AJOTE
African Journal of Teacher Education
ISSN 1916-7822. A Journal of Spread Corporation

Volume 10. No. 2 2021 Pages 121-143

Analysis of Item Writing Flaws in a Communications Skills Test in a Ghanaian University

Ato Kwamina Arhin,[1] Jonathan Essuman,[2] & Ekua Arhin[3]
[1,2] Faculty of Education and Communication Sciences, AAMUSTED, Kumasi, Ghana
[3]Department of Education, Ola College of Education, Cape Coast, Ghana

Abstract

Adhering to the rules governing the writing of multiple-choice test items will ensure quality and validity. However, realizing this ideal could be challenging for non-native English language teachers and students. This is especially so for non-native English language teachers because developing test items in a language that neither they nor their students use as their mother tongue raises a multitude of issues related to quality and validity. A descriptive study on this problem was conducted at a Technical University in Ghana which focused on item writing flaws in a communication skills test. The use of multiple-choice test in Ghanaian universities has increased over the last decade due to increasing student intake. A 20-item multiple-choice test in communication skills was administered to 110 students. The test items were analyzed using a framework informed by standard item writing principles based on the revised taxonomy of multiple-choice item-writing guides by Haladyna, Downing and Rodriguez (2002). The facility and discrimination index (DI) was calculated for all the items. In total, 60% of the items were flawed based on standard items writing principles. The most violated guideline was wording stems negatively. Pearson correlation analysis indicated a weak relationship between the difficulty and discrimination indices. Using the discrimination indices of the flawed items showed that 84.6 % of them had discrimination indices below the optimal level of 0.40 and above. The lowest DI was recorded by an item with which was worded negatively. The mean facility of the test was 45%. It was observed that the flawed items were more difficult than the non-flawed items. The study suggested that test items must be properly reviewed before they are used to assess students' knowledge.

Keywords: Discrimination index, Facility, Flawed item, multiple-choice item

Background

Multiple-choice items (MCIs) are one of the most commonly used item types for classroom assessment (Haladyna & Rodriguez, 2013). Because of its widespread use in the classroom multiple-choice items are highly indispensable regarding testing students at all levels of education. There is hardly any subject that cannot use MCI. Test results are often used to make decisions that determine the future of students and teachers. It is, therefore, imperative that MCIs are properly handled at the construction, administration, scoring and in analyzing the test scores.

Moreover, increasing enrolment in Ghanaian tertiary institutions, multiple-choice items (MCIs) have become the preferred mode of assessing students because of the greater ease and speed of grading of multiple-choice questions compared with other testing formats. They also cover a wide scheme of work or syllabus adequately. When assessing a large population of students, it will be very difficult to ignore multiple-choice items (MCI). In 2015, the gross enrolment ratio in tertiary education for Ghana was 16.2 %. The gross enrolment ratio in tertiary education in Ghana increased from 0.7 % in 1972 to 16.2 % in 2015 growing at an average annual rate of 28.47%. This indicates a substantial growth in students' intake at the tertiary level in Ghana. Due to increasing student intake many faculty members have resorted to the use of multiple-choice items to meet students' assessment needs. Previously, MCIs were rarely used in our tertiary schools. Restricted response type of test items dominated mid and end of semester examinations because students' enrolment was not as high as today. Essay was the ideal means of assessing students' knowledge. The challenge now is how to construct good quality MCIs that have minimal flaws to elicit the knowledge possessed by the students. Essay test items are relatively easy to construct compared to MCIs. Multiple-choice item is made up of the stem- which possess the problem to be resolved, a set of options that consist of the key and distractors, respectively, the answer and the options that suggest the wrong alternative to the test taker.

McKeachie (1999) notes that multiple-choice test is a staple of higher education because it provides an efficient and effective measure of student learning. The acceptance of multiple-choice test has increased over the years, partly due to improvements in technology in scoring multiple-choice items quickly and easily. The multiple-choice test is also highly reliable across scorers, unlike essay tests. For these reasons and others (Frederiksen, 1984), many educators consider the multiple-choice format as an optimal method of testing. However, multiple-choice tests have spawned substantial controversy mainly because questions of this

type are limited to measuring recall of knowledge. Despite all the weaknesses associated with it, multiple-choice (MC) tests are preferred in educational settings in Ghana.

The teacher's aim in crafting MCIs is not to confuse students, but to yield scores that accurately reflect the extent to which students have obtained an acceptable working knowledge of the content. It is worthy to note that students who pass a poorly designed test, may not necessarily possess adequate knowledge of the topic and this may constitute a real threat to their future academic progression. Well-constructed multiple-choice items represent a versatile assessment tool with the potential to assess students for sufficient evidence of the knowledge of the tested content (Rush, Rankin and White, 2016). A required characteristic of a multiple-choice item is its power to be able to discriminate between the test takers who have learnt the material they are being tested on and those who have not learnt it. The discrimination index can differentiate between students of different ability levels. Poorly constructed MCIs also contain cues that allow students to guess the correct answer without prerequisite knowledge (Downing, 2002).

