher levels. The first two levels (A1 and A2) are for basic users, whereas the next two are for independent users, and the final two for professional users (Council of Europe, 2001). For English as a foreign or second language users, the minimum requirement for university entrance and future academic achievements in the university career is the B2 level (Carlsen, 2018; Fleckenstein et al., 2018; Waluyo, 2019). The present study targeted the B2 level; hence, the design of the test focused on this level.

The test includes two passages that meet the readability analysis requirements for being qualified for the CEFR B2 level. The English language learning website Linguapress (2020) mapped Flesch-Kincaid reading scores onto CEFR levels (Natova, 2019). According to this website, the ranges of Flesch-Kincaid reading ease values between 60 and 70 are synthesized to be at the CEFR B2 level. Table 3 presents a summary of the readability analysis of the two passages selected based on the Flesch-Kincaid reading ease level. Text Inspector software was utilized to obtain the readability indices. The Flesch-Kincaid readability formula indicates that the higher the index, the easier the text.

**Table 3: Readability index**

| | Passage 1 | Passage 2 |
|---|---|---|
| Flesch reading ease | 63.43 | 67.47 |
| Flesch-Kincaid grade | 8.17 | 8.15 |
| Gunning fog index | 10.14 | 11.44 |

The Flesch reading ease score for passages 1 and 2 were 63.43 and 67.47, respectively. This suggests that the passages can be aligned with the CEFR B2 level.

Since the passages selected for this test were taken from the CEFR B2 level, most items were items on the level of HOT skills. Three items measured I (inferencing), five items evaluated BMM, and none of the items tested the most challenging HOT skills, such as CTLS and CITR, as discussed in the research conducted by the First Certificate in English (FCE) test (B2 level) of Cambridge assessment (Khalifa & Weir, 2009).

The questions were developed focusing on the cognitive processing in reading and were later validated by subject matter experts. We made sure that the test items were targeted at the B2 level. Two types of test methods (item formats), MCQ and multiple matching, were used, with 18 and 7 items, respectively.

Passage 1 used reading for information text belonging to the expository text. The passage was adapted from the www.sundayobserver.lk website. The first part of the article *What happens when you do nothing?* was adapted concerning test administration purposes and comprised seven MCQ four-option items and six fill-in-the-blank items of the selected-response type.

Passage 2 used reading for pleasure belonging to the narrative text. The passage was taken from a travel blog available at https://atasteoftravelblog.com/my-favourite-cities-in-the-world/. These items comprised seven multiple matchings of the selected-response method and five MCQ items.

Thus, a distribution of 25 items was tested using MCQ and multiple matching response methods. Responses to these items were used to examine the level of difficulty, the discrimination index for revision, distractor analysis to identify the malfunction distractor, and the reliability of the test.

### 3.2 Research Participants

This reading comprehension test was completed by 50 Faculty of Arts and Culture (FAC) students in the second and third year (semester 2, 2020/2021 academic year) from mixed-ability groups. Compared to science, technology, engineering, and mathematics (STEM) students, students from this faculty (FAC) are believed to be low achievers in English language proficiency, which hinders their enrolment in the job market (Dundar et al., 2017). Therefore, selecting students from a faculty of social sciences and humanities was the better choice. The participating students belonged to a multi-ethnic and multi-regional society whose mother tongue is Tamil. Out of the 50 participants, 44 were female, whereas 6 were male (88% and 12%, respectively). In addition, 30 participants were Muslim, 18 were Hindu, and 2 were Christian. Participants were provided 40 minutes to complete the task.

## 4. Results

The research findings are discussed in detail under the relevant subheadings below.

### 4.1 Item Difficulty Indices

This section describes the level of difficulty for each item to find out which cognitive processes are considered as easy, moderate, and difficult based on participants' responses. Table 4 provides a summary of the findings.

