How to cite this article:

Suppiah Shanmugam, S. K., Wong, V., & Rajoo, M. (2020). Examining the quality of English test items using psychometric and linguistic characteristics among grade six pupils. *Malaysian Journal of Learning and Instruction*, *17*(2), 63-101. https://doi.org/10.32890/mjli2020.17.2.3

EXAMINING THE QUALITY OF ENGLISH TEST ITEMS USING PSYCHOMETRIC AND LINGUISTIC CHARACTERISTICS AMONG GRADE SIX PUPILS

[1]S. Kanageswari Suppiah Shanmugam,
[2]Vincent Wong & [3]Murugan Rajoo

[1]School of Education and Modern Languages, College of Arts and Sciences, Universiti Utara Malaysia, Malaysia

[2]Primary Chinese National Type School Chung Hwa Wei Sin Kuala Terengganu, Terengganu, Malaysia

[3]Research and Development Division, SEAMEO RECSAM Penang, Malaysia

[1]Corresponding author: kanageswari@uum.edu.my

Received: 1/9/2019 **Revised**: 15/5/2020 **Accepted**: 17/5/2020 **Published**: 31/7/2020

ABSTRACT

Purpose - This study examined the quality of English test items using psychometric and linguistic characteristics among Grade Six pupils.

Method - Contrary to the conventional approach of relying only on statistics when investigating item quality, this study adopted a mixed-method approach by employing psychometric analysis and cognitive interviews. The former was conducted on 30 Grade Six pupils, with each item representing a different construct commonly found in English test papers. Qualitative input was obtained through cognitive interviews with five Grade Six pupils and expert judgements from three teachers.

Findings - None of the items were found to be too easy or difficult, and all items had positive discrimination indices. The item on

idioms was most ideal in terms of difficulty and discrimination. Difficult items were found to be vocabulary-based. Surprisingly, the higher-order-thinking subjective items proved to be excellent in difficulty, although improvements could be made on their ability to discriminate. The qualitative expert judgements agreed with the quantitative psychometric analysis. Certain results from the item analysis, however, contradicted past findings that items with the ideal item difficulty value between 0.4 and 0.6 would have equally ideal item discrimination index.

Significance -The findings of the study can serve as a reminder on the significance of using Classical Test Theory, a non-complex psychometric approach in assisting classroom teacher practitioners during the meticulous process of test design and ensuring test item quality.

Keywords: Classical test theory, item difficulty index, item discrimination index, test item quality, psychometric properties.

INTRODUCTION

Research on assessment has been gaining popularity over the last decade as people become more conscious of its close relation to learning. Whether summative or formative, assessments provide valuable feedback on learners' understanding of subject matter and their response to certain teaching methods, from which review and improvement can be done to better effect (Haladyna, 2002, as cited in Koçdar, Karadağ & Şahin, 2016). One of the most straightforward methods of obtaining such feedback is via testing. A good test is able to provide quality feedback on the intended construct, and in order to determine the quality of the test, its items must be analysed in terms of their difficulty and ability to discriminate between the pupils (Koçdar et al., 2016).

Item analysis is essential to any test construction as it examines "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, Pande, Parate, Nikam & Agrekar, 2013, p. 46). It also provides a better depiction of the test items' characteristics (Salkind, 2010). According to Thompson and Levitov (1985, as cited in Bichi & Embong, 2018), item analysis looks at the performance of an item in relation to other factors to better understand its characteristics and

identify its flaws, if any. Item analysis can provide test developers with useful information in constructing items with better quality and accuracy. For classroom practitioners, it also provides knowledge on students' strengths and weaknesses, which can have direct implications on classroom pedagogy. According to Brown and Hattie (2012) good items can inform pupils how they are doing and what they may still need to learn, and further motivate them as they are made aware of their gaps in learning. Teachers can also use this information to provide accurate and effective feedback and learning goals to students.

The novelty of this study is that using a case study design, it adopted Classical Test Theory and cohesively incorporated substantive analysis to obtain qualitative insights on the quality of test items, an issue that is commonly investigated using quantitative data only (see Kehoe, 1994; Olatunji et al., 2007; Singh et al., 2009; Gajjar, Sharma, Kumar & Rana, 2014; Mehta & Mokhasi, 2014). Quantitative data alone has proven to be inadequate as deeper issues in education require more holistic and in-depth explanations; hence the need for an injection of qualitative insights to complement statistics has become more prominent (Zainal, 2007). A case study design utilizes both quantitative and qualitative methods to collect data, which not only adds more credibility to the study, but also provides a better perspective on the complexities of real-life phenomenon often lacking from purely quantitative studies. Duff and Anderson (2016) defines case study design as an approach that constitutes a qualitative, interpretive method to formulate grounded understandings of issues through holistic and in-depth characterization of the individual components in context.

While the groundwork in this study still closely resembled the typical approach to item quality analysis, qualitative data was introduced via cognitive interviews with pupils, interviews with teacher experts and researcher's reflection. Quantitative data was obtained through pupils' test scores on selected test items while qualitative data played a pivotal role as a complement to the raw numbers. The inclusion of multiple sources of qualitative input was to provide diverse perspectives from different angles on individual items, thus giving a better description of the quality of the items. Participants who were interviewed, both teachers and pupils, were meticulously selected to cover a range of interpretations for each item, and interviews were done using open-ended questions to allow maximum liberty for participants to express themselves.

Standardised Assessment

Standardised assessments are not new within the sphere of education, and have been widely and heavily relied upon as a fair means of measurement. They are formal assessments that are administered to a large group, aimed at generating a clear score or scores if different areas are measured, which can be used to compare an individual to others as to how they rank in terms of their ability (Mons, 2009). As defined by the National Council for Curriculum and Assessment (2005, p.2), a standardised assessment "is an instrument…that contains standardised procedures for its administration and scoring and for the interpretation of its results [with] objectively-scored items… that are normed-referenced." Many countries such as England, France, Germany, the United States, Canada, Hong Kong and Israel have been documented utilising high-stakes large-scale standardised assessment as a method of grading their students (see Black & William, 2005; Mullis, Martin, Kennedy & Foy, 2007; Mullis & Mullis, 2003).

The Malaysian Education Ministry also employs several means of standardised national assessments to gauge students' progress throughout their education. One of them is the Primary School Evaluation Test or *Ujian Penilaian Sekolah Rendah* (UPSR), a summative assessment that Grade Six pupils have to undergo at the end of primary level education. It is an important benchmark for Grade Six pupils nationwide as the results are a barometer of pupils' academic performance and aptitude after six years of formal education. The number of subjects tested and the format differ between national and national-type schools as their syllabus and curriculum are slightly different. For example, in national schools, pupils sit for six papers while in national-type schools, pupils have to take extra papers in their mother tongue. The focus of this study is on the English test, which is not too dissimilar in terms of layout and structure.

The English test is divided into two separate papers, taken at two separate time slots. Paper 1 consists of two sections: Section A contains 20 multiple-choice items on vocabulary, grammar, idioms or proverbs, synonyms/antonyms and a comprehension text with questions. The multiple-choice items are single-best types. Section B features items on short social exchanges and comprehension questions which include true or false and short answer items based

on linear and non-linear texts. Section B allows a certain degree of freedom with the language; however, candidates still have to address the requirement stipulated by the stems.

