Available online at IJEE (Indonesian Journal of English Education) Website:

http://journal.uinjkt.ac.id/index.php/ijee

ANALYSIS OF A RESEARCH INSTRUMENT TO MAP ENGLISH TEACHERS' PROFICIENCY

Siti Mina Tamah, Anita Lie
Widya Mandala Catholic University Surabaya, Indonesia
(mina@ukwms.ac.id)

Received: 28th March 2019; Revised: 27th May 2019; Accepted: 28th June 2019

ABSTRACT

Teachers' English proficiency can be measured by designing a research instrument in a form of test. The devised test must fulfill the requirement of a good test. This article is aimed at discussing item analysis centering on multiple choice questions used to measure the proficiency of Indonesian High School teachers involved in English instruction. The first set of syllabus oriented test is tried out to 20 subjects, and the second set – general English oriented – to 28 subjects. The test analysis indicates the item difficulty indices range from .20 to 1 for the first set and .07 to .89 for the second set. With regard to item discrimination analysis, the study finds the *d* values range from -0.33 to 1.0 for the first set, and -0.11 to .78 for the second set. It is found that the whole test has 'average' level of difficulty and is 'good' at discriminating between high and low achieving test takers; to be used for the actual research, a revision of the test is done to eliminate the 'bad' items.

Key Words: item analysis; test; difficulty level; discrimination power; English proficiency; teacher

ABSTRAK

Kecakapan bahasa Inggris guru dapat diukur dengan merancang instrumen penelitian dalam bentuk tes. Tes yang dirancang harus memenuhi persyaratan tes yang baik. Artikel ini bertujuan membahas analisis soal yang berpusat pada pertanyaan pilihan ganda yang digunakan untuk mengukur kemahiran guru-guru SMA Indonesia yang terlibat dalam pengajaran Bahasa Inggris. Tes set kesatu yang berorientasi silabus diujicobakan pada 20 subjek. Set kedua - berorientasi Bahasa Inggris umum - diujicobakan ke 28 subjek. Analisis tes menunjukkan bahwa indeks kesulitan soal berkisar dari .20 hingga 1 untuk set pertama dan .07 hingga .89 untuk set kedua. Terkait analisis diskriminasi item, studi ini menemukan bahwa nilai D berkisar dari -0,33 ke 1,0 untuk set pertama, dan -0,11 hingga 0,78 untuk set kedua. Ditemukan bahwa keseluruhan tes memiliki tingkat kesulitan 'rata-rata' dan 'baik' dalam membedakan antara peserta tes berprestasi tinggi dan rendah. Untuk digunakan dalam penelitian aktual, revisi tes dilakukan dengan menghilangkan soal 'buruk'.

Kata Kunci: *analisis soal; uji; tingkat kesulitan; kekuatan diskriminasi; kemahiran bahasa Inggris; guru*

How to Cite: Tamah, S. M., Lie, A. (2019). Analysis of a Research Instrument to Map English Teachers' Proficiency. *IJEE (Indonesian Journal of English Education), 6*(1), 48-64. doi:10.15408/ijee.v6i1.11888

P-ISSN: 2356-1777, E-ISSN: 2443-0390 | DOI: http://doi.org/10.15408/ijee.v6i1.11888

INTRODUCTION

Teachers' subject matter mastery and teaching competence will affect the attainment of instructional objectives. Their skills and knowledge have been highlighted as a key component associated with clear objectives for student learning and accomplished teaching (OECD, 2005 cited in Caena, 2011). Teacher quality is in fact the key to enhance students' achievement (Barber & Mourshed, 2007; Chetty, 2011; Rasmussen & Holm, 2012; Harjanto et al., 2017). It is, therefore, crucial that research on teacher competence be conducted.

With the increasing importance of English as a language of global communication, the quality of English instruction in schools has drawn research interest particularly in countries where English is not the lingua franca. A number of studies on teachers' English proficiency have been conducted. Author (20xx) urged that to set advanced competencies in the English curriculum, Indonesian teachers' English proficiency first had to be improved. Tsang (2011) investigated to what extent 20 primary school English teachers in Hong Kong were aware of English metalanguage and found the need for regular or systematic use of metalanguage among school teachers. Sharif (2013) was concerned that limited English proficiency of teachers distorted students' understanding of the content taught. Othman and Nordin (2013) studied the correlation between the Malaysian University English Test (MUET) and academic performance of English teacher education students. Earlier, Lee (2004) criticized the use of the high-stake MUET as a driver to improve English proficiency and suspected that the very traditional approach to teaching reading with the focus on discreet skills may have been the result of teachers' preoccupation with getting their students to pass MUET. More recently, Nair and Arshad (2018) examined the discursive construction of Malaysian English language teachers in relation to the Malaysian Education Blueprint action plan from 2013 to 2015 and argued for ways to help teachers achieve the desired proficiency and make changes to existing classroom practices that are aligned with the government agenda.