**Table 4: Item difficulty indices**

| Difficulty level | Item no. | Cognitive processing | Item difficulty | Number of items | % |
|---|---|---|---|---|---|
| Difficult (0.0–0.29) | Q19 | EPM | 0.28 | 1 | 4 |
| Moderate (0.30–0.79) | Q22 | EPM | 0.3 | 17 | 68 |
| | Q11 | WR | 0.38 | | |
| | Q1 | EPM | 0.42 | | |
| | Q24 | I | 0.42 | | |
| | Q9 | LA | 0.44 | | |
| | Q17 | I | 0.56 | | |
| | Q20 | EPM | 0.56 | | |
| | Q13 | BMM | 0.58 | | |
| | Q23 | EPM | 0.58 | | |
| | Q2 | EPM | 0.64 | | |
| | Q12 | EPM | 0.64 | | |
| | Q14 | BMM | 0.64 | | |
| | Q21 | EPM | 0.64 | | |
| | Q8 | EPM | 0.68 | | |
| | Q15 | BMM | 0.68 | | |
| | Q3 | EPM | 0.76 | | |
| | Q5 | EPM | 0.76 | | |
| Easy (0.80–1.00) | Q18 | I | 0.82 | 7 | 28 |
| | Q25 | BMM | 0.82 | | |
| | Q7 | BMM | 0.84 | | |
| | Q6 | EPM | 0.88 | | |
| | Q16 | EPM | 0.94 | | |
| | Q10 | LA | 0.96 | | |
| | Q4 | EPM | 0.98 | | |

In terms of level of difficulty in percentage, 4% of the items were difficult, 68% were moderate, and 28% were easy. The mean difficulty of the 25 items was 0.65. Of the 25 items, only one (Q19) was identified as a difficult item. This item belonged to the EPM cognitive process, which is the most difficult skill among the LOT skills. Although some items in the test evaluate other difficult cognitive processes such as I (inferencing) and BMM, item 19 was still found to be at the most difficult level. A possible reason for this could be due to the test format used. For this item, the multiple matching formats with an excessive option may make the item challenging.

Out of the 17 moderate items, 12 belonged to the LOT skills by Khalifa and Weir (2009), including WR, LA, and EPM. Of the 5 remaining items, 2 belonged to I (inferencing) and 3 to BMM.

The seven easy items had indices ranging between 0.8 and 1.0, indicating that these were easy for the participants to attempt. Easy items are made up of different cognitive processes, namely I (inferencing), BMM, EPM, and LA. Four of the items evaluated LOT skills, which involve explicitly stated information, whereas three evaluated HOT items.

The items in the test covered all suitable subskills needed for the B2 level, containing both explicit and implicit items. Nonetheless, as the mean difficulty of the 25 items was 0.65, it can be concluded that the reading test was moderately difficult for the participating group of test-takers.

### 4.2 Item Discrimination Indices

Table 5 presents the results on the discrimination indices for each test item based on four categories, namely poor item, fair item, good item, and very good item. Item discrimination provides information about how good the item is in distinguishing the strong students from the weak ones.

Since 50 students participated in taking the test, 27% out of 50 is equal to 14 participants. We therefore had to take the number of correct answers from the top 14 students (H), deduct the number of correct answers from the bottom 14 students (L), and then divide it by 14.

**Table 5: Item discrimination indices**

| Z | Item no. | Cognitive processing | Discrimination | Items (n) | % |
|---|---|---|---|---|---|
| Poor item (0.09–0.19) | Q4 | EPM | 0 | 3 | 12 |
| | Q16 | EPM | 0.143 | | |
| | Q10 | LA | 0.143 | | |
| Fair item (0.20–0.29) | Q1 | EPM | 0.214 | 3 | 12 |
| | Q9 | LA | 0.286 | | |
| | Q6 | EPM | 0.286 | | |
| Good item (0.30–0.39) | Q2 | EPM | 0.357 | 3 | 12 |
| | Q18 | I | 0.357 | | |
| | Q25 | BMM | 0.357 | | |
| Very good item (0.40–0.99) | Q19 | EPM | 0.429 | 16 | 64 |
| | Q22 | EPM | 0.429 | | |
| | Q14 | BMM | 0.429 | | |
| | Q23 | EPM | 0.500 | | |
| | Q3 | EPM | 0.500 | | |
| | Q5 | EPM | 0.500 | | |
| | Q7 | BMM | 0.500 | | |
| | Q20 | EPM | 0.571 | | |
| | Q13 | BMM | 0.571 | | |
| | Q12 | EPM | 0.571 | | |
| | Q21 | EPM | 0.571 | | |
| | Q17 | I | 0.643 | | |
| | Q11 | WR | 0.714 | | |
| | Q24 | I | 0.714 | | |
| | Q8 | EPM | 0.714 | | |
| | Q15 | BMM | 0.786 | | |