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

Research Objective

Due to the importance of assessment and its outcome to all parties involved (see Moodley, 2015; Mogapi, 2016; Norafizah, 2018; Tayeb, Aziz, & Ismail, 2018), it is essential that careful measures are taken to ensure that the items used can accurately reflect pupils' English competence. However, the quality of test items is often questioned (Haliza, 2017). The main purpose of this study is to investigate the use of Classical Test Theory (CTT) in order to determine the quality of test items in English Paper 1, which consists of multiple-choice and subjective items. The rationale underlying this study is to encourage the versatile use of CTT in the classroom among teacher practitioners to better understand the quality of test items and detect flawed items from the psychometric perspective, on which items are deemed easy or difficult from the pupils' perspective. It is hoped that CTT could become a viable assessment tool in the classroom. Accordingly, the use of CTT by a teacher practitioner on a classroom-administered teacher-devised test is the focus of this study. The aims of this study are to:

i) determine the psychometric properties of test items using item analysis, and
ii) identify linguistic characteristics of test items with input from pupil cognitive interview and teacher expert judgement.

Linguistic Characteristics of English Test Items

As discussed in the previous section, the English standardised tests are separated into two papers. This study takes a closer look at Paper 1 of the whole assessment, which are made up of 20 multiple-choice items and 5 items on social expressions. The items for the former section may test on a range of linguistic characteristics, such as vocabulary, possessive pronouns, subject-verb agreements, conjunctions, forms of verb, prepositions, adverbs, proverbs, synonyms/antonyms, punctuation. In the latter section, pupils normally encounter short-answer items that require them to exhibit language properties, i.e., to ask, to be polite, to apologize, to congratulate, etc., the ability to extract information from a reading text and give comprehensive short answers, and higher-order thinking skills (HOTS) (Mokshein, 2019).

The items chosen for this study tested vocabulary (Obj. 1 – Item 1), forms of verb (Obj. 2 – Item 2), idioms (Obj. 3 – Item 3), synonyms (Obj. 4 – Item 4) and spelling (Obj. 5 – Item 5) from Section A. These items were selected as they represented the more popular types of item types in standardised tests. Vocabulary items test pupils' ability to identify the correct vocabulary associated with the context proposed by the stem and the structure. In the item chosen, the pupils were required to demonstrate their knowledge of aviation-related vocabulary, with key words such as *plane*, *towed* and *hangar.* Items testing on forms of verb assess pupils' mastery of English grammar; in this case, it was the to-infinitive. Items on idioms present contexts that pupils have to match with the most suitable idioms, whereas items testing synonyms and spelling further test pupils' linguistic knowledge, with the former focused on semantics of the vocabulary and the latter on knowledge of the English orthography.

The two subjective items (Subj. 1 and Subj. 2) chosen from Section B of English Paper 1 were Item 6 (Subj. 1), which was a comprehension item closely related to the reading text and Item 7 (Subj. 2) that required pupils to engage their HOTS by inferencing from the text. For the comprehension item, pupils needed to elicit the answer from the reading text based on the requirements of the stem, which was '*Why do you think Kenny's mother screamed when she opened the letter?*'. HOTS items are items that require pupils to generate their own responses from their understanding of the overall plot as well as

specific details in the text, and make appropriate inference from it. In all these test items, careful consideration was given to the construction of the stem. This is because in addition to the content of the test items and choice of options, considerable attention needs to go into the stem (Zimmaro, 2014). Stems should be clearly written and the words chosen should help the examinee understand what is being asked instead of confusing them with the possibility of ambiguity. This is to avoid any other factors not related to the test construct being measured confounding students' ability to select the answer. Zimmaro (2014) further explained that the element of testing should not only be confined to choosing the key from the given options; if examinees find difficulty in understanding the stem due to the use of high-level vocabulary beyond their comprehension or ambiguity in the construction of the stem, the outcome of the test will be affected as pupils are not tested fairly.

Classical Test Theory

Classical Test Theory (CTT), also known as 'True Score Theory', is a psychometric approach used to evaluate the quality of measures, such as questionnaires, surveys and achievement tests. Central to the theory are three concepts: observed test scores (O), which is the result of true score (T) and error score (E) (Magno, 2009). True scores are the examinees' real score if there are no errors in measurement instruments; however, this 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 (Kaplan & Saccuzzo, 1997, as cited in Magno, 2009); the larger the standard error of measurement, the less accurate the measurement of the intended attribute, and vice versa.

CTT operates based on the assumption that the differences between the responses of examinees 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 attracts the scrutiny of results is that all other sources of variation, such as external factors of the surroundings or physical and mental conditions of the examinees are constant throughout repeated standardization procedure, or just random and unsystematic occurrences (Magno, 2009).

Historically, principles of CTT have pioneered methods of analysis used to evaluate tests by looking at four criteria: frequency of correct responses to indicate item difficulty, frequency of each of the responses to analyse distractors for their functionality; the correlation between the items and responses between higher and lower achieving groups of examinees (Impara & Plake, 1997, as cited in Magno, 2009). As is common with theories that have existed for some time, CTT is not without its detractors. It does have its own limitations, which mostly revolve 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 on which they take the test. To address the shortcomings of CTT, Item Response Theory (IRT) was introduced in 1969 (van der Linen & Hambleton, 2013)

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. An advantage of CTT over IRT is the assumptions of the data by the theory allows for a simpler concept of the model, making it a friendlier model to be implemented in the classroom. The focus of the psychometric analysis in CTT is on measuring at the 'test' level, in contrast with the item-level focus of IRT. Despite its limitations, CTT is still widely used as it represents a more economical and practical method of developing quality test items. Hence, in this study, the items were analysed using CTT, and going beyond the psychometric analysis, the objective was also to substantiate the psychometric findings with qualitative analysis from the cognitive input and expert judgements, in addition to personal reflections.

Item Analysis

The item analysis was conducted by computing i) Item Difficulty Index, *p*, ii) Item Discrimination Index, D, and conducting iii)

Distractor Analysis without using any expensive, advanced or sophisticated software. The advantage of this approach using CTT is that these indices that inform the psychometric properties of test items can be easily calculated using a calculator or Microsoft Excel by any teacher practitioner without the need to have advanced knowledge or software in psychometrics. Four research papers were reviewed 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 from 0.0 and 1.0 derived from pupils' correct responses (Bichi & Embong, 2018). The higher the *p*-value, the easier the items are and vice versa.

Item Discrimination Index, D, measures how an item is able to discriminate the more able from the less able pupils (Mehta & Mokhasi, 2014). An index value of +1 means that the item is very effective, whereas 0 indicates that the item is unable to discriminate at all. If the discrimination index is negative, then it is a faulty item where more pupils from the lower ability group had managed to select the correct responses more frequently than pupils from the higher ability group (Bichi & Embong, 2018).

A multiple-choice item has a stem and four options. The four options contain a key and three distractors. Distractor Analysis looks at how effective the distractors are in influencing the pupils' judgement in identifying the key (Mehta & 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). NFDs must be revised, removed, or replaced with better options. Studies from Mukherjee and Lahiri (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 Lahiri (2015) looked at elements of multiple-choice questions in tests and proposed that items with *p*-values between 0.2 – 0.9 are good items. Items with *p*-values between 0.4 – 0.6 are excellent items. Item valued at less than 0.2 or above 0.9 are too difficult or too easy, respectively. Mukherjee and Lahiri (2015) also claimed that items valued between 0.4 and 0.6 have maximum discrimination index. As for D, values of 0.40 and above are excellent items; items with D between 0.30 and 0.39 are reasonably

good, whereas 0.20 to 0.29 means that they need to be reviewed. Value of 0.19 and below would rank them as poor items and could be rejected. Mukherjee and Lahiri (2015) further suggested 5% or more pupils are needed for a distractor to be deemed effective.

