REVIEW ARTICLE
published: 29 February 2020
doi: https://doi.org/10.21070/pedagogia.v9i1.123

The Role of Classical Test Theory to Determine the Quality of Classroom Teaching Test Items

Peran Teori Uji Klasik untuk Menentukan Kualitas Item Tes Pengajaran di Kelas

Wong Vincent [1]***, S.Kanageswari Suppiah Shanmugam*** [2]

[1] *College of Arts and Sciences, Universiti Utara Malaysia, Malaysia,* [2] *College of Arts and Sciences, Universiti Utara Malaysia, Malaysia*

The purpose of this study is to describe the use of Classical Test Theory (CTT) to investigate the quality of test items in measuring students' English competence. This study adopts a research method with a mixed methods approach. The results show that most items are within acceptable range of both indexes, with the exception of items in synonyms. Items that focus on vocabulary are more challenging. What is surprising is that the short answer items have an excellent item difficulty level and item discrimination index. General results from data analysis of items also support the hypothesis that items that have an ideal item difficulty value between 0.4 and 0.6 will have the same ideal item discrimination value. This paper reports part of a larger study on the quality of individual test items and overall tests.

Keywords: Classical Test Theory, Test Item Quality

OPEN ACCESS
ISSN 2548 2254 (online)
ISSN 2089 3833 (print)

Edited by:
Machful Indra Kurniawan

Reviewed by:
Deni Adi Putra

****Correspondence:***
Wong Vincent
wongvinc90@gmail.com

Received: *16 November 2019*
Accepted: *20 January 2020*
Published: *29 February 2020*

Citation:
Vincent W and Shanmugam SS (2020) The Role of Classical Test Theory to Determine the Quality of Classroom Teaching Test Items. PEDAGOGIA: Jurnal Pendidikan. 9:1. doi: https://doi.org/10.21070/pedagogia.v9i1.123

Tujuan dari penelitian ini adalah untuk menggambarkan penggunaan Teori Uji Klasik (CTT) untuk menyelidiki kualitas item tes dalam mengukur kompetensi bahasa Inggris siswa. Penelitian ini mengadopsi metode penelitian dengan pendekatan metode campuran. Hasilnya menunjukkan bahwa sebagian besar item berada dalam jangkauan yang dapat diterima dari kedua indeks, dengan pengecualian item dalam sinonim. Item yang fokus pada kosa kata lebih menantang. Yang mengejutkan adalah bahwa item jawaban pendek memiliki tingkat kesulitan item yang sangat baik dan indeks diskriminasi item. Hasil umum dari analisis data item juga mendukung hipotesis bahwa item yang memiliki nilai kesulitan item ideal antara 0,4 dan 0,6 akan memiliki nilai diskriminasi item ideal yang sama. Makalah ini melaporkan bagian dari studi yang lebih besar pada kualitas item tes individual dan tes keseluruhan.

Keywords: Teory Test Klasik, Uji Kualitas Item

INTRODUCTION

Assessment is, without a doubt, a key aspect of education. It is used to provide valuable feedback on an individual's and the groups' understanding of syllabi acceptance of certain teaching methods, of which review and improvisation can be done to better effect Koçdar et al. (2016) . One of the most straightforward methods of obtaining such feedback is via testing.

Designing a test is a complicated and elaborate process. It involves a lot of aspects, first of which requires test items to address the contents of the subject matter as equally as possible with aid of the Table of Specification, to minute attention to the finishing touches such as instructions and the duration of the test as to not put a blemish on the credibility and reliability of the test. A good test will be able to provide quality feedback on the intended construct; and in order to determine whether the items used to build are of high quality, they must be analyzed in terms of their difficulty and how well they are able to distinguish or discriminate between the pupils Koçdar et al. (2016) .

Item analysis is an important step in any test building as it looks at "students' responses to individual test items in order to assess the quality of those items and the quality of the test as a whole" Pande et al. (2013) .It also provides a better representation of the characteristics of items used in a test Salkind (2010) . According to Bichi and Embong (2018) , item analysis looks at the performance of items in relation to other factors and items to better understand its characteristics and, if there are, identify its flaws. Item analysis will provide useful information on whether or how to improve the quality and accuracy of items.

Primary School Evaluation Test (UPSR)

The Malaysia's Primary School Evaluation Test, more commonly known as UPSR from its Bahasa Melayu abbreviation Ujian Penilaian Sekolah Rendah, is a summative assessment that primary school pupils have to sit for at the end of Year Six. It is an important benchmark for Year Six pupils nationwide as the results are, in the eyes of the society, accurate indicators of the pupil's academic performance and aptitude. The number of subjects tested and format differs between national and national-type schools as their syllabus and curriculum are slightly different, for example the pupils of the national schools have to sit for six papers while those studying in the latter have to take an extra two papers for Mandarin.

The focus here, however, is on the English paper, which are not too dissimilar in terms of layout and structure. There are two separate papers to be taken at two different time slots. Paper 1 consists of two sections: Section A contains 20 multiple-choice items on vocabulary, grammatical items, idioms or proverbs, synonyms/antonyms and a comprehension text with questions, whereas items on short social exchanges i.e. 'Thank you.' as a reply to 'Congratulations!' and true/false and short answer items based on linear and non-linear texts form the structure of Section B. The multiple-choice items are single-best types, which means that of the four options available to the candidate, there is only one key for each item, with the other three functioning as distractors. Section B allows a certain degree of freedom with the language; however, candidates still have to address the requirement stipulated by the stems.