It is time-consuming and energy-sapping to construct an item that is good enough to discriminate among testees. According to Rush, Rankin and White (2016), it takes about 20 to 60 minutes to couch a quality multiple-choice item free from errors. Despite the importance of classroom assessment, studies suggest some deficiencies in teacher-made tests, (Mehrens and Lehmann, 2009). According to Lane et al. (2016), most teachers craft flawed items that measure the ability to recall basic facts and concepts. Some effects of item-writing flaws on students are; items may be easier or more difficult than intended, clues that will allow unprepared students to guess the correct answer and unnecessarily complex or esoteric test items prevent prepared students from demonstrating their knowledge (Case and Swanson, 2002; Downing, 2005). A poorly constructed item can inflate or deflate the student's score on a test and this represents a false picture of the student's performance. Also, these flaws are capable of clouding the results obtained from the test. The effect of the clouding of results is that it changes the interpretation of the results and it contributes to unwanted evidence getting into the test data.

There are many factors to consider when evaluating the quality of MC items. Firstly, one can examine the extent to which items conform to widely accepted item-writing guidelines, such as avoiding negatively worded items and avoiding the use of longer options as the answers. Writing MCIs without following the guidelines can result in lowering the quality of individual items and the test as a whole (Downing, 2005; Tarrant & Ware, 2008).

Specific research-based principles guide the development of effective MCIs (Downing & Haladyna, 1997; Haladyna, 2004). The use of these research-based principles makes item writing a science. In a review by *Haladyna, Downing and Rodriguez (2002*) a taxonomy of 31 item-writing principles, based on an analysis of 27 current educational measurement textbooks and 27 empirical research papers, have been identified. Deviation from established item-writing principles may result in a decrease in validity evidence for tests (Downing, 2002). Items that violate one or more of the standard items writing principles—flawed items—tend to produce construct irrelevant easiness which refers to a contaminating influence on test scores that tend to systematically increase test scores for a specific examinee or a group of examinees; construct-irrelevant difficulty does the opposite. It systematically decreases test scores for a specific examinee or a group of examinees (Haladyna, Downing, 2004). These effects are called construct-irrelevant variance (CIV).

Similarly, Multiple-choice test items tend to have high grading reliability, however, creating valid MC items that perform reliably is difficult and requires skill to do that properly (Pellegrino, Chudowsky & Glaser, 2001). Many teachers usually have little to no formal training regarding appropriate assessment practices. For example, most pre-service teachers in Ghana take a three-hour course in assessment in schools which is woefully inadequate to prepare them for the enormous task ahead of them. In addition to a lack of training, another reason is creating MC items can be difficult because there are numerous ways to lessen an examination's validity based on how it is designed.

Although multiple-choice items are commonly used in tertiary institutions and other levels of language instruction and other subject areas in Ghana, there has not been enough evidence about the item analysis of multiple-choice tests in the area of communication skills. It is important to note that “the quality of a test largely depends on the quality of the individual items” (Oluseyi & Olufemi, p.240). Therefore, this study attempts to fill this gap by answering the following research questions: (1) What is the difficulty level (item facility) of each item on the communication skills test? (2) What is the discrimination index (item discrimination) of each item on the communication skills test? (3) What is the relationship between the facility and the discrimination index of the item on the communication skills test?

The present study was undertaken in the first semester of the 2019/20 academic year to assess some item writing flaws observed in a Communication Skills test in a Technical University in Ghana. The observed flaws were, "longer sentences as answers among the options", "use of negative words", "starting a statement with a blank" and "options not arranged

in alphabetical order". These flaws are capable of introducing testwiseness in the answering of the test by students. Testwiseness is defined as a student's capacity to utilize the characteristics and formats of a test and/or the test-taking situation to receive a high score (Millman, Bishop and Ebel, 1965). Flawed test items are capable of providing test-wise cues to the items, thereby distorting the true performance of the student. Given the widespread use of MCIs in tertiary educational settings, it is practically important to look carefully at the quality of the MC items on classroom tests, and this was the specific purpose of this study. Evaluating the quality of MC items involves many factors. Firstly, one can examine the extent to which items conform to widely accepted item-writing guidelines, such as avoiding negatively worded items and avoiding the use of longer options as the answers. Writing MCIs without following the guidelines can result in lowering the quality of individual items and the test as a whole (Downing, 2005; Tarrant & Ware, 2008). Secondly, analyzing the responses from testees is another approach used in the research presented here. Specifically, we analyzed a teacher-made Communication Skills test administered to first-year IT students at a Technical University in Ghana and focused on how students score on the flawed items were affected by two major characteristics of MC items: facility and discrimination index.