The competence of Indonesian teachers of English has also been the focus of a number of studies. A study (Lengkanawati, 2005) examining the English proficiency of teachers in West Java used a TOEFL-equivalent test and found that the majority of the teachers did not demonstrate a satisfactory proficiency level. Aniroh (2009) discussed the need for ESP teachers to

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://doi.org/10.15408/ijee.v6i1.11888
49-64
P-ISSN: 2356-1777, E-ISSN: 2443-0390 | This is an open access article under CC-BY-SA license

have a set of qualities, one of which is proficiency in English but she did not further elaborate on the proficiency issue. Anugerahwati and Saukah (2010) studied professional competence of English teachers in Indonesia and presented a profile of exemplary teachers based on qualitative descriptions of the four research subjects. They argued that satisfactory competence in English

"may seem to be taken for granted by many people other than the English teachers themselves. They tend to put a lot of pressure on themselves to excel in the subject matter. Actually this competence is already guaranteed by the requirement that a teacher has to have an S1 or D-IV degree qualification, and as such, it is understandable that other people view subject matter competence as something given by their formal education (p. 55)."

The guarantee of subject matter competence through the teachers' formal education is still very much debatable as graduate competence standards are still yet to be established and enforced in English teacher education.

Assessing English teachers' competence remains a salient issue. Soepriyatna (2012) investigated and assessed competence of high school teachers of English in Indonesia and set three dimensions of English language competence domain (language skills, linguistic, and sociocultural), two dimensions of content knowledge domain (text types and grammar points), and seven dimensions of teaching skills domain (objectives, material development, learning management, teaching techniques, learning styles, learning strategies, and qualities of an engaging teacher). He developed performance tasks to assess the twelve competence dimensions. The language proficiency covered in the first two domains is addressed in performance indicators statements such as "uses vocabulary correctly and appropriately" and "maintains grammatical accuracy." Soepriyatna did not address how those indicators can be determined reliably. A test specifically constructed to assess the English proficiency of high school teachers is yet to be developed in Indonesia. The Ministry of Education has been administering annual Teacher Competency Test for all teachers as part of the certification process. The online test comprises of subject area and pedagogy items. Therefore, it does not specifically address language proficiency. Furthermore, there have been concerns that the test was not adequately constructed (Prasetyo, 2017; Putra, 2017). In line with these concerns, it is reported that of the eight

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://doi.org/10.15408/ijee.v6i1.11888
P-ISSN: 2356-1777, E-ISSN: 2443-0390 | This is an open access article under CC-BY-SA license

national education standards, three standards—teacher standard, learning resources and facilities standard, and graduate competence standard—are the weakest. Toni Toharudin, chair of the National School Accreditation Council, urges that the government should play a more concrete role in enhancing teacher competence and distributing high-quality teachers equally in the regions (Eln, 2018).

An essential requirement for a test to be employed especially for conveying teachers' proficiency is that the test should be a good one for a research instrument. The test devised ought to be valid and reliable. One extensively used way to perform as the step to fulfill this requirement is analyzing the test items—Gronlund (1982:101) simply puts it "studying the students' responses to each item". Plakans and Gebril (2015) assert that item analysis is a checking procedure to see that test questions are at the right level of difficulty. It is also a procedural entity to check that test questions distinguish test takers appropriately.

Test item analysis based on classical measurement theory functions as an analysis tool to measure *item difficulty* index, *item discrimination* index, and *distractor effectiveness* (Hughes, 1989). Classical test theory has less demand on the number of test takers whose answers will be the ones to analyze. This theory is consequently more practical since no formal training is needed prior to analysis undertaking. The item analysis is more easily performed manually—by taking, for instance, a calculator-assisted analysis or by using a simple program in a computer. The weakness of this theory is that there is an interdependency between test takers and item difficulty level.

Item response theory appears as a response to the weakness of classical measurement theory. Based on this item response theory – also called "Rasch analysis" (Hughes, 1989: 163), test item difficulty is ideally constant, taking no notice of whichever group is being tested. This theory performs item analysis by calculating difficulty index only (commonly termed as a one-parameter logistic model), item difficulty index and item discriminating index (prevalently termed as a two-parameter logistic model), and difficulty index, discriminating power, and speculation element (labelled a three-parameter logistic model). The more elements to be analysed, the more test takers will be engaged for their answers to analyse. In conclusion, classical test theory is more practical than item response theory. Classical test theory is more easily conducted as it does not require lots of test takers. It

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://doi.org/10.15408/ijee.v6i1.11888

can be applied more effortlessly by teachers or researchers.

This article presents the result of test item analysis. The analysis is delimited to item difficulty and item discrimination. The analysis is carried out to contribute to revealing the reliability of an instrument to measure high school teachers' English proficiency.

Difficulty level is most often paired with other terms having the same meaning like *difficulty index, index of item difficulty*, or *facility value* as used by Hughes (1989), Brown (2004), Brown and Abeywickrama (2010), or *Item Facility* as used by Brown (1996). They all refer to the same construct.