Bichi and Embong (2018, p. 98) recommended "values of difficulty no less than 30% correct and no greater than 70%." Items with *p*-values smaller than 0.3 are too difficult and those larger than 0.7 are too easy; these items are consequently weaker in their ability to discriminate high scorers and low scorers. The Item Discrimination Index, derived from the works of Ebel and Frisbe (1991, as cited in Bichi & Embong, 2018) classified items with values of 0.4 and above as 'very good' and 0.3 to 0.39 as 'reasonably good' but subject to improvement. Items with values between 0.2 to 0.29 are usually subjected to revision and items below 0.19 are 'poor'. However, the study rated a distractor as effective as long as it garnered one response.

Similar to Bichi and Embong (2018), Mehta and Mokhasi (2014) rate items with *p*-values between 0.3 and 0.7 as acceptable, and further advocate values between 0.5 and 0.6 as ideal. Items placed in the two extremes ($p < 0.3$ and $p > 0.7$) need modification as they are unacceptable as they are. In terms of D, items with an index of more than 0.35 are considered as excellent; a D-value between 0.2 and 0.35 is 'good' and those with any index less than 0.2 are 'poor' items. As for distractors, Mehta and Mokhasi (2014) deemed a distractor as effective if it is selected by 5% or more pupils.

METHODOLOGY

The study adopted a mixed method research design (Creswell, 2013) utilizing both quantitative and qualitative data to address the objective of the research, which aimed to study English standardised test items for the purpose of improving the test items. For quantitative data, psychometric analysis was used on the scores of the test items, which were obtained during a 40-minute test administration. Qualitative input on the test items were obtained using pupils' feedback from cognitive interviews and expert judgements from three teacher experts. To substantiate the psychometric and qualitative analyses, the study also triangulated data from the researcher's personal reflections.

As both expert judgments and psychometric analyses have their own limitations, the merging of the two methods for a fair, reliable and valid test is recommended. Therefore, validity and reliability were ascertained by conducting a two-step analysis involving expert judgement and psychometric analyses (Hambleton & de Jong, 2003; Hambleton & Patsula, 1999) and were further strengthened using self-reflections.

Test Items

The items in the test paper were chosen from a revision workbook that published past year English papers and were similar to items commonly found in English standardized tests. To preserve the authenticity of the items, no changes were made to the original stem or the options. 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) and two short answer items from a comprehension text. The items were printed on a single sheet of A4 paper, with five multiple-choice items forming the objective section on one side and the subjective section containing a comprehension text with its short answer items on the other.

Content Validity

As with all kinds of assessment, a test will not be seen as an effective form of measurement without evidence of content validity. In this study, content validity is defined as how well the items of the assessment cover the content, knowledge or skills that it claims to cover (Messick, 1975, as cited in Fitzpatrick, 1983). Cronbach (1971, as cited in Fitzpatrick, 1983) further illustrated this point by stating that an achievement test has to reflect the content domain outlined in a test manual. In short, tests demonstrate content validity when they test what they are meant to test.

To establish content validity, the test items were compared to a Table of Specification based on the Curriculum Standard of Grades 4 – 6. A Table of Specification is a test blueprint prepared by classroom teachers as a basis for a test, describing the topics to be covered on the test and the number of items associated with each topic as well as their respective cognitive level (Fives & DiDonato-Barnes, 2013).

Table 1

Table of Specification

Content Standard	Learning Standards	Bloom's Taxonomy						Total Count
		Remem-bering	Under-standing	Apply-ing	Analyz-ing	Evaluating	Creat-ing	
2.2 By the end of the 6-year primary schooling, pupils will be able to demonstrate understanding of a variety of linear and non-linear texts in the form of print and non-print materials using a range of strategies to construct meaning.	**Year 5** 2.2.1 Able to apply word attack skills by: (a) using contextual clues to get meaning of words: (i) before the word (anaphoric) (ii) after the word (cataphoric) (b) identifying idioms			Obj. 1 **(Item 1)**		Obj. 4 **(Item 4)**		**2**
	2.2.2 Able to read and understand phrases and sentences from: (a) linear texts (b) non-linear texts			Obj. 3 **(Item 3)**				**1**
	Year 6 2.2.3 Able to read and demonstrate understanding of texts by: (a) giving main ideas and supporting details (b) drawing conclusions with guidance				Subj. 1 **(Item 6)**			**1**

Content Standard	Learning Standards	Bloom's Taxonomy						Total Count
		Remembering	Understanding	Apply-ing	Analyz-ing	Evaluating	Creat-ing	
3.2 By the end of the 6-year primary schooling, pupils will be able to write using appropriate language, form and style for a range of purposes.	**Year 4** 3.2.4 Able to spell words by applying spelling rules.	Obj. 5 **(Item 5)**						**1**
3.3 By the end of the 6-year primary schooling, pupils will be able to write and present ideas through a variety of media using appropriate language, form and style.	**Year 4** 3.3.1 Able to create simple texts using a variety of media with guidance: (a) non-linear (b) linear						Subj. 2 **(Item 7)**	**1**
5.1 By the end of the 6 year primary schooling, pupils will be able to use different word classes correctly and appropriately.	**Year 4** 5.1.3 Able to use verbs correctly and appropriately: (a) irregular verbs (b) verbs that do not change forms				Obj. 2 **(Item 2)**			**1**
	Total Count	**1**	**0**	**2**	**2**	**1**	**1**	**7**

Note. Adapted from Anderson and Krathwohl (2001), and Yoong, Lee, and Kanagamani (2015).

The Table of Specification (see Table 1) was constructed by studying the skills that were tested by the test items, mainly Reading, Writing and Grammar and their learning standards, and comparing them to the revised Bloom's Taxonomy (2001). Content and learning standards for Listening and Speaking and Language Arts were omitted as these skills do not appear explicitly in separate sections in the UPSR. The constructs of the test items were as exhibited in parentheses for each item in the subsequent section.

Sample

The participants who took this test were 30 Grade Six pupils who, at the time of the study, were enrolled in a national-type Chinese school in one state in Malaysia. They were from a mixed ability group and had varying levels of English proficiency. Most of their input in English were during the English lessons. Based on their marks on the test, five pupils were selected to give their feedback about the content and quality of the test items from a cognitive perspective. Two pupils were selected from the high band, one pupil from the middle band, and two from the low band. This was to ensure that input from the pupils' perspective can represent different groups of pupils with varying levels of competence.

Three English teachers from the same school were also involved in providing their expert judgement on the linguistic characteristics of the test items. In order to gather high-quality responses that give an accurate appraisal of the items, the teacher experts selected fulfilled the criteria of: i) having taught for more than ten years at the same school where the sampled pupils studied, ii) familiar with the standardized summative assessments that the test was based on, and iii) familiar with the Grade Six pupils and their ability in general.

Test Administration

Prior to the test administration, 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 given 40 minutes to complete it. Pupils were also reminded that the results of the test would not be reflected in their academic achievement, so they should just try their best to answer without too much pressure.

Scoring

The scoring for each item was different, depending on the type of answer it elicited from the pupils. For multiple-choice items, there were four options; only one of them was the key and the other three were distractors. Scoring for the multiple-choice items was dichotomous – choosing the key would earn the pupil one mark, whereas selecting one of the three distractors would result in zero mark.

The subjective section of the test consisted of short answer items that were polytomous. The items had scores ranging from zero (inaccurate response or no response) to one (response partially correct or contained grammatical errors) and finally two marks, which was the maximum a pupil could get for an accurate and coherent response. Scoring was based on a rubric adapted from the original rubric designed for the specific sections in standardized tests.