As seen in Table 5, 16 of the 25 items were identified as very good that are functioning well, amounting to 64% of the test items. The remaining nine items were reported equally (three each) as good, fair, and poor items, at 12% each. From this statistic, more than 88% of the items can be recycled and reused for the next test paper. Moreover, the most used LOT skills (WR, EPM, I, and BMM) were also tested in the "very good item" category. If the test items are going to be reproduced in future tests, the three items that fell under the "poor item" category with the low indices of 0 and 0.143, which are in the range of 0.09–0.19, should be revisited and eliminated if there is no clear justification to place them in the test. This is especially true for item Q4, where 49 participants selected the key (D) and none selected the distractors A and B. This item can be removed in future or else the distractors should be modified, and the key should not be direct. However, for items Q16 and Q10, slight changes can be made to the distractors to recycle them for reuse.

### 4.3 Distractor Analysis

This section analyses the malfunction distractors based on the participants' answers. Table 6 shows the 18 MCQ items each with their four options. The green highlighted boxes are the key answers, while the yellow highlighted boxes indicate the malfunction distractors.

**Table 6: Distractor analysis for MCQ items**

| | MCQ items | | | |
|---|---|---|---|---|
| **Item no.** | **A** | **B** | **C** | **D** |
| Q1 | 17 | 21 | 5 | 7 |
| Q2 | 6 | 32 | 10 | 2 |
| Q3 | 6 | 38 | 3 | 3 |
| Q4 | 0 | 0 | 1 | 49 |
| Q5 | 6 | 3 | 38 | 3 |
| Q6 | 3 | 1 | 2 | 44 |
| Q7 | 42 | 1 | 3 | 4 |
| Q8 | 9 | 5 | 34 | 2 |
| Q9 | 15 | 22 | 5 | 8 |
| Q10 | 0 | 48 | 1 | 1 |
| Q11 | 10 | 9 | 19 | 12 |
| Q12 | 8 | 4 | 6 | 32 |
| Q13 | 29 | 6 | 5 | 10 |
| Q21 | 0 | 32 | 10 | 8 |
| Q22 | 15 | 10 | 12 | 13 |
| Q23 | 0 | 29 | 8 | 13 |
| Q24 | 7 | 9 | 21 | 13 |
| Q25 | 2 | 4 | 3 | 41 |

Table 7 indicates the seven multiple matching items, each consisting of one correct answer and four distractors.

Table 7: Distractor analysis for multiple matching items

| | Multiple matching items | | | | |
|---|---|---|---|---|---|
| Item no. | Paris | New York | Istanbul | All three cities | None of the cities |
| Q14 | 7 | 32 | 6 | 2 | 3 |
| Q15 | 3 | 8 | 4 | 1 | 34 |
| Q16 | 1 | 2 | 47 | 0 | 0 |
| Q17 | 12 | 28 | 4 | 4 | 2 |
| Q18 | 4 | 41 | 3 | 0 | 2 |
| Q19 | 14 | 15 | 6 | 10 | 5 |
| Q20 | 11 | 8 | 28 | 2 | 1 |

The tables above clearly portray that items Q4, Q10, Q16, Q18, Q21, and Q23 contained malfunction distractors (distractors not chosen by any student). Item 4 had options A and B as malfunction distractors. Item 16 also had two malfunction distractors. Therefore, these items should be eliminated for future reproduction. However, all the other items can be kept for reproduction, with changes to some of the distractors. The pie chart below indicates the percentages of the options for item Q4, which contained two malfunction distractors (Figure 1).

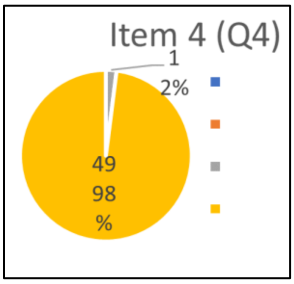

Figure. 1: Distractor analysis – item 4

Item 4 was answered correctly by 49 participants, possessing 98%, with option 4 or D being the key to the item. Only one participant answered wrongly, selecting option 3 or C. Options 1 (A) and 2 (B) were not selected by participants. In contrast, item 12 is a good example of an item with suitable distractors and a key (Figure 2).