As for Paper 2, there are three sections. Section A requires candidates to transfer information from a linear or non-linear text to another text, usually linear, correctly; Section B is further divided into two parts, the first part of which is a direct transfer of information based on the stem, while candidates are expected to read and understand the stem's instruction and create a short text of 50-80 words for the second part. The last part of the paper is note expansion based on short notes and graphics. Candidates are given the choice of a one-picture stimulus with words or a three-picture series with words to guide their writing. The scoring depends on the weightage attributed to each section, and examinees are awarded two separate grades for Paper 1 and Paper 2 respectively.

Research Objective

Due to the importance of the exam and its outcome to all parties involved, it is essential that careful measures are taken to ensure that the items used are of high quality and accurate. The purpose of this study is to explore the use of Classical Test Theory (CTT) to investigate the quality of test items in English Paper 1, which consists of multiple-choice and short answer items. The rationale behind this is to better understand which items are deemed easy or difficult from the pupils' perspective, with hopes that results of this test can better aid the design of for future English papers with a good balance between easy and challenging items that can offer a better barometer of the pupils' English competence. The aims of this study are to 1). investigate the role of Classical Test Theory the classroom testing; 2). determine the characteristics of different types of items designed for standardized tests using item analysis; 3). identify irregularities in the current tests setup with input from teacher experts

Classical Test Theory v. Item Response Theory

Classical test theory (CTT), also referred to as the "true score theory", operates based on the assumption that the differences between the responses of examines are systematic; they are affected by the variation in the ability of the examinees. The theory focuses its attention on only the ability of interest, and one of the biggest assumptions that often attract scrutiny of the results is that all other sources of variation, such as external factors of the surrounding or physical and mental conditions of the examinees are constant throughout repeated standardization procedure, or just random and unsystematic occurrence in Magno (2009) .

The model central to the theory are the three concepts: observed test scores (TO), which is the result of true score (T)

and error score (E) in Magno (2009) . True scores are the examinees' real score if there were no errors in measurement instruments; however, that is highly improbable as instruments are rarely perfect, thus the observed test scores for each individual is the outcome of the examinee's true ability influenced by error, either higher or lower. CTT also introduces the concept of standard error of measurement to account for how much the error has affected the reading on true scores in Magno (2009) ; the larger the standard error of measurement, the less accurate the measurement of the intended attribute, and vice versa.

Historically, principles of CTT have pioneered methods of analysis used to evaluate tests by looking at the four criteria: frequency of correct responses to indicate item difficulty, frequency to each of the responses to analyze distracters for their functionality; the correlation between the item and responses between higher and lower achieving groups of examinees in Magno (2009).

As is common with theories that have been around for some time, CTT is not without its detractors. It does have its own limitations, which mostly circle around its dependency on the test itself and the samples. Most of the results gained from the methods derived from the theory can only be attributed to the samples who are taking the test or that particular test and are unable to be generalized to other examinees or tests. For examples, the item difficulty index, *p* derived from a particular sample of examinees may change with a different sample taking the test, which is also the case with the item discrimination index, D and distractor analysis. The ability scores of examinees are also dependent on the test. Examinees' ability changes depending on different tests or the different occasions in which they take the test.

Hence, the item response theory (IRT) is created, partly to address the shortcomings of CTT. The theory is focuses on the chances of getting an item right or wrong based on the each items' characteristic curve, which looks at the probability of getting each item right or wrong in relation to the examinees' ability in Magno (2009) . It forms a boundary between the chances of getting an item correctand vice versa. The Rasch model is one of the many products of IRT, mainly used for dichotomous scoring under the assumption that the discrimination value is equals to one.

IRT has the upside of treating reliability and error of measurement through computation of the item information function in Magno (2009) . The item information function takes in account the parameters and displays the items efficiency for different ability levels. Another advantage of the IRT is its results do not depend on the types of samples in Magno (2009) . Its uniform scale of measurement can be relied upon to give accurate readings even for different samples. It also means that scores from different individuals tested with a different set of items that are appropriate to their ability can be compared Magno (2009).

Although IRT proves to be a significant step-up in terms of reliability and generalizability compared to CTT, it represents a more complicated method of analysis as a lot of factors come into play. Just by analyzing a one-dimensional parameter alone, users will have to look at the examinees' ability and the test's difficulty which is estimated by the total number of errors in said test, and it involves running the numbers through an algorithm, the process of which requires the help of digital computation. Although CTT has its own weaknesses, it is still widely used as it represents a more economical and practical way to generate statistics, especially when the assessment involved do not carry as much weight. Hence, in this study, the items will be analyzed using CTT as it was carried out in the school setting, and the objective is to do a quick analysis on the different types of items based on the pupils' impression.

Item Analysis

The key criteria to look at in this study are i) item difficulty index, *p*, ii) item discrimination index (D) and iii) distractor analysis. Four research papers were reviewed to lay down the groundwork for the three key criteria in order to provide a sound analysis for the current study. Item difficulty index, the *p*-value, represents how easy or difficult an item is based on the value ranging on 0.0 and 1.0 derived from pupils' correct responses Bichi and Embong (2018) . The higher the *p*-value, the easier the items are and vice versa.

Item discrimination index, D, as defined by Mehta and Mokhasi (2014) , measures how an item is able to discriminate the more able pupils from the less able ones, with a +1 index meaning the item is very effective, whereas 0 shows that the item is unable to discriminate at all. In the rare cases that the discrimination index is -1, it is an indication that more pupils from the group with lower overall score are selecting the key responses more frequently than pupils who perform better Bichi and Embong (2018) .

A multiple-choice item contains a stem and four options, which contains a key and three distractors. Distractor analysis looks at how effective are the distractors in affecting the pupils' judgement in identifying the key (Mehta and Mokhasi, 2014) Mehta and Mokhasi (2014) . The general interpretation of a functioning distractor is when the distractor is selected by 5% or more pupils. If a distractor is not working, it is classified as a Non-Functioning Distractor (NFD) and are revised, removed, or replaced with better options. The three research papers from Mukherjee and K (2015), Bichi and Embong (2018) and Mehta and Mokhasi (2014) were reviewed to derive the most suitable interpretation of *p*-value, D and Distractor Analysis for this study.