Cognitive Interview

The five pupils involved in this part of the study were chosen according to their test scores and their willingness to continue participating in the cognitive interview. Two pupils were selected from the high band (6-7 marks), one pupil from the middle band (3-5 marks), and two from the low band (0-2 marks). The chosen pupils were gathered in a classroom and interviewed individually. They were first asked about their opinions on each item, i.e., whether they found the items challenging and what part made it challenging. After they were briefed on the basic procedure of the cognitive interview, the researcher inquired further on the findings, i.e., why some items received an equal amount of responses from both high-achieving and low-achieving pupils and why most/some pupils answered wrongly. The pupils were allowed to respond in their mother tongue so that they could better express their opinions and thoughts. Their responses were recorded using a recording device on a mobile phone, and in point form using pen and paper. Each interview took approximately 20 – 30 minutes.

Teacher Expert Judgement

Sireci and Geisinger (1995) point out that as few as three subject-matter experts are adequate for a comprehensive expert judgement.

Being a small-scale study, only three teacher experts from the same school were chosen to provide judgement on the quality of test items. The three teachers were approached individually and the issues were discussed in an informal manner at their desks. The teachers were first briefed very simply on the purpose of the study and given a clean copy of the test items to read through. Then they were asked for their expert opinion on how the pupils would approach the test items and to what extent the pupils would find the items challenging. After they gave their input on each item, the researcher informed them of the findings from the quantitative analysis and inquired further on their elaboration. The interviews took around 15 – 20 minutes to complete. Their responses were recorded with pen and paper.

Data Analysis

This study utilized the *p* and D values as advocated by Bichi and Embong (2018). The acceptable item difficulty index, *p*, was between 0.3 and 0.7. A value below 0.3 was treated as too difficult and above 0.7 as too easy. The *p* value was calculated using responses from all the pupils taking the test. As for discrimination item, D, items with a value above 0.4 were considered excellent in terms of discriminating between high and low achievers. Items with D-values between 0.3 and 0.39 were regarded as reasonably good but could still be improved, whereas items with D-values ranging from 0.2 to 0.29 were only marginally acceptable and should be modified in order to be included on a test. Items with values 0.19 and below should be rejected. Due to the small sample size, Bichi and Embong's (2018) interpretation of at least one (n=1) pupil selecting an option to be sufficient as a functioning distractor was used. This is consistent with the 5% recommended by Mukherjee and Lahiri (2015) and Mehta and Mokhasi (2014), and for n=30, the value obtained was 1.5, which also amounted to one or two pupils. Nevertheless, comparisons were made with other literature during quantitative analysis when necessary.

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 the summation of the frequency of pupils on the scores 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 was two marks, while X_{min} is the minimum points that the pupil can get, which in this case was zero. Item discrimination index, D, was calculated using the formula $D = \frac{U_p - L_p}{U}$, U_p is the number of pupils answering the item correctly in the high achieving group while L_p represents the number of pupils in the low achieving group who answered the item correctly. *U* symbolises the number of high performers or low performers. Although the usual division of the high achieving and low achieving group is 27% of the pupils at the top and 27% at the bottom ranked according to their test scores (Bichi, 2015), it is recommended that for tests administered to only one class of pupils, the papers should be ranked from the highest to the lowest, whereby the top half of the papers would be the high achieving group while the lower half would be the low achieving group (Veloo & Awang-Hashim, 2016). In this study, the total number of pupils was 30, hence the number of pupils was set at 15 for both the high achieving and low achieving groups.

RESULTS

Item Analysis

Each item in the test was analysed for its quality using distractor analysis, item difficulty index, *p*, and item discrimination, D. The distractor analysis would present the data on the pupils' responses to all the options given for each item in order to identify how effective the distractors were. This was followed by calculations of *p* to identify the difficulty index for each item. Lastly, the D-value for each item was calculated to determine its efficiency in distinguishing the more able and less able pupils in relation to the construct on which each item is based. The results are presented for each item along with its interpretation and analysis.

Table 2

Item 1

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

Table 3

Pupils' Responses for Item 1

		Response				Total
		A	B	C*	D	
p	Number of pupils	3	6	6	15	30
D	Upper group	2	2	6	5	15
	Lower group	1	4	0	10	15
	Total	3	6	6	15	30

$$p = \frac{6}{30}$$
$$= 0.2$$

$$D = \frac{6}{15} - \frac{0}{15}$$
$$= 0.4 - 0$$
$$= 0.4$$

Table 3 shows that 15 pupils, or 50% of the class, chose the Distractor D, and only 6 pupils, or 20% of the group, managed to identify Distractor C as the key. The two other distractors were also functioning, although Distractor B had the same number of responses as the key. This was reflected in the *p*-value of 0.2, which ranked the item as 'difficult'. The discrimination index, D was valued at 0.4, which was 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 could discriminate well, but it was a difficult item, bordering on

'needs to be rejected or modified entirely'. The distractor analysis revealed that 50% of the class actually selected D, which contained the distractor 'station' to be paired with the subject 'plane' and the verb 'towed'. Instead of deducing that the distractor was functioning well, it would seem that the key 'hangar' was too unfamiliar to most of the pupils for them to select that as their answer. In such cases, the key or stem should be modified to use vocabulary that is of high frequency or familiar to the pupils.

Feedback from the cognitive interview reported two difficulties that pupils faced with the item. At the vocabulary level, words such as *towed*, *failure*, *runway* and *hangar* were 'too difficult for [them] to understand'. One of the pupils interviewed also claimed that "[he didn't] understand the sentence… [it was] very long." Upon inquiry about the high number of responses towards the Distractor D *station*, the reasons given were: 1) familiarity with the word and its common association with transportation, and 2) guessing.

Opinions from teacher experts echoed the pupils' views. The item was deemed difficult with regards to both vocabulary and sentence level. The key words such as *towed*, *runway* and *hangar* were highly technical terms and pupils were unlikely to have encountered them frequently enough for them to learn those words. "[They] will go for *station*, because they know *station*", as claimed by one teacher expert. The sentence structure also posed a challenge to the pupils as it was connected by both *because* and *and*; the higher-achieving pupils might not have much problem understanding the sentence, but it would have confused pupils with lower English competency.

Table 4

Item 2

Construct	5.1.3 Able to use verbs correctly and appropriately focused on different forms of verb
Item	Jordan needs at least an hour to _____ the home-work. A. complete B. completes C. completed D. completing

Table 5

Pupils' Responses for Item 2

		Response				Total
		A*	B	C	D	
p	Number of pupils	24	2	4	0	30
D	Upper group	15	0	0	0	15
	Lower group	9	2	4	0	15
	Total	24	2	4	0	30

$$p = \frac{24}{30} = 0.8 \qquad D = \frac{15}{15} - \frac{9}{15} = 1 - 0.6 = 0.4$$

Table 5 shows that Item 2 had a very high *p*-value of 0.8, which meant that it was too easy. While it was still acceptable by Mukherjee and Lahiri's (2015) estimation, it had already lost its potency as an 'ideal' item, which should be between $0.3 < p < 0.7$ (Bichi & Embong, 2018). The data indicated that more than three quarters of the pupils managed to locate the key, with the Distractor D completely ignored by all of them.

Item 2 functioned as well as Item 1 in terms of their discrimination index, D, which was also 0.4. Again, it showed that the item was doing well in distinguishing the high achievers from the low achievers, as seen in Table 5. 100% of the pupils in the upper group chose the key, while only 9 from lower group were able to do so.