Methodology

Participants

All 110 respondents were first-year undergraduate students pursuing a degree in Information Technology Education. These students were purposively selected because the instructor agreed to allow us to use his test items for the study. The test was conducted at a Technical University in Ghana. To ensure fairness, students were informed ahead of time to prepare for the test.

Instrument

The test consisted of 20 multiple-choice test items. The test items were used to assess students' communication, paragraphing and writing skills. The topics covered in the test constituted what had been taught in that semester. The items had four options, one of them being the correct answer and the other three being distractors. One of the items had only two options because it was a true/false item. In scoring the test no penalty was employed for guessing and the correct answer was awarded a mark of 1. Thus, the maximum possible score of the test was 20 and a minimum of 0. A copy of the test is included as an appendix.

Time Period and Procedure

Data was collected during the first semester of the 2019/20 academic year. The test was conducted under examination conditions. The test was administered by the English language instructors of the university and the students were supposed to answer the questions in 25 minutes.

Data Analysis

Students' responses from the MCIs were analyzed using Microsoft Excel. The MCIs were analyzed to obtain the facility (p-value), the discrimination index (DI), and distractor analysis for all non-correct options. The Kuder–Richardson formula (KR-20) was used to assess the internal reliability of the test scores. Data was analyzed based on the three research questions. Research questions #1 and #2 were answered using Microsoft Excel template for the facility and discrimination indices. Research question #3 was answered using Pearson product-moment correlation.

Item Evaluation Procedure

There are several methods available for evaluating multiple-choice items. Specifically, for this study, the aim was to determine which items exhibited the best quality in terms of option performance. The evaluation consisted mainly in inspecting the facility and discrimination indices for each test item. The result of the examinees' performance in the test was used to analyze the facility and the discrimination indices (DI) of each multiple-choice item. The facility is calculated as a percentage of the total number of correct responses to the test items. It is calculated using the formula $p = \frac{R}{T}$, where p is the facility, R is the number of correct responses, and T is the total number of responses (which includes both correct and incorrect responses).

According to Hotiu (2006), the p (proportion) value ranges from 0 to 1. When multiplied by 100, the p-value converts to a percentage, which is the percentage of students who got the item correct. The higher the p-value, the easier the items. This means the higher the facility, the easier the item is understood to be. It needs to be conceptualized that a p-value is a behavioural measure. Instead of explaining the facility in terms of some intrinsic characteristic of the item, the facility is defined in terms of the relative frequency with which those taking the test choose the correct response (Thorndike, Cunningham, Thorndike, & Hagen, 1991).

The item DI is the point biserial correlation between getting the item right and the total score on all other items. The discrimination index is the point biserial correlation between item score and corrected total score. This was computed using a Microsoft Excel sheet. The advantage derived from using this procedure is that it provides a more accurate assessment of the discrimination power of items because they take into account the responses of all students rather than just high and low scoring groups.

Discrimination index reflects the degree to which an item and the test as a whole are measuring a unitary ability, values of the coefficient will tend to be lower for tests measuring a wide range of content areas than for more homogeneous tests. Item discrimination indices must always be interpreted in the context of the type of test which is being analyzed. The higher the DI the better the test item discriminates between the students with higher test scores and those with lower test scores. According to Haladyna and Rodriguez (2013) guidelines for evaluating MC items based on classical test theory is provided in Table 1.

Table 1: Guidelines for evaluating test items (adapted from Haladyna & Rodriguez, 2013, p. 350)

Type	Difficulty	Discrimination	Comment
1	.60 to .90	Above .15	Ideal item; moderate difficulty and high
2	.60 to .90	Below .15	Poor discrimination
3	Above .90	Disregard	High-performance item; usually not very
4	Below .60	Above .15	Difficult but very discriminating
5	Below .60	Below .15	Difficult and non-discriminating
6	Below .60	Below .15	Identical to type 5 except that one of the distractors has a pattern

Results

In this study, the flawed items were more difficult than non-flawed items measuring the same content. The mean test score was 9. The lowest score was 3 and the highest was 15. A quick synopsis of the test results showed that 12 of 20 items were flawed when assessed in the light of the standard item forms. This represents 60% of the items. This observation was based on the revised taxonomy of multiple-choice item-writing guides by Haladyna, Downing and Rodriguez (2002). It is known that teacher-made tests are filled with a lot of flaws. Four kinds of flaws were observed namely, longer option as the answer, negatively worded item, options not arranged in alphabetical order and starting with a blank. The frequently violated rule was 'negatively worded item', there were 8 out of 12 flawed test items. The reliability estimate for the test measured by KR-20 was 0.41. According to Rudner and Schafer (2002), a teacher-made assessment needs to demonstrate reliability coefficients of approximately 0.50 or 0.60.