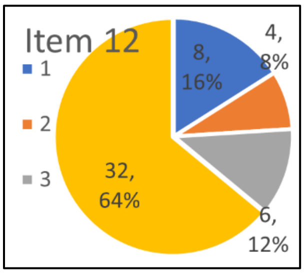

Figure. 2: Distractor analysis – item 12

As can be seen from Figure 2, the distractors and key of item 12 were selected in reliable proportions, which indicates that the distractors all functioned well. Each distractor and key of item 12 was selected by a good number of participants. Option 4 or D was the correct answer to item 12 and the rest were the distractors, with a distractor selection ratio of 8:4:6.

### 4.4 Reliability of the Test

The reliability index for the test analyzed in the current research was identified using the KR-20 formula (Table 8). To calculate the reliability index, first, the probability of the correct answers (P) for each item was calculated and then the probability of the wrong answers (Q) was identified. Thereafter, the sum of PQ was derived. Finally, the variance of the total scores (σ2) was calculated to determine the reliability of the test.

Table 8 shows the reliability index of the test to be at 0.82, which is a good value for a cognitive test (Samad, 2004; Tamil, 2015). The mean of the item difficulty indicates a good value of 64.8%. The standard error of measurement (SEM) of 2.006 is also an acceptable index (Tamil, 2015).

**Table 8: Reliability index**

| Item no. | Reading skill | Difficulty (%) | Discrimination | P | Q | PQ |
|---|---|---|---|---|---|---|
| Q1 | EPM | 42 | 0.214 | 0.42 | 0.58 | 0.244 |
| Q2 | EPM | 64 | 0.357 | 0.64 | 0.36 | 0.23 |
| Q3 | EPM | 76 | 0.500 | 0.76 | 0.24 | 0.182 |
| Q4 | EPM | 98 | 0.000 | 0.98 | 0.02 | 0.02 |
| Q5 | EPM | 7676 | 0.500 | 0.76 | 0.24 | 0.182 |
| Q6 | EPM | 88 | 0.286 | 0.88 | 0.12 | 0.106 |
| Q7 | BMM | 84 | 0.500 | 0.84 | 0.16 | 0.134 |
| Q8 | EPM | 68 | 0.714 | 0.68 | 0.32 | 0.218 |
| Q9 | LA | 44 | 0.286 | 0.44 | 0.56 | 0.246 |
| Q10 | LA | 96 | 0.143 | 0.96 | 0.04 | 0.038 |
| Q11 | WR | 38 | 0.714 | 0.38 | 0.62 | 0.236 |
| Q12 | EPM | 64 | 0.571 | 0.64 | 0.36 | 0.23 |
| Q13 | BMM | 58 | 0.571 | 0.58 | 0.42 | 0.244 |
| Q14 | BMM | 64 | 0.429 | 0.64 | 0.36 | 0.23 |
| Q15 | BMM | 68 | 0.786 | 0.68 | 0.32 | 0.218 |
| Q16 | EPM | 94 | 0.143 | 0.94 | 0.06 | 0.056 |
| Q17 | I | 56 | 0.643 | 0.56 | 0.44 | 0.246 |
| Q18 | I | 82 | 0.357 | 0.82 | 0.18 | 0.148 |
| Q19 | EPM | 28 | 0.429 | 0.28 | 0.72 | 0.202 |
| Q20 | EPM | 56 | 0.571 | 0.56 | 0.44 | 0.246 |
| Q21 | EPM | 64 | 0.571 | 0.64 | 0.36 | 0.23 |
| Q22 | EPM | 30 | 0.429 | 0.3 | 0.7 | 0.21 |
| Q23 | EPM | 58 | 0.500 | 0.58 | 0.42 | 0.244 |
| Q24 | I | 42 | 0.714 | 0.42 | 0.58 | 0.244 |
| Q25 | BMM | 82 | 0.357 | 0.82 | 0.18 | 0.148 |
|  | **Mean** | 64.8 | 0.451 | 0.65 | 0.35 | 0.189 |
|  | **Total** |  |  |  |  | 4.732 |
|  | **Variance** |  |  |  |  | 21.71 |
|  | **pKR20** |  |  |  |  | **0.815** |
|  | **SEM** |  |  |  |  | **2.006** |

This paper illustrates the utility of the CTT item analysis in providing *post-priori* information about test items in terms of difficulty, discrimination function, and the distractors' function. The results showed that many of the items are moderately difficult items. In fact, the test on average is moderately difficult, with a mean item difficulty of 0.65.