The analysis showed that Item 2, which was a grammar item testing the to-infinitive, was an easy item, as reflected by the high *p*-value (*p=80)*. The data from the distractor analysis also supported this finding as only 6 pupils chose the distractors, and Distractor D *completing* was a non-functioning distractor. However, despite not having the ideal value for *p*, it still had quite a high D score, showing its potential as an item that is able to distinguish between the high and low achievers. One way to make the item better would be to replace D with a plausible distractor that could attract some of the pupils who chose the key, to lower its *p* to between 0.3 and 0.7. One recommendation by the teacher expert was the option *completion*, as

this noun form was not commonly encountered by the pupils and it could confuse some of them who "memorize" this type of question rather than understand the mechanics of the language.

The interviewed pupils' input also supported the statistical evidence from the data analysis, with most of them claiming that the item was "easy" and "[they] often answer this type of question." One of the pupils interviewed who answered wrongly claimed that he chose *completes* due to confusion over the tense forms. When inquired about *completing* which was not chosen by any of the pupils, their general response was that they had been taught not to choose the continuous tense form *–ing* when *to* preceded the word.

The teacher experts were of the same opinion. They reasoned that the options *completing* and *completed* would not be likely to receive a high number of responses from the pupils due to how they were taught, which was consistent with the data. For the former, pupils have been somewhat conditioned to only choose verbs in their continuous form only if they managed to locate the auxiliary verb *be*, and *completed* was unlikely to be the correct answer because the past tense form was not present in the stem. As for *completes,* one teacher expert claimed that some pupil would choose that distractor due to presence of *needs* in the stem, which suggested the possibility of the singular form.

Table 6

Item 3

Construct	2.2.2 Able to read and understand phrases and sentence from a) linear texts and b) non-linear texts
Item	Choose the most suitable **idiom**. 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 7

Pupils' Responses for Item 3

		Response				Total
		A	B	C*	D	
p	Number of pupils	2	4	15	9	30
D	Upper group	0	1	11	3	15
	Lower group	2	3	4	6	15
	Total	2	4	15	9	30

$$p = \frac{15}{30} = 0.5 \qquad D = \frac{11}{15} - \frac{4}{15} = 0.73 - 0.27 = 0.46$$

As displayed in Table 7, the *p*-value of Item 3 was 0.5, which, according to a few other studies, was an ideal score for multiple-choice items. It ranked as 'excellent' in terms of difficulty (Mehta & Mokhasi, 2014). When items have *p*-value of $0.4 < p < 0.6$, their discrimination index is also high (Mukherjee & Lahiri, 2015).

As suggested by Mukherjee and Lahiri (2015) and supported by the data, Item 3 scored the highest in D-value in this test. Table 7 shows a clear gap between the upper and lower group for the key, with 11 pupils in the upper group but only 4 in the lower group managing to locate the key. A more remarkable outcome is that there was an even distribution across the four options for the lower group, whereas responses from the higher achieving group centred on the key, with only one choosing the distractor B and three choosing distractor D.

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

Two of the pupils interviewed found this item to be easy, while three of them found it to be hard. The two respondents who said the

item was easy expressed their familiarity with the idiom *burning the midnight oil*, and the meaning complemented the stem perfectly. However, they acknowledged that a lack of knowledge of the idioms concerned made answering a challenge. As for the latter group who found the item to be challenging, they claimed that they resorted to guessing because the options were too difficult to understand. One respondent who chose the option *beating around the bush,* which had the second highest number of responses, explained that she used her own method of understanding the idioms and "imagined last-minute work as wildly swinging, beating around the bush in order to finish as soon as possible."

The teacher experts' responses reflected the results from the item analysis, claiming that the general perception of items testing idioms was that "good pupils are always able to get idioms correct" and "poor pupils always seem to struggle with idioms." The teacher experts stated that idioms were part of the syllabus; while high ability pupils seemed to be able to grasp the concept of idioms fairly quickly, low ability pupils could not seem to understand that idioms had a deeper meaning than at word level. One teacher expert believed that "it could be because their English is poor [and] idioms require them to know the words before they can comprehend the meaning between the lines." As the pupils were already struggling with words and their basic meaning, it would be too demanding for them to deal with idioms, which require a certain level of comprehension and imagination.

Table 8

Item 4

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
Item	Choose the word that has the **same meaning** as the underlined word. Ruhil has very <u>bitter</u> memories of her childhood. A. sad B. happy C. deadly D. unpleasant

Table 9

Pupils' Responses for Item 4

		Response				Total
		A	B	C	D*	
p	Number of pupils	10	9	3	8	30
D	Upper group	7	2	1	5	15
	Lower group	3	7	2	3	15
	Total	10	9	3	8	30

$$p = \frac{8}{30} = 0.27 \qquad D = \frac{5}{15} - \frac{3}{15} = 0.33 - 0.2 = 0.13$$

Table 9 shows the responses for Item 4. There was an equal distribution of responses among options A, B and D, which was the key, while B, despite not having as many takers as the other three, was not too weak as a distractor. This fact was further exemplified by the *p*-value of 0.27 for the item, which put it in the range of $0.2 < p < 0.29$; i.e., it was a marginally acceptable item, but needed to be modified for improvement.

Mukherjee and Lahiri (2015) point out that when the *p*-value is between 40% and 60%, the item also functions well in terms of discrimination. This particular postulation was used to classify Item 3 earlier as a well-made item. Expanding further, it is possible that as the difficulty index decreases, the item's ability to discriminate also weakens. The table shows that the same number of pupils from the upper group and lower group chose the key as their answer, with the larger number of pupils from the upper group and lower group distracted by Options A and B.

In short, Item 4 was a fairly difficult item with a *p*-value of 0.27, and its difficulty had affected its ability to discriminate between the more able pupils and the rest. Five pupils from the upper group and three pupils from the lower group managed to locate the key, resulting in 0.13 for D. A further look at the distractor analysis also revealed the same trend, with distractors A, *sad,* and B, *happy,* garnering more responses than the key. A viable option for distractor A could be a word that was not so close to *unpleasant* in its meaning.

Feedback from the cognitive interview reflected the findings from the data analysis on the item. The pupils interviewed reported the item as "difficult" and "confusing," as the word "bitter" was more commonly associated with taste in their repertoire. The second point raised was the ambiguity of the options. Two of the pupils interviewed expressed uncertainty when trying to identify the key: "I can't tell if *sad* or *unpleasant* is more suitable as the answer" and "Except for *happy*, the other three all looked like possible answers to me."

Teacher experts offered a more in-depth observation. Excluding the factor of carelessness when it came to items that set specific instructions asking for similar/opposite meaning, teacher experts deduced that the pupils were probably not as familiar and "comfortable" with the key *unpleasant*, and instead went for the "second best" option available to them. One teacher was fixated on the very similar nature of *sad* and *unpleasant* in terms of being synonymous to *bitter*, and argued that the item would be deemed to have two keys in certain cases. The teacher experts also felt that the poorer pupils would struggle to identify *unpleasant* as the key and would probably choose *sad*, which would be more familiar and "safe" for them. Pupils also tended to guess or select the longest option when in doubt, which might have contributed to the same number of responses from the lower achievers and the high achievers.