Mukherjee and K (2015) looks at elements of multiple-choice questions as an efficient form of tests in health sciences. The study proposes that items with *p*-value between 0.2 – 0.9 are considered good items, with those which value located between 0.4 – 0.6 further classified as excellent items. If an item is valued at less than 0.2 or above 0.9, they are not acceptable and require modification because they are too difficult or too easy respectively. An interesting claim made by Mukherjee and K (2015) is that items valued between 0.4 and 0.6 also have

maximum discrimination index. As for D, 0.40 and above are considered excellent items; items with D from 0.30 to 0.39 are reasonably good, whereas 0.20 to 0.29 would put them as items needing to be reviewed. Value of 0.19 and below would rank them as poor items with the possibility of rejection. Mukherjee and K (2015) further suggests count of 5% or more pupils is needed for a distractor to be deemed effective.

Another literature reviewed was Bichi and Embong (2018) whose study evaluates the quality of Islamic Civilization and Asian Civilizations Examination Questions. Bichi and Embong (2018) recommends "values of difficulty no less than 30% correct and no greater than 70%". Items smaller than 0.3 and bigger than 0.7 in p-value are too difficult or too easy; and will consequently be weaker in ability in discriminate high scorers and low scorers. The Item discrimination index, derived from the works of Bichi and Embong (2018) classified items with values of 0.4 and above as 'very good'; 0.3 to 0.39 'reasonably good' but subject to improvement; items between 0.2 to 0.29 are usually subjected to revisions and items <0.19 is 'poor'. However, the study rates a distractor as being acceptable as long as it attracts at least one candidate.

Mehta and Mokhasi (2014) also advocates similar views to Bichi and Embong (2018) , rating items between p-value between 0.3 and 0.7 as acceptable, and further suggests that items between 0.5 and 0.6 are ideal. Items placed in the two extremes ($p < 0.3$ and $p > 0.7$) are in need of modification as they are not acceptable as they are. In terms of D, items with index more than 0.35 are considered as excellent; D-value between 0.2 and 0.35 is 'good' and those with index less than 0.2 are 'poor' items. As for distractors, Mehta and Mokhasi (2014) deems a distractor as effective if it is selected by 5% or more pupils.

Content Validity

As with all kinds of assessment, a test will not be seen as an effective form of measurement without content validity. While experts have various opinions and views of what constitutes as content validity, this study will proceed with the definition that content validity refers to how well the items of the assessment cover the content, knowledge or skills that it claims to cover in Fitzpatrick (1983) . Fitzpatrick (1983) further illustrates this point when he states that an achievement test has to reflect the content domain outlined in a test manual. In short, based on this definition, tests have content validity when 'they test what they are meant to test".

To ensure that the test for this study achieve the standard of content validity, the items are compared to a Table of Specification based on the Curriculum Standard of Year 4, Year 5 and Year 6 from which the items for UPSR will be based upon. The Curriculum Standard is a document in which the underlying pedagogical principles of the English curriculum, the Content Standard, the Learning Standard are described in detail as it is a document that serves as a guideline to educators nationwide as we move to a more skill-based approach to language learning in Pendidikan (2013) .

The Table of Specification was done by looking at the skills that are able to be tested in a written exam, mainly Reading, Writing and Grammar, and their learning standards and comparing them to the revised Bloom's taxonomy. Content and learning standards for Listening and Speaking and Language Arts were omitted as both of those skills do not appear explicitly in their separate sections in UPSR.

Test Administration Procedure

For further field study, a few modifications can be made to make the whole process more effective, thus enhancing the reliability of the data collected: 1) have a special room or hall to emulate actual standardized testing scenario which they are familiar with Cook and Beckman (2006) , with each pupil having a table and seat of their own and spacing between each other to prevent cheating and comfort purposes, 2) while no time limit is advisable, participants need to be on task at all times to ensure validity and reliability at all times, 3) as suggested by Cook and Beckman (2006) , tests should be administered in a way as closely resembled to a real standardized test as possible, with as little indication that the whole process is more than a test. This can be done by mimicking the actual procedure of standardized tests, and the structure of the test as well by adding more items to the test designed for study.

METHOD

Test Items

The test items in this paper were chosen at random either from past years' UPSR papers or items that closely resemble the items in the actual paper from a revision workbook that publishes past year's UPSR papers and model papers for the same examination. No changes were done to the stem or the options during the process of lifting to preserve the authenticity of the items. A total of seven test items were chosen for this study. Five multiple choice items with different test focus (vocabulary, tense, idiom, synonym and spelling) form the objective section of single-best response items while two short answer items from a comprehension text make up the subjective part in Paper 1. The items were printed on a single sheet of A4 paper, with five multiple-choice items on one side and the comprehension text with its short answer items on the other.

A few constructs were represented in the test. Multiple choice items 1 and 4 tested pupils on Construct 2.2.1 Able to apply word attack skills by using contextual clues to get meaning of words using items on vocabulary and synonyms respectively. Multiple choice Item 2, which focused on different forms of verb, drew on Construct 5.1.3 Able to use verbs correctly and appropriately. Construct 2.2.2 was represented by Multiple choice Item 3 that tested on idioms. Multiple choice Item 5 which tested on their knowledge of the spelling of the word 'queue' was from the Construct 3.2.4 Able to spell

words by applying spelling rules. Subjective items were also designed based on the learning standards in the Table of Specification. Subjective Item 6 tested pupils on the ability to read and demonstrate understanding of texts by drawing conclusions with guidance, as stipulated by Construct 2.2.3. Subjective Item 7 represented the skill in Construct 3.3.1 Able to create simple linear texts using a variety of media with guidance.