Difficulty index is a score indicating whether a test item is difficult or easy. The level of item difficulty can be explained by the percentage of the test takers who answer a test item correctly. Gronlund (1982) points out that it is the percentage of answering the items correctly. Brown (1996: 64-65) similarly asserts that it is "a statistical index used to examine the percentage of students who correctly answer a given item."

Therefore, difficulty index which is symbolized as P value is one which is obtained after a measurement has been done on students who are able to answer the item correctly. The difficulty index functions as an indicator for test makers to know the quality of their test by determining whether the test is difficult or easy. Difficulty item analysis will reveal students' ability to the problem being analyzed.

With regard to good P value, the majority of test analysts would argue for the level of 'sufficient' or 'medium' (P value of 0.50) for a good test. Meanwhile, Hughes (1989: 162) claims, "There can be no strict rule about what range of facility values are to be regarded as satisfactory. It depends on what the purpose of the test is … The best advice … is to consider the level of difficulty of the complete test."

Discriminating power also has several terms like *discrimination index, item discrimination, level of discriminating,* and *index of discriminating*. They all refer to the same construct.

Some literature labels index of item discriminating power with the letter 'D', while some others use two letters 'DI'. This D value or DI value reveals the discrimination power of a test item. To be more specific, it indicates "the degree to which an item separates the students who performed well from those who performed poorly" (Brown, 1996: 68) therefore it allows test developer to contrast the performance of the high achievers and low achievers.

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://doi.org/10.15408/ijee.v6i1.11888
P-ISSN: 2356-1777, E-ISSN: 2443-0390 | This is an open access article under CC-BY-SA license

An item discrimination index of 1.00 is considered "very good as it indicates the maximum contrast between the upper group and lower groups of students—that is, all the high-scoring students answered correctly and all the low-scoring students answered correctly." (Brown, 1996: 68).

In light of the need for better quality of English instruction in Indonesia, our research team identified the research gap of mapping the content knowledge competence of English language teachers in Indonesia high schools and assessing their English proficiency. This study is a part of a bigger research project funded in 2018 by the Indonesian Ministry of Research, Technology and Higher Education to conduct a mapping of high school teachers of English. This article presents the construction of a test to assess their English proficiency as a preliminary step before assessing their English language teaching competences.

METHOD

As previously mentioned in the background, the test constructed by the research team will be used as a research instrument to map the English proficiency of high school teachers in Indonesia.

Design

This study which centers on item analysis is quantitative in nature. The statistical formula prevalently employed include the difficulty and discriminating power values.

In order for the test to be an accurate measure of what it is supposed to measure, and also more importantly in order that the test does not result in "a harmful backwash effect" (Hughes, 1989: 22-23), or in order for a test to be an effective strategy to determine the content of Multiple Choice questions (Plakans & Gebril, 2015), a test specification is prepared. A test specification is responsible for "the construct framework for operationalizing the test design through subsequent item development" (Kopriva, 2008: 65). Despite the counter-argument stating that Multiple Choice questions do not adequately simulate how language is used in real life, Multiple Choice questions occasionally provide better coverage of content than the nowadays performance based assessment (Plakans & Gebril, 2015). Furthermore, in spite of its drawbacks, Multiple Choice format offers efficiency of administration, particularly when it involves a large number of test-takers. These particular reasons lead the research team to include Multiple Choice type.

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://doi.org/10.15408/ijee.v6i1.11888

Subjects

There were 20 and 28 subjects involved in the first and second tests respectively. Some subjects consisted of pre-service teachers/fresh graduates of English Department of Teacher Training Faculty; they were not involved in the teaching field yet. Some other subjects were completing their last semester at the English Department of Teacher Training Faculty; they were finishing their thesis writing. The try-out subjects excluded those teachers who would be engaged in the following research.

Instrument

The test was developed to cover three main categories: the syllabus-oriented, the general English (grammar and reading comprehension), and essay. There were three test types utilized: Multiple Choice, Cloze test, and Writing. All together 65 items were developed. This paper presents only the analysis of 50 Multiple Choice items (the other test types – Cloze test amounting to 15 items and Writing test – are not analysed). Among the seven Multiple Choice formats (Haladyna, Downing, & Rodriguez, 2002), the one used in this study was Conventional MC. The first test set which consists of 30 items is presented in Table 1.

The test specification guiding the construction of the 30 items in the first test set is taken from the currently used 2013 English Curriculum for high school in Indonesia.

The second test set which is general English consists of 20 items covering 10 Grammar and 10 Reading Comprehension items as presented in Table 2 and Table 3 respectively.