The facilities for the test items ranged from .17 to .86. The mean facility for the test was 45% that is p = 0.45. The optimal facility for a classroom teacher-made test is .63. Comparing the optimal value with a facility for the test indicates that the test was difficult. A possible reason for the difficult nature of the test could be because 60% of the items were flawed. Items 1and 2 were very easy compared to the optimal facility for classroom achievement tests.

Odukoya et. al (2018), observed that majority of the items used in a private university in Nigeria (about 60 out of the 70 items fielded) did not meet psychometric standard (of appropriate difficulty and distractive index) and consequently need moderation or deletion. Approximately, 86% of the items failed to meet the suitable psychometric properties. The current shows that 12 out of 20 items were flawed when assessed in the light of the standard item forms. This represents 60% of the items. This therefore, collaborates the study by Odukoya et al. (2018) which suggests that the teachers need to improve their item writing skills. One danger associated with flawed items is that they introduce errors into the student's test score thereby making the difference between the observed score and the true score wider. It is these same results filled with errors that will be used to provide certificates for the students. This, therefore, calls on all involved in crafting of test items especially, MCIs to be abreast of the currents suggestions for writing multiple-choice test items.

Table 2: Item Properties

Item	Type of flaw	Facility	Discrimination index
Q 1	Longer option as the answer	.86	.09
Q 2	Negatively worded	.70	.17
Q 4	Negatively worded	.66	.10
Q 4	Options not arranged in alphabetical order	.66	.10
Q 5	Starting with a blank	.58	.40
Q 6	Negatively worded	.46	.11
Q 7	Negatively worded	.33	.11
Q 9	Negatively worded	.45	.15
Q 10	Starting with a blank	.49	.17
Q 15	Negatively worded	.46	-.02
Q 16	Negatively worded	.39	.08
Q 18	Negatively worded	.33	.22

Discussions

Longer option as the answer

Responses should be similar in length, the shorter the better (if one option is much longer than the others, students will assume that is either the correct answer or blatantly the wrong answer, which gives them better odds at "guessing". Responses differ in length because the teacher would like to add qualifying phrases to make sure the keyed option is correct. Many novices and experienced test constructors make this mistake to respond free from disputations (Haladyna & Downing,1989). Test-wise students look for the "longest responses" to choose as answers during tests when they are unsure of the correct option. Making the responses almost the same length reduces the bias of such items and improves the validity of the measurement. The item analysis showed that the item was very easy. A total of 86 % of the students answered the item correctly. The facility of the item was .86. The distracters for these items were not good enough to discriminate among the students.

Figure 1: An example of an item with the longer option as the answer from the test is:

Q.1

Communication is a universal activity because it……

A. is a credible source of data collection
B. create the right atmosphere of dialogue
C. enables people to give out or receive information
D. is therapeutic

Response key is C

According to Haladyna and Downing (1989), 8 of 9 studies suggest that using long correct options makes items easier; Q.1 shows no difference. This study shows that the item with a longer option as the answer was the easiest of the 20 items used for the test. In this study, this particular item was the first and it can be argued that the first few items on a test be easier to avoid students losing confidence. This agrees with the notion regarding the arrangement of test items, that easier items start, put the difficult items in the middle and conclude with the easy items. In as much as a test developer will fulfil this condition, they should not compromise on item writing rules to justify the wrong.

It recorded a p-value of .86 and this validates the rule that the long options are often selected as the answer. This finding, therefore, collaborates with Haladyna and Downing (1989). Also, this item was less discriminatory between knowledgeable and non-knowledgeable students (DI=0.09). Board and Whitney (1972) as cited in Haladyna and Downing (1989) posit that low-achieving students were inclined to take advantage of the option-length clue, whereas higher achievers do not. Higher achievers are disadvantaged when such items are prevalent in a test. It is, therefore, important that tests are rid of these items.

Starting a question with a blank

The stem can be written in two forms - as a question or partial sentence that requires completion. Research comparing these two formats have not demonstrated any significant difference in test performance (Violate, 1991, Haladyna, 1999, Masters et al 2001). To facilitate understanding of the question to be answered, it is recommended that if a partial sentence is to be used, a stem with parts missing either at the beginning or in the middle of the stem should be avoided (Haladyna, 1999). It is recommended that the blank should be towards the end of the stem or sentence. It is natural that when conversing with someone you do not start with a blank for the other person to fill in. This does not make the communication effective. Therefore, starting a stem or sentence with a blank distorts the meaning of the question and makes it difficult to answer. From the test, questions 5 and 10 started with a blank and recorded facility of .58 and .49 respectively. From Table 1 it is obvious that their facilities are not within the range of ideal items.