Table 10

Item 5

Construct	3.2.4 Able to spell words by applying spelling rules
Item	Choose the word with the correct **spelling**. We _____ up to get tickets to the theme park. A. queue B. qeueu C. qieue D. qeeue

Table 11

Pupils' Responses for Item 5

		Response				Total
		A*	B	C	D	
p	Number of pupils	22	2	6	0	30
D	Upper group	14	1	0	0	15
	Lower group	8	1	6	0	15
	Total	22	2	6	0	30

$$p = \frac{22}{30} = 0.73$$

$$D = \frac{14}{15} - \frac{8}{15} = 0.93 - 0.53 = 0.4$$

Table 11 shows the pupils' responses for Item 5. Its *p*-value revealed that the item was too easy. At 0.73, it did not need to be discarded, but having a *p*-value > 0.7 indicated the item needed some modification, i.e., replacing the non-functioning distractor D with one that would be able to distract some of the pupils from locating the key too easily.

All but one pupil from the upper group converged on the key, with no one distracted by the other options, while eight from the lower group managed to answer correctly. The item scored a D-value of 0.4, which proved it was adequate in discriminating between high and low scores.

The data collected on the *p*-value of Item 5, which tested spelling, showed the item as a slightly easy item (> 0.7). However, the item could still be regarded as a quality item due to its high D-value, which showcased the high probability that most pupils would flock to the key while weaker pupils, distracted by the different arrangements of letters which was quite similar to the correct spelling, would chose the wrong option. Table 11 also shows that 26.67% of the pupils selected the distractors, except Distractor D which was an NFD. Replacing it with a better distractor would probably yield a better *p*-value.

All five of the pupils interviewed found this item to be easy; one even said that he would be very happy for this type of question to

appear in the actual UPSR paper. According to the pupils, items that test the spelling of words were very easy because they knew the spelling, and even if they did not know the spelling, they still had a high probability of getting the correct answer by guessing from the arrangement of the alphabets and how the word was pronounced. They felt that it was possible for some pupils to get this item wrong, either due to carelessness or not knowing the spelling of *queue*; however, the number would be minimal.

The responses from teacher experts regarding Item 5 were generally similar to the interviewed pupils; however, the focus of the interview with teacher experts was on the NFD, i.e., Distractor D *qeeue*. The teacher experts' input was consistent with the data. The rationale given by one of the teacher experts was that most pupils would never choose that option as they are not exposed to that particular combination of letters, as opposed to the combinations presented in the other distractors. The pupils who chose *qeueu* might have been confused with the arrangements of *u* and *e* for the spelling, and *qieue* was probable from the way the word is read.

The next section will discuss the subjective items selected for the study, and the analysis of the discrimination index, D that was done using the formula proposed by Bichi and Embong (2018) for subjective items.

Table 12

Item 6

Construct	2.2.3 Able to read and demonstrate understanding of texts by: a) giving main ideas and supporting details and b) drawing conclusions with guidance
Item	Why do you think Kenny's mother screamed when she opened the letter? _____ _____ [2 marks]

Table 13

Pupils' Responses for Item 6

Score (x)	p		D			
	Number of pupils (f)	fx	Upper group (f_U)	$f_U x$	Lower group (f_L)	$f_L x$
0	12	0	3	0	9	0
1	13	13	7	7	6	6
2	5	10	5	10	0	0
	30	$\sim fx = 23$	15	$\sim f_U x = 17$	15	$\sim f_L x = 6$

$$p = \frac{23 - 30(0)}{30(2-0)} = 0.38 \qquad D = \frac{17 - 15(0)}{15(2-0)} - \frac{6 - 15(0)}{15(2-0)} = 0.37$$

The results shown in Table 13 indicate that Item 6, the first subjective item (Subj. 1), was a good item in terms of difficulty and ability to discriminate. The *p*-value for Item 6 was 0.38, between 0.30 and 0.70, which made it a very good item in terms of difficulty. However, its D value of 0.37 was a little lower than the ideal value, which implied that it could still discriminate between the high achievers and the low achievers, although some modifications would make it better.

Item 6 was a comprehension item that required pupils to read and understand a linear text and locate the answers to the stem. As it involved reading skills and the ability to pinpoint the key information, it was perhaps unsurprising that the item had a *p*-value of 0.38, making it an excellent item in terms of difficulty although leaning slightly towards being difficult because pupils needed to modify the key from the text in order to construct a grammatically accurate sentence. Its challenge as an item was also reflected in its D value, which was only 0.38 due to the fact that only five pupils from the upper group and none from the lower group scored full marks. Item 6 could be made slightly easier to achieve optimum value for *p* and D.

Four of the pupils interviewed found this item to be easy, as the answer could be lifted from the text with minimal modification, while one of them found the words in the stem too difficult to comprehend and he "could not find the answer in the text." When

asked if this or other items of a similar nature could be challenging to some pupils, a pupil from the High band reasoned that some of his peers "[did] not like reading [and] have problems understanding long sentences," which proved to be a stumbling block in tackling comprehension items.

The teacher experts also agreed that the item was straightforward, as the key word used in the stem (*screamed*), was quite similar to the words used in the text (*screams with fright*), hence providing sufficient hint for most pupils to locate the answer. Also, the structure of the stem did not require much modification; the pupils could just copy the sentence structure used in the stem as part of their full-sentence response and add in the answer found in the text. The teacher experts suspected that the pupils who received zero marks were "not careful with their reading," and those who received one mark were probably not as adept with compound sentence structure and made "grammatical errors when answering."

Table 14

Item 7

Construct	3.3.1 Able to create simple linear texts using a variety of media with guidance
Item	Why is saving money a good habit? _____ _____ [2 marks]

Table 15

Pupils' Responses for Item 7

Score (x)	p		D			
	Number of pupils (f)	fx	Upper group (f_U)	$f_U x$	Lower group (f_L)	$f_L x$
0	7	0	1	0	6	0
1	11	11	6	6	5	5
2	12	24	8	16	4	8
	30	$\sim fx = 35$	15	$\sim f_U x = 22$	15	$\sim f_L x = 13$

$$p = \frac{35 - 30(0)}{30(2-0)} = 0.58 \qquad D = \frac{22 - 15(0)}{15(2-0)} - \frac{13 - 15(0)}{15(2-0)} = 0.3$$

Table 14 presents the results for Item 7, the second subjective item (Subj. 2). It was an item that required HOTS from the pupils, as evident from the Table of Specification (refer to Table 1). While it had a higher *p*-value than Item 6 (0.58), it was still within the range of $0.4 < p < 0.7$, exhibiting characteristics of a good item on the Difficulty Index.

With a D-value of 0.3, Item 7 was a borderline good item (Bichi & Embong, 2018). Although its *p* value was excellent, the D value showed that it needed improvement. Being a HOTS item, pupils had to come up with their own answer from the stem, which acted as a stimulus. Unlike Item 6, the text did not provide the key. Instead it prompted pupils to come up with their own logical responses using critical thinking skills. Normally, pupils from the lower achieving group would struggle to come up with responses that earned full marks. This is reflected in Table 15, where six pupils from the lower group scored zero marks. However, 27% of the responses from the higher ability group merited full marks, showing a gulf in their capabilities.

Feedback from the cognitive interview highlighted that the greatest difficulty for pupils trying to answer HOTS items was expressing their own thoughts in a clear and concise manner. An interesting response from one of the interviewed pupils was, "I understand the question… the question is easy… I know the answer but I don't know how to write it down." This remark was similar to the response in Sariay (2017, p.28): "… I get stuck when I do not have the words to answer." Other pupils also expressed the same opinion, except for one person who stated that he "didn't know what the question was asking." However, the pupils interviewed did not rate the item as difficult, only that providing the desired response was more challenging.