Table 1. Table of Test Specification (Syllabus Oriented)

Basic Competence	Items Prepared
1. Implement social function, text structure, and language feature ... involving giving and asking personal (family and relative) information based on the appropriate context (Focus on *pronoun: subjective, objective, possessive*).	1. My mother's brother in-law is my … *aunt / uncle / cousin / grandfather*
2. Implement social function, text structure, and language feature ... involving giving and asking information related to future intention based on the appropriate context (Focus on *be going to, would like to*).	2. Shinta … married next year. *is going to get / would like to get / got / are getting* 3. Doni … a new job. *getting / would like to get / have got / are getting*

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://doi.org/10.15408/ijee.v6i1.11888
P-ISSN: 2356-1777, E-ISSN: 2443-0390 | This is an open access article under CC-BY-SA license

Basic Competence	Items Prepared
3. Distinguish social function, text structure, and language feature ... involving giving and asking information related to famous historical building based on the appropriate context (Focus on e.g. adverbs *quite*, *very*).	4. Borobudur Temple is … beautiful. *quite / quiet / quitely / quietly*
4. Implement social function, text structure, and language feature ... involving giving and asking information related to past event based on the appropriate context (Focus on e.g. *simple past tense vs present perfect tense*).	5. He … his leg in a car accident last year. *is breaking / broke / has broken / breaks* 6. I left home at 7 a.m. and I … here at 1 p.m. *am getting / got / has gotten / get* 7. I cannot go out because I … my work yet. *am not finishing / didn't finish / haven't finished / don't finish*
5. Distinguish social function, text structure, and language feature ... involving recount texts based on the appropriate context (Focus on e.g. transitional words like *first, then, after that, before, when, at last*.	8. … the movie ends, we head out for a late night snack. *Before / Then / After that / When*
6. Distinguish social function, text structure, and language feature ... involving narrative texts based on the appropriate context (Focus on e.g. simple past tense, past continuous).	9. Once upon a time, there was a little boy, who was poor, dirty, and smelly, … into a little village. *comes / is coming / coming / was coming* 10. Kancil … quick-witted, so that every time his life was threatened, he managed to escape. *was / were / is / be*
7. Implement social function, text structure, and language feature ... involving giving and asking information related to suggestion and offering based on appropriate context (Focus on e.g. *model auxiliary should* and *can*).	*11.* Giving suggestion: *Can I help you? / I can walk that far. / I should go. / You should study harder.* 12. Offering something: *Should I go to your house tonight? / Can I help you? / You can do it. / He should go to the doctor today.*
8. Implement social function, text structure, and language feature ... involving giving and asking information related to giving opinion based on appropriate context (Focus on e.g. *I think, I suppose).*	13. Giving opinion: *In my opinion, she's pretty. / Can you give me your opinion? / He is thinking about her everyday. / He should go.*
9. Distinguish social function, text structure, and language feature ... involving actual issues based on the appropriate context (Focus on transitional words like *therefore, consequently*).	14. Madeline is rich, …, her cousin is poor. *however / otherwise / so / therefore* 15. The students didn't study. …, they failed the course. *however / otherwise / so / therefore*

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://doi.org/10.15408/ijee.v6i1.11888

P-ISSN: 2356-1777, E-ISSN: 2443-0390 |

Basic Competence	Items Prepared
10. Implement social function, text structure, and language feature ... involving giving and asking information related to events or activities with the focus not on the doers based on appropriate context (Focus on e.g. passive voice).	*16.* What is the passive voice of this sentence: Somebody stole my pen. *My pen has been stolen. / My pen was stolen. / My pen had stolen by somebody. / My pen is stolen.* 17. What is the passive voice of this sentence: Have you finished the report? *Has the report been finished? Has the report finished? / Has the report finished by you? Has the report been finish?* 18. This experience will never … by me. *forget / forgot / be forgot / be forgotten* 19. The girl … by the boy. *was tease / tease / was teased / teases* 20. Choose the correct sentence: *Her duty done by her. / Was her duty done by her? / Did she done her duty? / She was done her duty.*
11. Implement social function, text structure, and language feature ... involving giving and asking information related to cause-effect based on appropriate context (Focus on e.g. *because of, due to*).	21. His defeat was … the lottery issue. *due to / because / since / thanked to* 22. The crash occurred … the erratic nature of the other driver. *due / because / because of / thanked to*
12. Distinguish social function, text structure, and language feature ... involving nature or social issues based on the appropriate context (Focus on transitional words like *if –then, so, as a consequence, since,* and passive voice).	23. The snowfall came … the effects of El Nino. *as a consequence / due / since / because of* *24.* Serious threats … by genetic engineering. *is posed / will be posed / can be posed / pose* 25. Deforestation … some rainforest ecosystems.*has been destroyed / have been destroyed /has destroyed / have destroyed*
13. Distinguish social function, text structure, and language feature ... involving news based on the appropriate context (Focus on Tenses like Past tense, Present Perfect Tense, Future Tense, passive voice, direct-indirect speech, preposition).	*26.* President Joko Widodo … to depart for Surakarta, Central Java, on Tuesday evening to pay his last respects to his in-law, Didit Supriyadi, who passed away in the morning. *set / sets / is set / are set* *27.* He asked her … him a cup of water. *give / giving / to give / gave* *28.* She told the boys … on the grass. *not to play / don't play / not play / doesn't play* *29.* Who are you waiting … *by / in / for / at* 30. Where's Martin? Is he … work today? *for / on / in / at*