An example of an item that starts with a blank is:

Q.5 ……...are to move the reader to make a particular choice or to take a particular course of action.

A. Expository paragraphs
B. Mainstream paragraphs
C. Narrative paragraphs
D. Persuasive paragraphs

The key is D

The distractor analysis of item 5 indicates no selection for option C. In the entire test, item 5 was the only item that recorded a zero selection for an option. This particular option C was

rendered implausible and unattractive even to the lower achievers because the item began with a blank. Three options for this particular item would have worked well instead of padding the question options that did not work. Because putting options A and B together results in less than 40%, clearly the options were not plausible at all.

On the contrary, item 5 which started with a blank recorded a discrimination index of 0.40 among all the 20 items which are considered as the optimal level of DI for multiple-choice test items. Even though item 5 is flawed based on the guidelines for writing MCIs, its item indices were all excellent (p=.58, DI=.40). This finding does not conform to outcomes from other studies indicating that beginning the stem with a blank cannot discriminate well between the high achievers and lower achievers. Also, this finding does not conform to outcomes from other studies indicating that beginning the stem with a blank makes the item difficult.

However, with all the excellent item characteristics for item 5 it clearly violates the 'cover options test'. An important indicator of a well-written MCI, according to many writing guidelines, is that the question should allow testees to formulate a correct response without needing to first look at the available options – a criterion commonly referred to as the 'cover options test' (Case and Swanson, 2002). This guideline is violated by beginning a stem with a blank hence the respondent needs to fit in the options one after the other before the correct option can be selected.

Negatively worded items

In this study, 7 of 20 items were constructed in the negative sense. Cautiously, these items can be deemed as difficult based on the established range of p-values considered excellent, that is 40% and 60%. Item 15 which was constructed negatively recorded a discrimination index of -.02. A negative value indicates an inverse relationship between the item and test performance. A total of 55 students got the item wrong. This makes the item difficult. This study is not consistent with Harasym, Price, Brant, Violato and Lorscheider (1992), who posited that negatively couched items stems were less difficult. The seven negatively worded items in this study recorded low discriminatory indices. These items were confusing for the higher achievers but rather favoured the lower achievers. A survey of authors' guidelines on multiple-choice items showed that 63% are in favour of wording items positively, 19% were in support of using negatively worded items and 19% did not discuss this (Haladyna, Downing and Rodriguez, 2002). Dudycha and Carpenter (1973) speculated that a negative orientation in the stem makes the item more difficult because this requires a negative-to-positive shift

in mental orientation to answer questions. The purpose of MCIs is to measure the achievement of learning objectives thus to achieve instructional validity, a teacher must test what has been taught and since most learning objectives are not stated negatively, it means that writing items in the negative will not help the teacher to test students' knowledge on what has been taught. Negatively worded stems should only be used when a student knows what to avoid or what is not the case. Most often, students learn what to do and what is the case, and thus item stems should be positively worded (Harasym, Price, Brant, Violato & Lorscheider, 1992).

A student who reads very fast may miss the 'not' keyword and, consequently, the entire meaning of the question. However, negatively couched items are useful when necessary to measure an appropriate objective (e.g., what is not correct) and with negative words (*at least, except* and *not*) are underlined, highlighted, **boldface**, *italicized* and CAPITALIZED to caution the individual taking the test. It is also important to word each option positively to avoid forming double negative with the stem. There are some legitimate uses of negative terms, such as the case of medications or procedures that are contraindicated; this use may be legitimate in that "contraindication" is a straightforward concept in health care domains.

Q. 15 Which of the following is NOT a feature of a good note?

A. Descriptive
B. Readable
C. Reflects the source
D. Understandable

The key is A

Options arranged in order (Alphabetical or sequential)

One of the most ignored multiple-choice item writing guidelines is to arrange options in alphabetical, chronological, or conceptual order. The purpose of this guideline is for easy reading and to make options appear attractive to the test taker. Options can be arranged in either ascending or descending order. It becomes difficult to observe this guideline especially, for novice teachers when the options are in sentences. Also, it is not easy to obey this, because

observing this guideline could create a discernible pattern for the correct options on a test. In analysing these two items, it was valuable to look at the discrimination index (DI) and distractor analysis was important. The DI indicates the relationship between performance on an individual item and performance on the overall test. The discrimination index (DI) was .01 for item Q4. The DI indicates that Q4 falls below the range where the items are to be rejected or improved. Arranging options in an order is not considered important by most class teachers when constructing items, but the indices obtained from this study shows that it must be considered seriously. The p-value for Q4 was .66. An example of the item with options not arranged in alphabetical order is:

Q4. The concluding paragraph has the following functions EXCEPT to……..