In terms of what made up difficult and easy items in the test, we found that easy items are not necessarily made up of LOT skills and easy cognitive processes, a finding corroborated by Khalifa and Weir (2009). For example, EPM was found to be both easy and difficult in this test, where it is theoretically considered an LOT skill. Similarly, according to Khalifa and Weir (2009), although WR is the easiest and CITR the most difficult skill, the results of this research showed the contrary. Therefore, the present study cannot confirm the hierarchy of cognitive processes as stipulated in literature.

The discrimination analyses conducted in this study yielded percentages for items regarding being considered as very good, good, fair, or poor items. This is important in situations where the items are to be reused. For example, items Q4, Q10, and Q16 were poorly functioning items with low item discrimination indices between 0.09 and 0.19 and therefore need to be reworked before they can be reused. The same can be said about the utility of the distractor analysis. Some items were found to have options that were not plausible for any participating test taker.

All these findings provide support that developing test items is not an easy task. Test designers need to consider a few aspects when designing tests. These include the purpose and objective of the test; how the test specifications will reflect both the purpose and objectives; selection of test tasks; arrangement of the separate items; and what kind of scoring, grading, or feedback is expected. Furthermore, creating a question paper is challenging and time-consuming. The designer has to start with the text input, create the stem and options, and consider the length, vocabulary range of the students, and formats. Although MCQ items are believed to be "easy", designing them is "time-consuming" (Powell & Gillespie, 1990, p. 1) and it is difficult to come up with plausible and alternative distractors.

## 5. Conclusion

This paper has illustrated the efficacy of item difficulty index, item discrimination index, and distractor analysis in identifying the quality of items used in assessments. In addition, it provided ample data on the reliability of the whole test analyzed using the reliability index. Examining the item analysis and student performance can help to improve the course and curriculum as well as shape teachers' professional development. This study has an implication for test item writing. Test writers need to consider the facets that can contribute to the difficulty of an item besides the cognitive processes and thinking levels. Further research is needed and should include more items, examine a larger sample size, and focus on the constructed responses. In addition, employing item response theory (IRT) instead of CTT may provide different findings.

## 6. Acknowledgement

We wish to express our gratitude to the AHEAD project funded by the World Bank in collaboration with the UGC, Sri Lanka and the Ministry of Higher Education, Sri Lanka for enabling the first author to pursue her PhD program.

## 7. References

Alderson, J. C. (2000). *Assessing reading*. Cambridge Assessment English.

Bax, S., & Chan, S. H. C. (2016). Researching the cognitive validity of GEPT high-intermediate and advanced reading: An eye-tracking and stimulated recall study. *LTTC-GEPT Research Reports*, *7*, 1-47. www.lttc.ntu.edu.tw/lttc-gept-grants/RReport/RG07.pdf

Bichi, A. A., & Embong, R. (2018). Evaluating the quality of Islamic civilization and Asian civilizations examination questions. *Asian People Journal (APJ)*, *1*(1), 93-109. www.uniszajournals.com/apj

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (Vol. 10). Pearson Education.

Carlsen, C. H. (2018). The adequacy of the B2 level as university entrance requirement. *Language Assessment Quarterly*, *15*(1), 75-89. https://doi.org/10.1080/15434303.2017.1405962

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. https://rm.coe.int/1680459f97

Creswell, J. W. (2012). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (4th ed.). Pearson.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Eric.

Deygers, B., Zeidler, B., Vilcu, D., & Carlsen, C. H. (2018). One framework to unite them all? Use of the CEFR in European university entrance policies. *Language Assessment Quarterly*, *15*(1), 3-15. https://eric.ed.gov/?id=EJ1171980

Dundar, H., Millot, B., Riboud, M., Shojo, M., Goyal, S., & Raju, D. (2017). *Sri Lanka education sector assessment: Achievements, challenges, and policy options*. World Bank Group. https://doi.org/10.1596/978-1-4648-1052-7

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice-Hall.