Table 2 Table of Test Specification (General English; Grammar)

	Grammar Category	Items Prepared
1.	Verb; Tense (Past Tense)	Your niece used to help you quite often, … ? didn't she / wouldn't she / doesn't she / hadn't she
2.	Verb; Tense (Future Tense)	If Anton . . . with us, he would have had a good time. would join / had joined / would have join / joined
3.	Verb; Subjunctive	Honestly, I'd rather you … anything about it for the time being. Do / don't / didn't do / didn't
4.	Verb;	Since he isn't answering his telephone, he . . .

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://doi.org/10.15408/ijee.v6i1.11888
P-ISSN: 2356-1777, E-ISSN: 2443-0390 | This is an open access article under CC-BY-SA license

	Grammar Category	Items Prepared
	Modal Auxiliary	must have left / need have left / should have left / can have left
5.	Verb; Tense (Perfect Tense)	We were hurrying because we thought that the taxi . . . had already came / had already come / has already came / have already coming
6.	Pronoun (Object pronoun)	Let you and … agree to straighten out our own problems. I / me / myself / my
7.	Pronoun (Relative Pronoun)	If you had told us earlier … he was, we could have introduced him at the meeting. Who / whom / which / whoever
8.	Pronoun (Relative Pronoun)	The notebooks … Ben had lost at the bus station were returned to him. what / which / who / whose
9.	Pronoun (as object of a sentence)	They didn't seem to mind … TV while they were trying to study. my watching / me watching / that I watch / me to watch
10.	Verb; Tense (Passive Voice)	My pictures … until next week. won't develop / don't develop / aren't developing / won't be developed

Table 3 Table of Test Specification (General English-Reading Comprehension)

Barret Taxonomy	Items prepared	
Reorganization	1.	Which of the following is the best title for this passage? What the Eye Can See in the Sky / Bernard's Star / Planetary Movement / The Ever-moving Stars
Inferential Comprehension	2.	The expression "naked eye" in line 1 most probably refers to … a telescope / a scientific method of observing stars / unassisted vision / a camera with a powerful lens
Literal Comprehension	3.	According to the passage, the distances between the stars and Earth are … barely perceptible / huge / fixed / moderate
Inferential Comprehension	4.	The word "perceptible" in line 5 is closest in meaning to which of the following? Noticeable / Persuasive / Conceivable / Astonishing
Inferential Comprehension	5.	In line 6, a "misconception" is closest in meaning to a(n) … idea / proven fact / erroneous belief / theory
Literal Comprehension	6.	The passage states that in 200 years Bernard's star can move … around Earth's moon / next to Earth's moon / a distance equal to the distance from Earth to Moon / a distance seemingly equal to the diameter of the Moon
Inferential Comprehension	7.	The passage implies that from Earth it appears that the planets … are fixed in the sky / move more slowly than the stars / show approximately the same amount of movement as the stars / travel through the sky considerably more rapidly than the stars
Inferential Comprehension	8.	The word "negligible" in line 8 could most easily be replaced by … negative / insignificant / rapid / distant

Barret Taxonomy	Items prepared	
Inferential Comprehension	9.	Which of the following is NOT true according to the passage? Stars do not appear to the eye to move. / The large distances between stars and the earth tend to magnify movement to the eye. / Bernard's star moves quickly in comparison with other stars. / Although stars move, they seem to be fixed.
Inferential Comprehension	10.	The paragraph following the passage most probably discusses … the movement of the planets / Bernard's star / c. the distance from Earth to the Moon / why stars are always moving

Data Collection

The test was tried out using two ways of administration: on-line version (making use of google form) and off-line version commonly known as paper-based test. A week period of test administration was given to the subjects who did the timed on-line version. A 60-minute classroom session at a university in Nusa Tenggara Timur province in East Indonesia was administered off-line due to the poor internet connection.

Data Analysis Procedure

The result of the test try out having been collected is analysed quantitatively using two types of statistical formula. The first prevalently employed formula to find difficulty level is taken from Gronlund (1982).

$$P = R/T$$

Where P = the percentage who answered the item correctly
R = the number who answered correctly
T = the total number who tried the item

The second employed formula to calculate the index of discriminating power is taken from Brown (1996).

$$D = IF\ upper - IF\ lower$$

Where D = item discrimination power for an individual item
IF upper = item facility or p value for the upper group on the whole test
IF lower = item facility or p value for the lower group on the whole test

FINDINS AND DISCUSSION

The analysis on the first set of test indicates that the item difficulty indices (P value) range from .75 to 1.00 for **easy** items which amounts to 33.3%, .35 to .70 for **average** items amounting to 56.7%, and .20 to .25 for **difficult** items reaching only 10%, the smallest percentage (See Figure 1). It is revealed that the average items occupy the highest percentage rank. Calculating the average percentages of difficulty level for the test with regard to the syllabus oriented test – the first test set, the writer finds it to be .64 revealing **average** level of difficulty.