A. introduce new research idea

B. refer to cause or effect of issues

C. summarize main ideas

D. suggest solutions to issues

The key is A

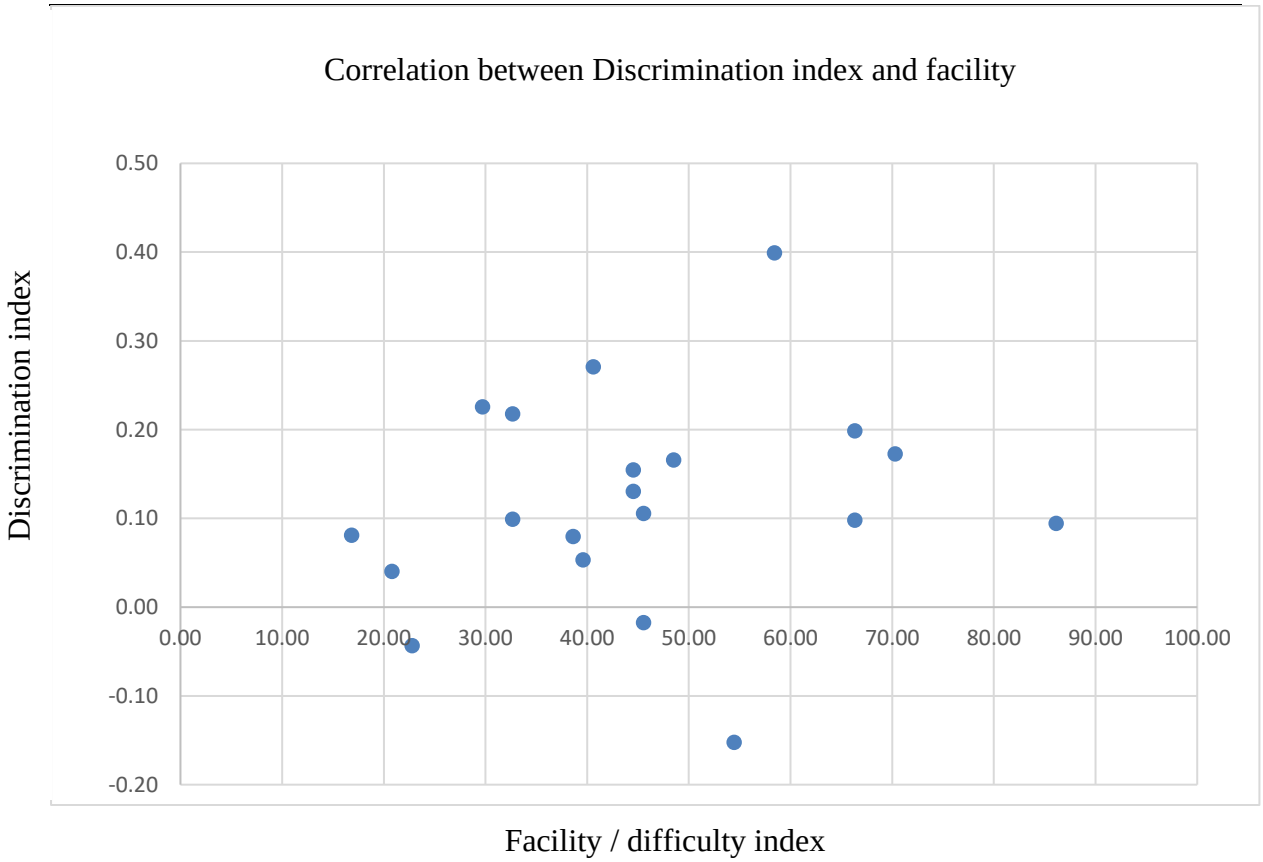
The first two options were correctly ordered, but option (d) should have come before option (c).

Correlation between the difficulty index and discrimination index

A Pearson product-moment correlation coefficient was computed to assess the relationship between the facility and the discrimination index. There was a positive correlation between the two variables, r = 0.162, n = 20, p = 0.496. The maximal discrimination (D =0.4). A scatter plot (Figure 1) represents the relationship between the difficulty index (P) and discrimination index (D) of 20 MC items. The plot is not linear, rather very scattered in shape which indicates a weak relationship between the difficulty index and discrimination index. Increasing difficulty indices do not correlate with increases in the discrimination index. It is seen from the scatter plot that only three items (15 %) recorded negative discrimination.