Teacher experts also considered HOTS items to be the most challenging for pupils, especially those with a poor command of English. According to them, pupils are often unable to provide an answer that is clear and straight to the point. One of the teacher

experts commented: "Pupils seem to feel that they need to write a lot when it comes to HOTS questions… They didn't realize that the more they write, the more mistakes they make, [and] the more unclear their answer is." Nevertheless, the teacher experts emphasized the necessity of HOTS items as they are useful in identifying the more able pupils who under normal circumstances will not be too hindered by this type of items and can express themselves fairly well.

Researcher's Reflection

The purpose behind the study was for its findings to act as a catalyst for teacher practitioners to become aware of the potential of CTT as a classroom assessment tool that can raise the quality of test items. While standardized tests go through a long procedure of design and review and hence its quality is rarely disputed, summative assessments in school are often made by the teachers themselves. When it comes to item building, a common practice at school level is to select items that seem appropriate from available workbooks, and to put them together to form a complete test paper. Modification is rarely, if ever done on these items; nor is any kind of analysis or review ever conducted. As a result, there is a possibility that the quality of summative assessments in school sometimes suffers, and at times, they are unable to act as an appropriate indicator of pupils' capability.

In order to add credibility to the study, a mixed-method design was chosen, consisting of psychometric analysis of the data from the test and cognitive interviews with the parties involved. Building the test items according to the table of specification was a real eye-opener, as the items not only had to mirror the general item types in standardized tests but also had to be a fair representation of the constructs listed in the Curriculum Standard. The cognitive interviews with the pupils and teacher experts were also very fruitful, as a lot of insights from a first-person perspective (pupils) and third-person perspective (teacher experts) were gained to complement the quantitative data.

Rumi, a 13th-century Persian poet once said, "Yesterday I was clever, so I wanted to change the world. Today I am wise, so I am changing myself." In order to bring about a wave of change, one must be the first drop in the ocean and change himself. Before conducting this study, one of the researchers was one of the teachers described in the

earlier paragraph. Item building was more of a duty than a means to assess pupils. Although marks were produced from the tests, they served no further purpose than to see if the pupils were scoring higher than on the previous test, and to rank the pupils. Undertaking this research from a new perspective proved to be both challenging and informative. The experience has opened the eyes of the researcher-cum-teacher practitioner as to how teacher practitioners can do so much with the data procured from teacher-made items in order to construct tests and develop test items that can validly, better assess pupils' competence. It is hoped that the study is able to provide some insights on the use of CTT by teacher practitioners without having to resort to complicated CTT or IRT software. By applying knowledge about test construction obtained during their initial teacher education programme, teacher practitioners can develop quality test items, using item analysis to support the process of constructing quality teacher-made tests in the classroom.

Discussion

The focus of this study was 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 discrimination, and iii) functioning or non-functioning distractors.

By definition, quality items are items that have ideal values in their difficulty index, *p* and discrimination index (D). The difficulty index *p* should be between 0.3 and 0.7, whereas the discrimination index D should be 0.4 or higher. In general, most items in this study were either too easy or too difficult, except for Item 3 (Idioms) and Item 7 (HOTS). On the other hand, more items fared better in terms of D, with Item 1 (Vocabulary), Item 2 (to-infinitive), Item 3 (Idioms) and Item 5 (Spelling) achieving ideal scores in the Item Discrimination Index. Of the five items that obtained ideal D values, Item 1, Item 2 and Item 5 leaned towards the extreme in their *p*-values, which were 0.2, 0.8 and 0.73 respectively. This meant that minor modifications were needed for them to be considered quality items. However, even items which did not score within Bichi and Embong's (2018) acceptable range were still within the wider range proposed by other researchers such as Mukherjee and Lahiri (2015) who proposed a range of 0.2 to 0.9. The best item on the test was Item 3 (Idioms)

which was within the acceptable range for *p* and had a high D value. It was surprising that despite the negative perception towards subjective items, both short answer items had acceptable *p* and D values; even Item 7, which tested pupils' HOTS, actually fared quite well on the Difficulty Index. Nevertheless, some findings in this study were inconsistent with Mukherjee and Lahiri's (2015) view that items with the ideal *p*-value of between 0.4 and 0.6 will have a Discrimination Index of 0.4 and above. While Item 3 ($p = 0.5$; D = 0.46) lent support to the hypothesis, Item 7 showed otherwise—with an ideal *p*-value of 0.58, Item 7 scored only 0.3 for D, which pointed to a need for item modification.

Slight modifications could also be done to Items 2 and 5, which had NFD and *p*-values that were slightly higher than 0.7. By introducing more appealing distractors in place of the NFDs, their *p*-values of 0.8 and 0.73 respectively, could fall within the ideal range of $0.3 < p < 0.7$. The only item that possibly needed a thorough revision was Item 4 ($p = 0.27$, D = 0.13). Its *p*-value indicated that it was a little too difficult as an item, and D signified its inability to discriminate the pupils in any capacity.

A few conclusions could also be derived from the responses to the cognitive interview and the teacher expert input. Firstly, the wording of the stem should be carefully chosen so that it does not pose too much difficulty in comprehension. Zimmaro (2004) stated that directions in the stem and the wording should be clear as stems that are unclear due to high-level vocabulary or ambiguous due to poor language structure obstruct students' ability to choose the key. Similarly, the choice of options should also be clear. Sariay (2017, p.23) offers students' insight on being confused when faced with almost identical options: "I sometimes feel that two options are identical. You might get confused between them. But you have to choose one of them even though the other one is similar to the one you have picked." Item 4 posed such a challenge, as the interviewed pupils reported being confused with some of the options being too close in definition to one another.

Another reason for the meticulous design of stems and options is to cut down on the instances of guessing. While guessing can never be totally eradicated from tests, the outcome does not contribute to the understanding of pupils' competence. Participants interviewed

by Sariay (2017, p.23) highlighted how common guessing was in answering multiple choice items: "… if you do not know anything about the answer, but pick out one answer that then turns out to be correct, you get the mark"; "Most of the time, you are given the answers, and it is just as easy to guess or to know them." In their study on the effect of guessing on test scores, Ubulom and Amini (2012) also concluded that "guessing was the main factor responsible for the error in the test scores" (p. 19) and that it should be discouraged by teachers and examiners. While instructions may be given to pupils to not adopt guessing when answering, more needs to be done on the design of items. When pupils are able to understand the requirements of the stem and feel that they are able to answer it correctly with some thought, they are less inclined to resort to hasty methods.

However, this study was done using only a small sample size of thirty pupils as part of a larger study. Further studies will be needed to consolidate the claims made from the data analysis above.

CONCLUSION

The intention behind this study was to have a better understanding of the quality of test items used in the classroom so that they can validly measure pupils' competence in English. As a lot of factors hinge on the outcome of these assessments (i.e., placement in a better class, enrolment in 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.

This study has further consolidated the usefulness of Item Difficulty Index, *p*, Item Discrimination Index, D, and Distractor Analysis 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 approaching zero or negative for either index should be revised, modified or rejected. The values presented not only provide conclusive evidence of the characteristics of the items, but also serve as a useful platform should further analysis be

required to identify the factors that contribute to a poorly designed item, 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.

For further field studies of a similar design, a few modifications can be done to make the whole process more effective, thus enhancing the reliability of the data collected: i) have a special room or hall to emulate the actual standardized testing scenario that pupils are familiar with (Cook & Beckman, 2006), with each pupil having a table and seat of their own and spacing between them to prevent cheating and for comfort, ii) while no time limit is advisable, participants need to be on task at all times to ensure validity and reliability, iii) tests should be administered in a way as closely resembling a real standardized test as possible, with as little indication that the whole process is more than a test (Cook & Beckman, 2006). 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.