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://doi.org/10.15408/ijee.v6i1.11888
P-ISSN: 2356-1777, E-ISSN: 2443-0390 | This is an open access article under CC-BY-SA license

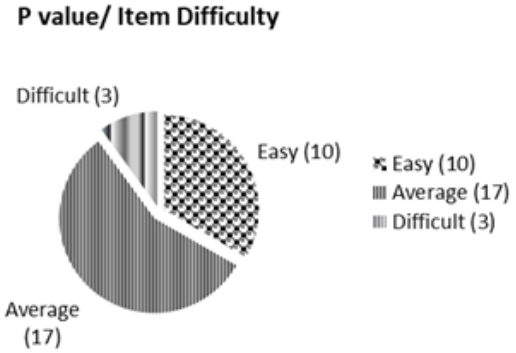

Figure 1. Item Difficulty of Syllabus-Oriented Items

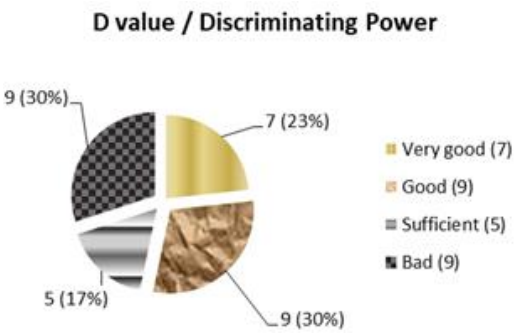

Figure 2. Discriminating Power of Syllabus-Oriented Items

Meanwhile as displayed in Figure 2 below, the indices of discriminating power range from -0.33 to 1.0. Having D value of .83 – 1, seven (23.3%) items are 'very good' at discriminating between the high achieving test takers and the low ones. Having D value of .5 to .67), nine (30%) items are 'good' at discriminating between the high and low achieving test takers. Five (16.7%) items have the D value of .33 indicating they are 'sufficient' in discriminating between the high and low achieving test takers. Nine (30%) items belong to 'bad' ones They cannot distinguish between the two groups well. One of those nine items has negative value (-0.33). The average index of discriminating power for the test with regard to the syllabus oriented test – the first test set – is .43 indicating 'good' discriminating power).

The analysis to the second set of test – as seen in Figure 3 – indicates that the item difficulty indices (P value) range from .79 to .89 for **easy** items which amount to 15%, and .68 to .32 for **average** items amounting to 75%. The item difficulty indices (P value) range from .07 to .29 for **difficult** items reaching 10%, the smallest percentage of the total. It is explicitly revealed that the average items occupy the highest percentage rank. Calculating the average percentages of difficulty level for the test with regard to the general English oriented test – the second test set, the writer finds it to be .55 revealing **average** level of difficulty.

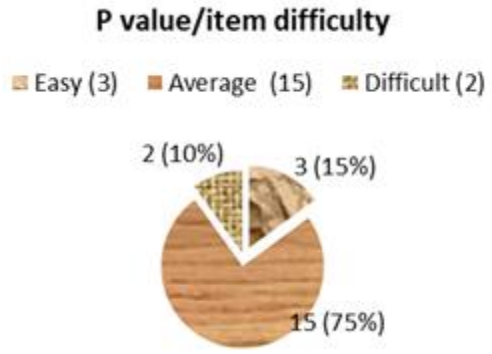

Figure 3. Item Difficulty of General English-Oriented Items

Meanwhile as seen in Figure 4 the indices of discriminating power range from -0.11 to .78 Having D value of .78, only one (5%) item is 'very good' at discriminating between the high achieving test takers and the low ones. Having D value of .44 - .67, ten (50%) items are 'good' at discriminating between the high and low achieving test takers. Five (25%) items have D value of .22 - .33 indicating they are 'sufficient' in discriminating between the high and low achieving test takers. Four (20%) items are found to be 'bad' ones. They cannot distinguish between the two groups well. One of those four items has negative value (-0.11). The average index of discriminating power for the test with regard to the general English oriented test – the second test set – is .39. This D value indicates 'sufficient' discriminating power.

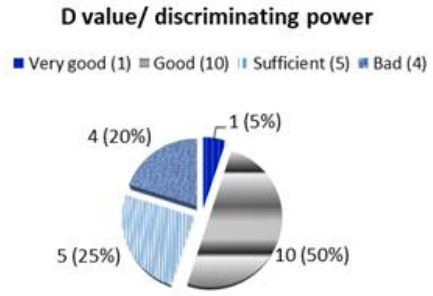

Figure 4. Discriminating Power of General English Items

When all 50 items are combined and analysed for their P value and D value, it is found – as seen in Figure 5 below – that 13 (26%) items belong to **easy** category (ranging from .75 to 1), 32 (64%) items belong to **average** category (ranging from .32 to .7), and 5 (10%) items belong to **difficult** category (ranging from .07 to .29).