Table 1: Correlation between facility and discrimination index

		Discrimination index	Facility
Discrimination index	Pearson correlation	1	.162
	Sig (2-tailed)		.496
	N	20	20
Facility	Pearson correlation	.162	1
	Sig (2-tailed)	.496	
	N	20	20



Recommendations for Improvement

There are many innovative and easy ways to implement strategies that can help teachers improve their knowledge and skills in constructing MCIs. For example, team item writing by way of leveraging the expertise of colleagues and senior faculty members can help construct well-composed test items. New faculty must be oriented toward constructing MCIs and be

assigned to experienced hands. Another strategy is 'nudging' and 'shoving' where distractors are easily manipulated to alter an item's facility. According to Quaigrain and Arhin (2017), the appropriate quality of MCI is based on the presence of quality distractors. This assertion collaborates with the suggestion of moving from the traditional four or five option responses to three options because this will assuage the challenges of producing more than two plausible distractors without affecting students' performance. On the other hand, reducing the number of options tends to increase students' chances of guessing which is better than padding the options with non-functional distractors.

Teachers must understand that crafting multiple-choice items is both science and art. Both science and art are to be employed fully to get efficient items that will produce valid results. Test items must be solely written based on learning objectives so that the teacher will know what exactly each item measured. This in a way will reduce the flaws in the items and will ensure instructional validity.

Jozefowicz, Koeppen, Case, Galbraith, Swanson and Glew (2002) posit that teachers spend substantial time planning their lectures and course materials for students and insufficient time is allowed for test preparation and review before administration. Consequently, many tests are administered to students without adequate pretest to check the quality of the items. Before a test is administered, a review by an examinations review board whose members have adequate knowledge in item writing can eliminate flawed items.

Conclusion

The purpose of the study was to analyze the responses obtained from a communications skills test in terms of facility and discrimination indices. Overall analysis of the items, show that most of the items were ideal- the items had acceptable facility and discrimination indices based on the guidelines for evaluating test items (Haladyna & Rodriguez, 2013, p. 350). On the other hand, some items were not ideal. These items were found to need revision to improve the discriminatory power and the quality of the examination.

It is worth noting that firmly following the guidelines for writing multiple-choice test items can reduce the number of flawed items on a test. Some teachers who write multiple-choice items are either ignorant or find it too laborious to use the guidelines. As a result, you find items on a test being flawed. The guidelines serve as a compass to the item writer to his destination without missing out. The ultimate aim of all item writers is to have good items that will produce valid test scores. After all, students who pass a poorly designed exam, although

they do not possess adequate knowledge of the content of the examination, may constitute a real threat for themselves and society at large.

Studies show that tests with item writing flaws tend to disadvantage high achieving students and lower their test scores (Tarrant & Ware, 2008). Contrastingly, tests with item writing flaws can improve the grades of weaker students, who are not familiar with the content of the test (Nedeau-Cayo et al., 2013; Tarrant & Ware, 2008). We, therefore, are of the view that, in contrast to these undisputed views of authorities, using flawed MCIs can play a valuable role in the development (and not simply in the measurement of academic performance) of students' critiquing abilities. Thus, students become active observers of the learned materials and objectives of the lessons taught. This will help in the development of multiple-choice items and making them stand the test of time and become robust. But this will come at a cost to both teachers and students.

Finally, the findings of this study have significance for practising teachers and test developers in that particular care should be taken when selecting or crafting new items to achieve an accurate measurement of students' behaviour. Also, Item analyses should be utilized to improve already existing test items.

Assessment Implications

On the strength of the findings made from this study using MCIs should begin with preparation of test blueprint that carefully adheres to the rules for writing multiple-choice items. Thereafter all items should be pre-tested, analysed, and subjected to item moderation to augment the overall content and construct validities. These processes will require the input of subject and psychometric specialists. To ensure that faculty uses quality test items in our tertiary institutions, it implies that these processes be established as statutory quality assurance procedures. In sum, anytime a teacher is deciding on a test, the following must be carefully considered to ensure maximum gains from the test:

(a) the test's specific purpose:

The traditional assumption is that tests are used to determine whether students have learned what they were expected to learn or the level or degree to which students have learned the material. Beyond this, the tests' purpose might be that the teacher intentionally plans to improve students' performance or plans to help students estimate what they are capable of doing outside the classroom. Hence, pre-testing students to determine their previous knowledge before introducing a new topic is an important teaching strategy. Teachers are

aware that continuously assessing students enables them to adjust their instruction appropriately.

(b) what kind of information is required from the test results:

Test results provide vital information to both the testee and teacher. The teacher must consider what type of information the test scores is to provide to the student.

(c) the impact of test results on students:

Test results must have a positive impact on students. The test results should provide feedback that will motivate students to learn and improve their learning.