Test Samples

The participants of this test were Year 6 pupils currently studying in a national-type school in Kuala Terengganu. A total of 30 pupils were chosen, with 22 pupils coming from the 4^{th} class out of seven and a further eight (n=8) were chosen at random from the 3^{rd} class. Since the school practices non-streaming except for the first class and the 4^{th} class is the middle in terms of ranking among the classes, it is a safe assumption that there was a nice range of pupils with different levels of mastery of the language ranging from poor to competent. Most of their input in English is only during English lessons, with the language being used minimally outside the classroom and only at the vocabulary level.

The test was administered in the classroom where the 4^{th} class has their lessons during one of their English periods. The other eight (n=8) participants were granted special leave from their own class to participate in the test.

Test Administration

The mixed method approach was employed to gather the necessary data. Pupils participating in the test were gathered in one of the Year 6 classrooms for the test.

Before they began, the pupils were given a simple briefing on the structure of the test. No time limit was imposed on them to complete the test; however, they were told not to delay unnecessarily as it was during school hours and seeing that some of them were seated with their peers (due to insufficient seating), preventive measures beforehand in the form of reminders were deemed necessary in order to cut down on the influence of outside factors such as distraction and time wasting, thus affecting the reliability of the test. Pupils were also reminded that the results of the test will not be reflected in their academic achievement, so they should just try their best in answering them without too much pressure. The whole testing process took around 15 minutes.

Scoring

The scoring for each item is different depending on the type of answer it elicits from the pupils. For the multiple-choice items, there are four options, with only one of them the key and the other three distractors. Scoring for the multiple-choice items is dichotomous, which means being able to choose the key would earn the pupil one mark, whereas selecting one of the other three distractors would result in zero mark.

The subjective section consists of short answer items that are polytomous in scoring. The items have scores ranging from zero (inaccurate response or no response), to one (response partially correct or contains grammatical errors) and finally two marks, which is the maximum a pupil can get for accurate and coherent response.

The key for the five objective items are C, A, C, D and A respectively. As for the two subjective items, scoring was done based a rubric adapted from the original rubric designed for the specific section in standardized tests. The general outline of the rubric has already been explained in the previous paragraph.

Data Analysis

This study utilizes the *p* and D values as advocated by Bichi and Embong (2018) . The acceptable item difficulty index, *p*, is between 0.3 and 0.7, with values below 0.3 rated as too difficult and above 0.7 being too easy. Items that are valued in the region of these two extremes are treated as poor items and should either be modified or removed. The *p* value is calculated using responses from all the pupils taking the test. As for discrimination item, D, items with value above 0.4 are considered excellent in terms of discriminating high and low achievers. Items with D-values between 0.3 and 0.39 are considered reasonably good but can still be improved, whereas items with D-values ranging from 0.2 to 0.29 are only marginally acceptable and should be modified in order to be included as an item. Poor items are items with values 0.19 and below and should be rejected. If the D-value of an item is in the negative, it means that more pupils in the lower achieving group are able to locate the key than pupils in the higher achieving group, which is highly illogical. These items should be investigated further to understand the reasons behind such irregularity. Due to the small sample size, Bichi and Embong (2018) interpretation of at least one (n=1) pupil to be sufficient as a functioning distractor is used, as 5% of n=30 is 1.5, which would be one or two pupils as well. However, comparison will be made with the other literatures when necessary.

Each data gathered was run through the formula to determine the *p* and D of each item. The formula for item difficulty index, *p* for objective items is $p = \frac{C}{N}$. *C* refers to the number of pupils who answered the item correctly and *N* is the total number of pupils taking the test. As for subjective items, the formula for *p* is $p = \frac{\sim fx - nX_{min}}{n\,(x_{max} - X_{min})}$ The symbol *~fx* is used to represent the total number of points earned by all the pupils on the item; *n* refers to the total number of pupils (n=30); *X max* is the maximum points available for the item, which is two marks, while *X min* is the minimum points that the pupil can get, which in this case is zero. Item discrimination index, D, is calculated using the formula $D = \frac{U_p - L_p}{U}$.*Up* is the number of pupils answering the item correctly in the high achieving group while *Lp* represents the number of pupils in the low achieving group who answered the item correctly.*U* symbolizes the number of high performers, which was nine (n=9) in this study.

Teacher Expert

Three teacher experts were chosen to address the irregularities in the data gathered. In order to gather high-quality responses that give accurate reflections on the issues, the teacher experts selected fulfilled the criteria of: 1) having taught at the same school that the pupils who were administered the test studied in for more than ten years, 2) familiar with the standardized summative assessments of which the test was based on, and 3) familiar with the Year 6 pupils in general. The three teachers were approached individually and the issues were discussed in a casual manner. Their responses were recorded with pen and paper.

RESULTS AND DISCUSSION

Results

Each item in the test was analyzed for its quality using Distractor Analysis, item difficulty index, *p*, and item discrimination, D. The Distractor Analysis will present the data on the pupils' responses to all the options of each item in order to identify how effective the distractors are, followed by calculations of *p* using the data from Distractor Analysis to identify the difficulty index for each item. Lastly, each items D-value will be calculated to determine its efficiency in distinguishing the more able and less able pupils in relation to the construct each item is based on. The workings for each item are presented below along with its interpretation and analysis.

[Table 1 about here.]