It needs to be noted that item analysis and subsequent interpretations are done on an item-to-item basis. However, with regards to summative assessments, it is widely accepted that there should be a balance of easy and difficult items for results to be reliable. The novelty of this study is that the examination of the psychometric properties of test items was supported by qualitative expert judgements by classroom practitioners. In addition, it also considered the students' perspective on the content of test items. One limitation of the study would be the small sample size.

In conclusion, item analysis using CTT should become a common practice among teacher practitioners because of its importance in providing vital information towards producing good quality test items that are valid and reliable. Data from item analysis can be a valuable ally in the classroom, as teacher practitioners can make more informed judgements of their pupils' ability, either to guide their classroom teaching or to construct better test items to achieve the objective of *assessment for learning*, instead of solely *assessment of learning*.

ACKNOWLEDGEMENT

This research received no specific grant from any funding agency.

REFERENCES

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing*. (Abridged edition). Boston, MA: Allyn and Bacon.

Bichi, A. A. (2015). Item analysis using a derived science achievement test data. *International Journal of Science and Research (IJSR), 4*(5), 1656- 1662.

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

Black, P., & William, D. (2005). Lessons from around the world: How policies, politics and cultures constrain and afford assessment practises. *The Curriculum Journal, 16*, 249–261.

Brown, G. T., & Hattie, J. (2012). The benefits of regular standardized assessment in childhood education. In S. Suggate & E. Reese (Eds.), *Contemporary Debates in Childhood Education and Development* (pp. 287-292). London: Routledge.

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *American Journal of Medicine, 119*, 166.e7–166. e16.

Creswell, W. J. (2013). *Research design; qualitative, quantitative, and mixed approach*. Yogyakarta: Pustaka Pelajar.

Duff, P., & Anderson, T. (2016). Case study research. In J.D. Brown & C. Coombs (Eds.), *Cambridge Guide to language research* (pp. 112-118). Cambridge, UK: Cambridge University Press.

Fitzpatrick, A. R. (1983). The meaning of content validity. *Applied Psychological Measurement, 7*, 3-13.

Fives, H., & DiDonato-Barnes, N. (2013). Classroom test construction: The power of a table of specifications. *Practical Assessment, Research and Evaluation, 18*, 2–7.

Gajjar, S., Sharma, R., Kumar, P., & Rana, M. (2014). Item and test analysis to identify quality multiple choice questions (MCQs) from an assessment of medical students of Ahmedabad, Gujarat. *Indian Journal of Community Medicine*, *39*(1), 17-

20.

Haliza, I. (2017). *Item analysis of English paper 1 (EPI) of 2014 UPSR trial examination using Rasch measurement model* (Unpublished doctoral thesis). Universiti Pendidikan Sultan Idris, Malaysia.

Hambleton, R. K., & de Jong, J. H. A. L. (2003). Advances in translating and adapting educational and psychological tests. *Language Testing, 20*(2), 127-134.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology 1*(1), 1-30. Retrieved from http://www.testpublishers.org/journal01.htm

Kehoe, J. (1994). Basic item analysis for multiple-choice tests. *Practical Assessment, Research, and Evaluation*, *4*(4), Article10.

Koçdar, S., Karadağ, N., & Şahin, M. D. (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(4),* 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), 1-11.

Mehta, G., & Mokhasi, V. (2014). Item analysis of multiple-choice questions - an assessment of the assessment tool. *International Journal of Health Sciences and Research, 4(7)*, 197-202.

Mokshein, S. E. (2019). The use of Rasch measurement model in English testing. *Cakrawala Pendidikan*, *38*(1), 16-32.

Mogapi, M. (2016). Examinations wash back effects: Challenges to the criterion referenced assessment model. *Journal of Education and e-Learning Research, 3*(3), 78-86.

Norafizah, B. M. (2018) The washback effect of Primary School Evaluation Test (UPSR) on teaching and learning: A case study of an English teacher in Kuala Terengganu, Malaysia. *International Research Journal of Education and Sciences (IRJES), 2*(2), 15-18.

Mons, N. (2009). *Theoretical and real effects of standardised assessment*. Retrieved from https://pdfs.semanticscholar.org/c252/4fa43a1b7d250d5700b842af1c002fde0ee2.pdf

Moodley, V. (2015). Visual literacy in high-stakes testing: Implications of washback for language teachers. *Literacy Information and Computer Education Journal*, *6*(4), 2054-2062.

Mukherjee, P., & Lahiri, S. K. (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*(12), 47-52.

Mullis, I., & Mullis, A. (2003). *PIRLS 2001 International Report*. Retrieved from http://pirls.bc.edu/pirls2001i/PIRLS2001_Pubs_IR.html.

Mullis, I., Martin, M., Kennedy, A., & Foy, P. (2007). *PIRLS 2006 International Report*. Retrieved from http://pirls.bc.edu/pirls2006/intl_rpt.html.

National Council for Curriculum and Assessment (NCCA). (2005). *Supporting assessment in schools: Standardised testing in compulsory schooling*. Dublin: NCCA.

Pande, S. S., Pande, S. R., Parate, V. R., Nikam, A. P., & Angrekar, S. (2013). Correlation between difficulty & discrimination indices of MCQs in formative exam in physiology. *South-East Asian Journal of Medical Education, 7*(1), 45-50.

Sariay, M. (2017). Teachers' and students' perceptions of multiple-choice and open-ended questions, along with the GSCE system. (Unpublished master's dissertation). School of Education and Lifelong Learning, University of East Anglia, UK. DOI: 10.13140/RG.2.2.10395.77608.

Salkind, N. J. (Ed.). (2010). *Encyclopedia of research design, Vol 1*. Thousand Oaks, CA: Sage.

Singh, A. N., Matson, J. L., Mouttapa, M., Pella, R. D., Hill, B. D., & Thorson, R. (2009). A critical item analysis of the QABF: Development of a short form assessment instrument. *Research in Developmental Disabilities*, *30*(4), 782-792.

Sireci, S. G., & Geisinger, K. (1995). Using subject-matter experts to assess content representation: An MDS scaling. *Journal of Applied Measurement in Education*, *19*(3), 241-255.

Tayeb, Y. A., Aziz, M. S. A., & Ismail, K. (2018). Predominant washback of the general secondary English examination on teachers. *International Journal of Engineering & Technology*, *7*(21), 448-456.

Ubulom, W. J. & Amini, C. M. (2012). Determining the effect of guessing on test scores. *Mathematical Theory and Modelling, 2*(12), 16-21.

van der Linen, W. J., & Hambleton, R. K. (2013). *Handbook of modern item response theory*. New York: Springer.

Veloo, A., & Awang-Hashim, R. (2016). *Teori ujian dan pentaksiran pendidikan*. Sintok, Kedah: UUM Press.

Yoong, Y. L., Lee, T. E., Kanagamani, K. (2015). *English Year 6 Sekolah Jenis Kebangsaan*. Kuala Lumpur: Percetakan Rina Sdn. Bhd.

Zainal, Z. (2007). Case study as a research method. *Jurnal Kemanusiaan*, *9,* 1-6.

Zimmaro, D. M. (2004). *Writing good multiple-choice exams*. Faculty Innovation Center, University of Texas, Austin, USA. Retrieved from https://facultyinnovate.utexas.edu/sites/default/files/writing-good-multiple-choice-exams-fic-120116.pdf