Eleje, L. I., Onah, F. E., & Abanobi, C. C. (2018). Comparative study of classical test theory and item response theory using diagnostic quantitative economics skill test item analysis results. *European Journal of Educational and Social Sciences*, *3*(1), 57-75. https://www.researchgate.net/publication/343557487

Fleckenstein, J., Leucht, M., & Köller, O. (2018). Teachers' judgement accuracy concerning CEFR levels of prospective university students. *Language Assessment Quarterly*, *15*(1), 90-101. https://doi.org/10.1080/15434303.2017.1421956

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. Routledge.

Halek, M., Holle, D., & Bartholomeyczik, S. (2017). Development and evaluation of the content validity, practicability and feasibility of the Innovative Dementia-Oriented Assessment System for Challenging Behaviour in Residents with Dementia. *BMC Health Services Research*, *17*(1), 554. https://doi.org/10.1186/s12913-017-2469-8

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38-47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Kastner, M., & Stangl, B. (2011). Multiple choice and constructed response tests: Do test format and scoring matter? *Procedia – Social and Behavioral Sciences*, *12*, 263-273. https://doi.org/10.1016/j.sbspro.2011.02.035

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, *30*(1), 17-24. https://doi.org/10.1037/h0057123

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge University Press.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151-160. https://doi.org/10.1007/BF02288391

Linguapress. (2020). *A comparison of different readability scales*. https://linguapress.com/teachers/flesch-kincaid.htm

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment*, *1*(1), 1-11. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1426043

Manalu, D. (2019). An analysis of students reading final examination by using item analysis program on eleventh grade of SMA Negeri 8 Medan. *Journal of English Teaching & Applied Linguistics*, *1*(1), 13-19. http://repository.uhn.ac.id/handle/123456789/2796

McNamara, T. F. (1996). *Measuring second language performance*. Longman Publishing Group.

Messick, S. (1989). Validity. In R. L. Linn (ed.), *Educational measurement* (3rd ed.; pp. 13-104). MacMillan.

Natova, I. (2019). Estimating CEFR reading comprehension text complexity. *The Language Learning Journal*, *49*(6), 699-710. https://doi.org/https://doi.org/10.1080/09571736.2019.1665088

Powell, J. L., & Gillespie, C. (1990). *Assessment: All tests are not created equally.* https://files.eric.ed.gov/fulltext/ED328908.pdf

Pratiwi, R., Antini, S., & Walid, A. (2021). Analysis of item difficulty index for midterm examinations in junior high schools 5 Bengkulu City. *Asian Journal of Science Education*, *3*(1), 12-18. http://www.jurnal.unsyiah.ac.id/AJSE/article/view/18895

Samad, A. (2004). *Essentials of language testing for Malaysian teachers*. UPM Press.

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. https://files.eric.ed.gov/fulltext/EJ1272266.pdf

Tamil, A. M. (2015). Calculating difficulty, discrimination and reliability index/standard error of measurement. *PPUKM*. https://ppukmdotorg.wordpress.com/2015/04/02/calculating-omr-indexes/

Turner, R. C., & Carlson, L. (2003). Indexes of item-objective congruence for multidimensional items. *International Journal of Testing*, *3*(2), 163-171. https://doi.org/10.1207/s15327574ijt0302_5

Urquhart, A. H., & Weir, C. J. (1998). *Reading in a second language: Process, product and practice*. Longman.

Waluyo, B. (2019). Thai first-year university students' English proficiency on CEFR levels: A case study of Walailak University, Thailand. *The New English Teacher*, *13*(2), 51-71. http://www.assumptionjournal.au.edu/index.php/newEnglishTeacher/article/view/3651

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Mesa Press.

Yusup, R. B. (2012). *Item evaluation of the reading test of the Malaysian University English Test (MUET)* (Master's thesis). The University of Melbourne. http://hdl.handle.net/11343/37608

Zimmerman, D. W. (1972). Test reliability and the Kuder-Richardson formulas: Derivation from probability theory. *Educational and Psychological Measurement*, *32*(4), 939-954. https://doi.org/10.1177/001316447203200408

Zubairi, A. M., & Kassim, N. L. A. (2006). Classical and Rasch analyses of dichotomously scored reading comprehension test items. *Malaysian Journal of ELT Research*, *2*(1), 1-20. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.535.2955&rep=rep1&type=pdf