[Table 2 about here.]

$$p = \frac{C}{N} = 0,2 \tag{1}$$

The**Table 1** ,**Table 2** and calculation in **Equation** (**1**) above show the difficulty, *p* of Item 1. Based on the table, we can see that 50% of the class chose the distractor option A, and only 20% of the group managed to identify the key. This is reflected the *p*-value of 0.2, which ranks the item as 'difficult', and needs to some improvement in order to be a better item suited for testing.

[Table 3 about here.]

$$D = \frac{U_p - L_p}{U} = 0,44 \tag{2}$$

The discrimination index, D is valued at 0.44, which is above the widely accepted cut-off point of >0.4 to be sufficient to be regarded as an excellent item in terms effectiveness in discriminating between high and low scores. Based on the predetermined values and the data collected, Item 1 cannot be entirely regarded as a quality item. Although it achieved standard as an item with excellent D, it scores poorly in its *p*-value, bordering on 'needs to be rejected or modified entirely'. Distractor Analysis reveals a vital piece of information – 50% of the class actually went for D, which contains the distractor 'station' to be paired with the subject 'plane' and the verb 'towed'. Instead of saying that the distractor was functioning well, it seems to be that the key 'hangar' was too unfamiliar to most of the pupils for them to choose that as their answer. In such cases, the key or stem should be modified to use vocabulary that are of high frequency or familiar to the pupils.

[Table 4 about here.]

[Table 5 about here.]

$$p = \frac{C}{N} = 0,8 \tag{3}$$

Item 2 in **Equation** (**3**) has a very high *p*-value of 0.8, which means that it is too easy; while it is still acceptable by Mukherjee and K (2015) estimation, it has already lost its potency as an 'ideal' item, which should be between $0.3 < p < 0.7$ Bichi and Embong (2018) . The table clearly indicates that more than three quarters of the pupils managed to locate the key, with the distractor D completely ignored by all of them.

[Table 6 about here.]

$$D = \frac{U_p - L_p}{U} = 0,44 \tag{4}$$

Item 2 functions as well as Item 1 in terms of their discrimination index, D, which is also 0.44 **Equation** (**4**) . Again, it shows that the item is doing a more than able job in distinguishing the high achievers from the low achievers, as shown by the table where all of the pupils in the upper group chose the key, while only 5 from lower group were able to do so.

The data collected shows Item 2, which is a grammar item testing on the infinitive to-, is an easy item, as reflected by the high *p*-value (*p=80)*. The data from the Distractor Analysis also supports this inference, as only 6 pupils answered with answers other than the key, and D 'completing' was a non-functioning distractor. However, despite not having the ideal *p*-value, it still scores quite high in D, showing its potential as an item that is able to distinguish the high and low achievers. A possible way to make the item better is to replace NFD with another plausible distractor that will attract some of the pupils who chose the key to lower its *p* to between 0.3 and 0.7.

The teacher experts' opinions were sought for this item as it has two NFDs, *completes* and *completing*. The teacher experts all agreed that the reason that the option *completing* did not receive any responses was because the pupils have been taught

and conditioned to only go for verbs in their continuous forms only if they see the auxiliary verb *be*. As for *complete,* one explanation was it was due to presence of *needs* in the stem, which has taken away the possibility of the singular form, while another was due to the small sample size; one teacher expert claimed that with more pupils taking the test, some would go for *completes*.

[Table 7 about here.]

[Table 8 about here.]

$$p = \frac{C}{N} = 0,5 \tag{5}$$

The *p*-value of Item 3 is 0.5, which, according to a few other studies, is an ideal score in more ways than one. Not only does it rank as 'excellent' in terms of difficulty Mehta and Mokhasi (2014) , Mukherjee and K (2015) also reports that when items have *p*-value of $0.4 < p < 0.6$, their discrimination index is also high.

[Table 9 about here.]

$$D = \frac{U_p - L_p}{U} = 0,67 \tag{6}$$

As suggested by Mukherjee and K (2015) and proven by data, Item 3 scored the highest in D-value in this test. The table clearly shows a clear gap between the upper and lower group, with a difference of 6. A more remarkable outcome was that there was a nice spread across the four options for the lower group, whereas responses from the higher achieving group centered on the key, with only one choosing the distractor D.

Item 3 is one of the best items in the test, with an ideal *p*-value of 0.5 and D of 0.67, indicating that it is an average item in terms of difficulty and is able to create a division the pupils who are better from those who are poor. Item 3's efficiency is further illustrated as the responses from all the pupils were taken into account. While a higher number of responses to the key C was to be expected, all the other distractors were found to be functional, with distractor D proving to be an appealing option to 30% of the pupils.

[Table 10 about here.]

[Table 11 about here.]

$$P = \frac{C}{N} = 0,27 \tag{7}$$

The**Table 11** shows a more or less equal share of responses among options A, B and D, of which the latter is the key, while B, despite not having as many takers as the other three, was not too weak as a distractor. This fact is further exemplified by the *p*-value of 0.27 of the item, which puts it at $0.2 < p < 0.29$, meaning that it is a marginally acceptable item, but needs to be modified to be better.

[Table 12 about here.]

$$D = \frac{U_p - L_p}{U} = 0 \tag{8}$$

Mukherjee and K (2015) points out that when *p*-value is between 40% and 60%, the item will also function ably in terms of discrimination. This particular theory was used to further illustrate Item 3 earlier as a well-made item. By the same argument, perhaps in some cases, as the Difficulty Index gets lower, the item's ability to discriminate also weakens. The table shows that the same number of pupils from the upper group and lower group choosing the key as their answer, with the larger number of pupils from upper group and lower group distracted by Options A and B.