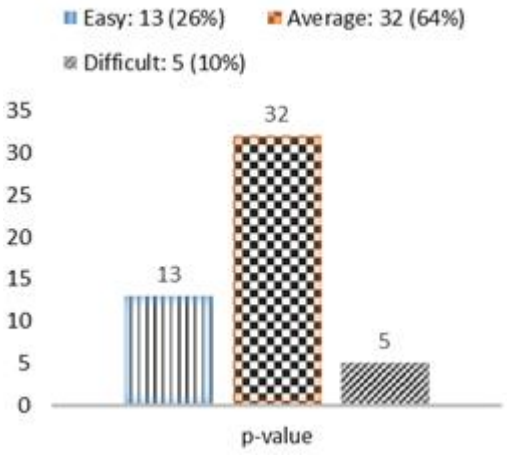

Figure 5. Item Difficulty of All Items

It is also found – as seen in Figure 6 – that 8 (16%) items belong to the category of 'very good' at discriminating test takers (D value

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://doi.org/10.15408/ijee.v6i1.11888
P-ISSN: 2356-1777, E-ISSN: 2443-0390 | This is an open access article under CC-BY-SA license

ranges from .83 to 1), 19 items belong to the category of 'good' at discriminating test takers (D value ranges from .44 to .66), 10 items belong to the category of 'sufficient' at discriminating test takers (D value ranges from .33 to .22), and 13 items belonged to the category of 'bad' at discriminating test takers (D value ranges from -.33 to 0). Two of these 13 items have negative values (-.33 and -.11).

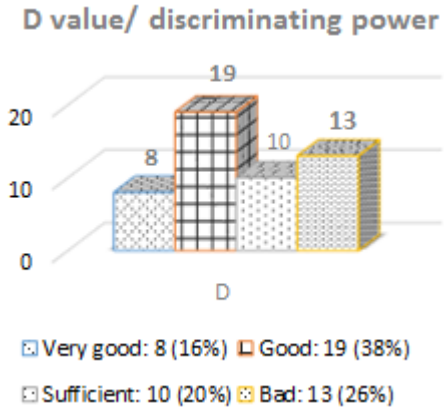

Figure 6. Discriminating Power of All Items

Having combined the detailed calculation of the two test sets – covering syllabus oriented and general English test, the writer finds that the average P value equals to .60 and the D value equals to .41. This finding makes it evident that the devised test has reached the category of **average** level of item difficulty and the classification of **good** at discriminating between the high and low achieving test takers. This particular finding of the study is congruent with Sim and Rasiah's (2006) stating that MCQ items that demonstrate **good** discrimination index tend to be **average** items for their item difficulty. They further claim that items that are in the **moderately difficult** to **very difficult** range are more likely to show negative discrimination.

Nevertheless, as it found that nine and four bad items appear in the first and second test sets respectively, the test devised for inclusion in the actual research should be reassessed. The bad items can simply be eliminated or improved by developing some more items. The items kept for inclusion in the actual research instrument should—following Boopathiraj and Chellamani (2013)'s suggestion – be arranged in such a way that items of higher indices of difficulty, of moderate indices of difficulty, and of lower indices of difficulty are organized in a balanced composition.

CONCLUSION AND SUGGESTIONS

This article is a report on test item analysis centering on Multiple Choice questions used to measure the proficiency of Indonesian High School teachers involved in English instruction. Restricted to the analyses of item difficulty and item discrimination, the study has found that with regard to the whole test (covering syllabus oriented and general English oriented

items) the average P value equals to .60 and the D value equals to .41. It is evident that the devised test has reached the category of **average** level of item difficulty and the classification of **good** at discriminating between the high and low achieving test takers. The complete test should, however, be improved for the actual research since some items—slightly above three quarters—are indicated as 'bad' at discriminating test takers.

The result of item analysis to the devised test in this study can hopefully become a section in a good item bank for the decision makers dealing with teacher professional development. Another suggestion might be for test developers to consider the need of the test takers by developing a test which attempts to see further the possibility of co-certification as exemplified by Newbold (2011).

Acknowledgements

The authors disclosed receipt of the financial support for the research, authorship, and/ or publication of this article from the Directorate of Research and Community Service, Indonesia Ministry of Research, Technology and Higher Education.

REFERENCES

Aniroh, K. (2009). From English as a general school subject onto English as a medium for learning specific subjects: The need to shift in the teaching orientation. *TEFLIN Journal. 20*(2), 169-179.