In short, Item 4 is a fairly difficult item with *p*-value of 0.27, and its difficulty has affected its ability to discriminate between the more able pupils and the rest. The same amount of pupils from the upper group chose the key as the number of pupils in the lower group, resulting in 0 for D. A further look at the distractor analysis also portrays the same trend, with distractors A *sad* and B *happy* clocking up more responses than the key. Perhaps the option for A can be replaced with a vocabulary of which the meaning is not so close to *unpleasant*.

Item 4 is interesting in many ways. A teacher expert said that that the pupils were not as familiar and 'comfortable' with the key *unpleasant*. The teachers also felt that *sad* and *unpleasant* were quite similar in terms of being synonyms to *bitter*, and would have been a double-key item in some cases. As for the option *deadly* that was not chosen by anyone, tt was most probably, according to the teacher experts, because when the pupils looked at the word and its parts, the word stem *dead* is highly unlikely to have the same meaning with *bitter*.

The most intriguing issue with the item, however, was its discrimination value of zero. Only three pupils from the higher achieving group managed to choose the key. It is a low number of responses by the higher achieving pupils if compared to the lower achieving group, and for same number of lower achievers to be able to identify the key makes it worth looking into further. The teacher experts offered the same reasoning for the low number of responses from the higher achieving group – that they were unfamiliar with *unpleasant*, hence they chose *sad*; as for the three response from the lower achieving group, it was more likely a result of guessing since it is still quite a low number. There is also a trend of selecting the longest option when in doubt, which might also have contributed to the three correct responses from the lower achieving group.

[Table 13 about here.]

[Table 14 about here.]

$$P = \frac{C}{N} = 0,73 \quad (9)$$

The *p*-value of Item 5 reveals that the item a bit too easy to be a good item on its own. With its value of 0.73, it probably does not need to be discarded, but having a *p*-value > 0.7 indicates the item needs to be some modification i.e. replacing the non-functioning distractor D with a more attractive one.

[Table 15 about here.]

$$D = \frac{U_p - L_p}{U} = 0,67 \quad (10)$$

All the pupils from the upper group converged on the key, with no one distracted by the other options, while 3 from the lower group managed to answer correctly. The item scored a D-value of 0.67, which is higher than the 0.4 it needs to be considered a very good index for its potential to discriminate high scores and low scores.

The data collected on the *p*-value of Item 5 on spelling shows the item is a slightly easy item (> 0.7). However, the item can be regarded as a quality item still due to its high D-value that showcases the high probability that the outcome of the item will see most pupils flock to the key while weaker pupils distracted by the different arrangements of spelling which is quite similar to the correct spelling and chose the wrong option. The table on Distractor Analysis above shows 27% out of all the pupils chose the other options except from the key, except for option D. It is a NFD; perhaps replacing it with a better distractor will yield a better value for *p*-value, making it a much better item.

Item 5 were also brought to the attention of the teacher experts to explain for the NFD *qeeue*. The rationale given by one of the teacher experts was that, excluding those who already located the key, pupils were never exposed to that particular combination of letters, as opposed to the other combinations by the other distractors. The pupils who chose *qeueu*might have been confused the arrangements of *u* and *e* for the spelling, and *qieue* could have been probable from the way the word is read.

[Table 16 about here.]

$$p = \frac{\sim fx - nX_{min}}{n(X_{max} - X_{min})} = 0,6 \quad (11)$$

The *p*-value for Item 6 puts the item at 0.6, between 0.30 and 0.70, which rates it as a very good item in terms of difficulty.

[Table 17 about here.]

D = Up - UL
= 0.56

The table for discrimination index shows that the item is rated as a very good item when it comes to discriminating high achievers and low achievers with D exceeding < 0.4.

Item 6 is a comprehension item whereby pupils have to read and understand a linear text and locate the answers to the stem. With the reading skill and the ability to pinpoint the key information involved, it is perhaps unsurprising that the item has a *p*-value of 0.6, rating it as an excellent item in terms of difficulty, erring a little towards being easy perhaps because the key can be lifted from the text with minimal modification. As suggested by Bichi and Embong (2018) where an item with decent *p* value will be able to discriminate well as well, its D-value of 0.56 also indicates it can work very well in discriminating the pupils that belong in the upper and those in the lower group. As things stand, Item 6is a quality item with no revision needed.

[Table 18 about here.]

$$p = \frac{\sim fx - nX_{min}}{n(X_{max} - X_{min})} = 0,4 \quad (12)$$

Item 7 has lower *p*-value than Short Answer Item 1,but is still within the range of 0.4 < *p* < 0.7, which exhibits its characteristics a good item on the Difficulty Index.

[Table 19 about here.]

D = Up - UL
= 0.61

The D-value for Item 7 is, surprisingly, almost the same as Item 6, given that there is a gap of 0.2 in their *p*-value. Item 7 is still considered to be an good item with *p* of 0.4; its D of even higher than Item 6 at 0.61, however, ranks the item as an excellent item. With both its *p* and D values, Item 7 can be regarded as a quality item. Being an item that requires higher-order-thinking skill, pupils have to come up with their own answer from the stem, which acts like a stimulus. The text does not provide key as with the case of Item 7; it merely prompts the train of thought for pupils to come up with their logical responses. It is only natural that pupils from the lower achieving group will struggle to come up with full mark responses, as shown by the table where 0 pupils from the lower group got full marks. However, 67% of the responses from higher group merited full marks, showing a gulf in their capabilities.

Lastly, Item 5, which tests on spelling of the vocabulary *queue*, has a NFD for the option *qeeue*. While spelling items are normally quite straightforward, what makes this worth examining is the two other distractors, *qeueu*and *qieue* functioned, with *qeeue* the only exception. The rationale given by one of the teacher experts was that, excluding those who already located the key, pupils were never expose to that particular combination of letters, as opposed to the other combinations by the other distractors. The pupils who chose *qeueu*might have been

confused the arrangements of *u* and *e* for the spelling, and *qieue* could have been probable from the way the word is read.

Discussion

The focus of this study is to explore the use of Classical Test Theory (CTT) to investigate the quality of test items in English Paper 1, which consists of multiple-choice and short answer items in terms of i) item difficulty, ii) item differentiation and iii) functioning or non-functioning distractions.

By definition, quality items are items that are valued favourably in their difficulty index, *p* and discrimination index (D). For *p*, it should be between 0.3 and 0.7; and D should be 0.4 or higher. In general, the 7 test items are within the acceptable range in both *p* and D, with the exception of Item 4 (Synonyms) with *p* at 0.27 and D at 0. Item 1 (Vocabulary) and 2 (Infinite to-) are also leaning towards the extreme in their *p*-values at 0.2 and 0.8 respectively, meaning minor modifications are needed for them to be considered quality items. However, even for those items who do not score within the Bichi and Embong (2018) acceptable range, they still are within the wider range proposed by other papers, such as Mukherjee and K (2015) who has their range from 0.2 to 0.9. The best items are 3 (Idioms), Item 6 (Comprehension) and Item 7(Higher-order-thinking items) for being within the acceptable range for *p* and having a high D value. What is surprising is that despite the negative perception towards subjective items, both short answer items have excellent *p* and D values; even Item 7, which tests on pupils' higher-order-thinking skills, did not fare too badly on its difficulty index. The general outcome of data analysis of the items also support the hypothesis espoused by Mukherjee and K (2015) that items that have the ideal *p*-value of between 0.4 and 0.6 will have discrimination index of 0.4 and above. Of all the items, three items have $0.4 < p < 0.6$; and of the three, Item 3 (p =0.5) has D of 0.67; Item 6 (p =0.6) has D of 0.56; Item 2 (p =0.4) has D of 0.61. The only anomaly is Item 5. Despite being rated as an easy item ($p = 0.73$), it still has a high index for discrimination.

The three more interesting items based on their *p*, D and the frequency of their distractors and looked further into with input from teacher experts are Item 2, Item 4 and Item 5. The conclusion that can be drawn from the analysis of the characteristics of Item 2 and input from the teacher experts is that distractors have to be well thought of in order to be functioning, especially when it comes to the format of items on grammar in standardized tests. There are only limited numbers of forms that can tested on i.e. root word, present tense, present continuous, past tense, past continuous etc. Even so, grammatical items need to be designed better so that correct responses from pupils indicate mastery and understanding, instead of guessing or a mere process of 'putting two and two together.

Item 4, which tested on synonyms, emphasized the importance of clear, precise options as to not create confusion. One of the points for discussion is that the distractor *sad* garnered more responses from the higher achieving group than the key *unpleasant*. The definitions between the former and the latter are quite similar, which made the options ambiguous. As aforementioned in Item 2, distractors should also be plausible and be attractive choices to the pupils, highlighted by the NFD*deadly* in this item, which, according to the teacher experts, was because the word stem *dead* is highly unlikely to be related to *bitter.* The item's inability to discriminate between the two contrasting ability levels of pupils further accentuates the significance of good working options. The ambiguous distractor and NFD contributed to the outcome where the same number of pupils from the higher achieving and lower achieving group to identify the key.

The results of Item 5 also stressed on the need to put some thought when creating or choosing distractors. The reason why*qeeue*is the NFDwas that, in reference to opinions from the teacher experts, pupils were never exposed to that particular combination of letters. Distractors have to play a role as one of the options.

Slight modifications can be done for Items 2 and 5 who have NFD respectively and have *p*-values that are slightly higher than 0.7. Perhaps by introducing more appealing distractors to the items in place of their NFDs, their *p*-values of 0.8 and 0.73 respectively, a little higher than the acceptable range, will drop to within the ideal range of $0.3 < p < 0.7$. The only item that needs a thorough revision or even being rejected as an item all together is Item 4 (p = 0.27, D = 0). Its *p*-value indicates that it is a little too difficult as an item, and D signifies its inability to discriminate the pupils in any capacity. Item 4 needs modification for the two different contrasting types of distractors. For the highly distracting *sad* who is quite close to the meaning of *bitter*, it can perhaps be replaced with an antonym, like *sweet*, to distract the pupils who do not read the instructions preceding the stem. As for its NFD *deadly*, a simpler and more probable vocabulary may be used, such as *happy* or *boring*.

However, it needs to be reminded that this study is done using only a small sample size of 30 pupils as it is a small part of a larger study. Further study will be needed to further consolidate the claims made from analysis of the data above.

CONCLUSION

The intention behind this study to have a better understanding of the items that are being used currently insummative assessments in schools around Malaysia and to see if they are effective in gauging the pupils' competence in English. Seeing that a lot of factors hinges on the outcome of these assessments (i.e. , placement in a better class, enrolment to better schools or institutions etc.), it is of paramount importance that these tests and items truly reflect each individual's competence so that the fairest evaluation of the pupils' capabilities can be made. It is with such thoughts that I have undertaken, frankly, the toughest task to date. From what was meant to be a mereassignment became a full-blown project to tackle the big issues regarding assessments and examinations.

It was my first attempt at a research in this discipline, and having to derive conclusions from formulas and calculations after a very long sabbatical was quite the challenge. There was also a lot of reading involved in order to form a solid understanding of the concepts behind this study, and for the sake of giving the study a more authentic and legitimate outlook, the layout of the paper was done based on a full-scale report of a similar study.