Anugerahwati, M. & Saukah, A. (2010). Professional Competence of English Teachers in Indonesia: A Profile of Exemplary Teachers. *Indonesian Journal of English Language Teaching. 6*(2), 47-59.

Barber, M. & Mourshed, M. (2007). *How the World's Best Performing Schools Come Out on Top*. London: McKinsey and Company.

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

Brown, D. H. (2004). *Language Assessment: Principles and Classroom Practices*. New York: Longman.

Brown, D. H., & Abeywickrama, P. (2010). *Language Assessment: Principles and Classroom Practices*. (2nd Edition). New York: Pearson Longman.

Brown, J. D. (1996). *Testing in Language Program.* New Jeersey: Prentice Hall Regents.

Caena, F. (2011). Teachers' core competences: Requirements and development. European Commission. http://ec.europa.eu/dgs/education_culture/repository/ education/policy/strategic-framework/doc/teacher-competences_en.pdf

Eln. (2018). Fokus pada Perbaikan Standar Mutu Guru (Focus on Improvement of Teacher Quality Standards). *Kompas,* 10 August, p. 12.

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://doi.org/10.15408/ijee.v6i1.11888
P-ISSN: 2356-1777, E-ISSN: 2443-0390 | This is an open access article under CC-BY-SA license

Gronlund, N. E. (1982). *Constructing Achievement Test* (3rd edition). New York: Prentice Hall.

Chetty, R., Friedman, J. N. & Rockoff, J. E. (2011). The Long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. Working Paper 17699 http://www.nber.org/papers/w17699 National Bureau of Economic Research.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education. 15*(3), 309–334.

Harjanto, I., Lie, A. Wihardini, D., Pryor, L., & Wilson, M. (2018). Community-based teacher professional development in remote areas in Indonesia. *Journal of Education for Teaching. 44*(2*), 212-231. https://doi.org/10.1080/02607476.2017.1415515*

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: Cambridge University press.

Kopriva, R. J. (2008). *Improving Testing for English Language Learners*. New York: Routledge.

Lee, K. S. (2004) Exploring the connection between the testing of reading and literacy: The case of the MUET. *GEMA Online® Journal of Language Studies. 4*(1).

Lengkanawati, N. (2005). EFL Teachers' competence in the context of English curriculum 2004: Implications for EFL teacher education. *TEFLIN Journal. 26*(1), 79-92.

Nair, R. & Arshad, R. (2018). The discursive construction of teachers and implications for continuing professional development. *Indonesian Journal of Applied Linguistics. 8*(1), 131-138. doi: 10.17509/ijal.v8i1.11472

Newbold, D. (2012). Local institution, global examination: Working together for a 'co-certification'. In Dina Tsagari / Ildikó Csépes (Eds.) *Collaboration in Language Testing and Assessment* (pp.127-142). Frankfurt: Peter Lang.

Othman, J. & Nordin, A. (2013) MUET as a predictor of academic achievement in ESL teacher education. *GEMA Online® Journal of Language Studies. 13*(1), 99-111.

Plakans, L. & Gebril, A. (2015) *Assessment Myths*. University of Michigan Press.

Prasetyo, S. (2017). Uji Kompetensi Guru, Tes Sesuaikan Kompetensi Guru. *Jawa Pos.* 2 July. Retrieved 9 August 2018 from https://www.jawapos.com/pendidikan/02/07/2017/uji-kompetensi-guru-tes-sesuaikan-kompetensi-guru

Putra, I. (2017). Pretest UKG (Ujian Kompetensi Guru) Ini Ujian Apa Ya? *Kompasiana.* Retrieved 9 August 2018 from https://www.kompasiana.com/indra-yahdi/59afb95aa32cdd1bae7721d3/pretest-ukg-ujian-kompetensi-guru-ini-ujian-apa-yaaa

Rasmussen, J. & Holm, C. (2012). In pursuit of good teacher education: How can research inform policy? *Reflecting Education. 8*(2), 62-71 http://reflectingeducation.net

Sharif, A. (2013) Limited proficiency English teachers' language use in science classrooms. *GEMA Online® Journal of Language Studies. 13*(2), 65-80.

Sim, D. S. & Rasiah, R. I. (2006). Relationship between item difficulty and discrimination indices in true/false type multiple choice questions of a para-clinical multidisciplinary paper. *Annals Academy of Medicine. 35*(2), 67-71.

Soepriyatna. (2012). Investigating and assessing competence of high school teachers of English in Indonesia. *Malaysian Journal of ELT Research. 8*(2), 38-49.

Tsang, W. L. (2011) English metalanguage awareness among primary school teachers in Hong Kong. *GEMA Online® Journal of Language Studies. 11*(1), 1-16..

http://journal.uinjkt.ac.id/index.php/ijee | DOI: http://doi.org/10.15408/ijee.v6i1.11888
P-ISSN: 2356-1777, E-ISSN: 2443-0390 | This is an open access article under CC-BY-SA license