This study has further consolidated the usefulness of item difficulty index, *p*, item discrimination index, D, and distractor analysis as in determining the quality of items used in assessments. As shown in the results, a good item must be moderate in difficulty and have a good discriminating power of more than 0.4. On the other hand, items that have values for either indices approaching zero or negative should be revised, modified or rejected. The values presented do not only just provide conclusive evidence of the characteristics of the items, but also serves as a useful platform should further analysis is required to identify the factors that contributed to a poor item design, be it internal or external. This is especially useful with the aid of distractor analysis, as we look at the relationships not only between the individual with the items and options, but between the options themselves. Interview with the teacher experts on items that offer interesting readings also reveals important considerations when building items and choosing distractors. The item needs to be viewed not only from the professional point of view; teachers can do a lot of worse than look at them from the pupils' perspective, as ultimately they are the ones being assessed with this items in order to gauge their competence.

It needs to be noted that the item analysis and the subsequent interpretations are done on an item-to-item basis. However, when it comes to summative assessments, it is widely accepted that there needs to be a balance of easy and difficult items for results to be reliable. The study aims to look at the different characteristics of item types in standardized assessments and to determine their *p* and D-values; further study needs to be done on the test as whole so the relationship between each item's *p* and D can be correlated to the difficulty index and discrimination index of UPSR as well.

To conclude, item analysis should be a common practice among item builders and test developers because of its importance in providing vital information in producing good, quality items that are valid and reliable.

REFERENCES

Bichi, A. A. and Embong (2018). Evaluating the quality of Islamic Civilization and Asian Civilizations Examination Questions. *Asian People Journal* 1, 93–109.

Cook, D. A. and Beckman, T. J. (2006). Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *The American Journal of Medicine* 119, 166.e7–166.e16. doi: 10.1016/j.amjmed.2005.10.036.

Fitzpatrick, A. R. (1983). The Meaning of Content Validity. *Applied Psychological Measurement* 7, 3–13. doi: 10.1177/014662168300700102.

Koçdar, S., Karadağ, N., and şahin (2016). Analysis of 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, 16–24.

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, 1–11.

Mehta, G. and Mokhasi (2014). Item analysis of multiple choice questions - an assessment of the assessment tool. *International Journal of Health Sciences and Research* 4, 197–202.

Mukherjee, P. and K, S. (2015). Analysis of Multiple Choice Questions (MCQs): Item and Test Statistics from an assessment in a medical college of Kolkata. *West Bengal. IOSR Journal of Dental and Medical Sciences* 14, 47–52.

Pande, S. S., Pande, S. R., Parate, V. R., Nikam, A. P., and Agrekar, S. H. (2013). Correlation between difficulty & discrimination indices of MCQs in formative exam in Physiology. *South-East Asian Journal of Medical Education* 7, 45–45. doi: 10.4038/seajme.v7i1.149.

Pendidikan, M. K. (2013). *Kurikulum Standard Sekolah Rendah Bahasa Inggeris SJK TahunEmpat* (Kuala Lumpur: Kementerian Pendidikan Malaysia).

Salkind, N. J. (2010). Item analysis. *Encyclopedia of Research Design*. doi: http://dx.doi.org/10.4135/978141296128.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

LIST OF TABLES

TABLE 1 | Item 1

1 The plane was towed because of engine failure and parked at the _____.
A runway
B garage
C hangar
D station

TABLE 2 | Response item 1

	Response			
	A	B	C*	D
Number of pupils	3	6	6	15

TABLE 3 | Goup Tes Results item 1

Group	Response			
	A	B	C*	D
Upper group	1	1	4	3
Lower group	1	2	0	6

TABLE 4 | Item 2

2 Jordan needs at least an hour to _____ the homework.
A complete
B completes
C completed
D completing

TABLE 5 | Response item 2

	Response			
	A*	B	C	D
Number of pupils	24	2	4	0

TABLE 6 | Goup Tes Results item 2

Group	Response			
	A*		C	D
Upper group	9	0	0	0
Lower Group	5	0	4	0

TABLE 7 | item 3

Choose the most suitable idiom.
3 Adam has to get his homework done by tomorrow so he will be _____ tonight.
A crying over spilt milk
B turning over a new leaf
C burning the midnight oil
D beating around the bush

TABLE 8 | Response item 3

	Response			
	A	B	C*	D
Number of pupils	2	4	15	9

TABLE 9 | Goup Tes Results item 3

Group	Response			
	A	B	C*	D
Upper group	0	0	8	1
Lower Group	2	3	2	2

TABLE 10 | Item 4

Choose the word that has the same meaning as the underlined word.
4 Ruhil has very bitter memories of her childhood.
A sad
B happy
C deadly
D unpleasant

TABLE 11 | Response item 4

	Response			
	A	B	C	D*
Number of pupils	10	9	3	8

TABLE 12 | Goup Tes Results item 4

Group	Response			
	A	B	C	D*
Upper group	5	1	0	3
Lower Group	1	5	0	3

TABLE 13 | Item 5

Choose the word with the correct spelling. 5 We _____ up to get tickets to the theme park. A queue B qeueu C qieue D qeeue

TABLE 14 | Response item B

	Response			
	A*	B	C	D
Number of pupils	22	2	6	0

TABLE 15 | Goup Tes Results item 5

Group	Response			
	A*	B	C	D
Upper group	9	0	0	0
Lower Group	3	1	5	0

TABLE 16 | Item 6

1. Why do you think Kenny's mother screamed when she opened the present? _____ _____ [2 marks]

TABLE 17 | item 6 score

Item score	No. of students in upper group	No. of students in lower group
2	7	2
1	2	2
0	0	5

TABLE 18 | item 7

2. Why is saving money a good habit? _____ _____ [2 marks]

TABLE 19 | item 7 score

Item score	No. of students in upper group	No. of students in lower group
2	6	0
1	1	2
0	2	7