Author Contribution

- Ato Kwamina Arhin: introduction, literature review, data analysis, discussion, and conclusion.
- Jonathan Essuman: literature review, methodology and conclusion.
- Ekua Arhin: data analysis, discussion and conclusion.

References

Alderson, J. C. (2000). *Assessing Reading*. Cambridge: Cambridge University Press.

Case, S. M., & Swanson, D. B. (2002). *Constructing written test questions for the basic and clinical sciences*. (3rd ed.). Philadelphia: National Board of Medical Examiners; p. 31–66. http://www.nbme.org/publications/item-writing-manual.html.

Dudycha, A. L., & Carpenter, J. B. (1973). Effects of item format on item discrimination and difficulty. *Journal of Applied Psychology, 58*, 116-121.

Downing, S. M. (2002). Construct-irrelevant variance and flawed test questions: Do multiple-choice item writing principles make any difference? *Academic Medicine 77*, 103–104.

Downing, S.M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education: Theory and Practice, 10(2),* 133–143.

Downing, S. M. (2006). Twelve steps for effective test development. In. S. M. Downing & T. M Haladyna (Eds.), *Handbook of test development* (pp. 3–25). Mahwah, NJ: Lawrence Erlbaum Associates.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10,* 61-82. http://dx.doi.org/10.1207/s15324818ame1001_4

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

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. American Psychologist, 39, 193-202

Haladyna, T. M. (1999). *When should we use a multiple-choice format*? Paper presented at the annual meeting of the American Educational Research Association, Montreal Canada.

Haladyna, T.M & Downing, S.M. (1989). Validity of a Taxonomy of Multiple-Choice Item-Writing Rules. *Applied Measurement in Education*, 2:1, 51-78, http://dx.doi.org/10.1207/s15324818ame2014_4

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

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, 309-344. http://dx.doi.org/10.1207/S15324818AME1503_5

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23(1),* 17–27.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and Validating Test Items*. New York, NY: Routledge.

Haladyna, T. M. & Downing, S. M. (2004). Construct-irrelevant Variance in High-Stakes Testing. *Educational Measurement: Issues and Practice*

Harasym, P. H., Doran, M. L., Brant, R., & Lorscheider, F. L. (1992). Negation in stems of single-response multiple-choice items. *Evaluation and the Health Professions*, *16(3),* 342–357. http://doi:10.1177/016327879201500205

Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002).

The quality of in-house medical examinations. *Academic Medicine, 77,* 156-161. http://dx.doi.org/10.1097/00001888-200202000-00016

Lane, S., Raymond, M. R., Haladyna, T. M., & Downing, S. M., Test development process. In S. Lane., M. R. Raymond., & T. M. Haladyna (Eds), Handbook of test development (pp. 3-18). New York: Routledge, 2016.

McKeachie, W. J. (1999). Teaching tips: Strategies, research, and theory for college and university teachers (10th ed.). Boston: Houghton Mifflin.

Masters, J. C., Hulsmeyer, B. S., Pike, M. E., Leichty, K., Miller, M. T., & Verst, A. L. (2001). Assessment of multiple-choice questions in selected test banks accompanying textbooks used in nursing education. *Journal of Nursing Education*, *40,* 25-32.

Mehrens, W. A. & Lehmann, I. J., Measurement and evaluation in education and psychology. New York: Harcourt Brace College Publishers, 1991.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of testwiseness. *Educational and Psychological Measurement, 25, 707-726.*

Nedeau-Cayo, R., Laughlin, D., Rus, L., & Hall, J. (2013). Assessment of item-writing flaws in multiple-choice questions. *Journal for Nurses in Professional Development, 29, 52–57.*

Odukoya, J.A., Adekeye, O., Igbinoba, A.O.& Afolabi, A. (2018). Item analysis of university-wide multiple choice objective examinations: the experience of a Nigerian private university. *Quality & Quantity 52, 983–997* https://doi.org/10.1007/s11135-017-0499-2

Oluseyi, A. E., & Olufemi, A. T. (2012). The Analysis of Multiple-Choice Item of the Test of an Introductory Course in Chemistry in a Nigerian University. International Journal of Learning, 18(4), 237-246. doi:10.18848/1447-9494/CGP/v18i04/47579.

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). Knowing what students know: The science and design of educational assessment. Washington, D. C.: National Academy Press.

Quaigrain, K & Arhin, A.K. (2017). Using reliability and item analysis to evaluate a teacher-

developed test in educational measurement and evaluation, *Cogent Education, 4:1,* 1301013 https://doi.org/10.1080/2331186X.2017.1301013

Rudner, L. M., & Schafer, W. D. (2002). *What teachers need to know about assessment.* Washington, DC: National Education Association. Retrieved from http://echo.edres. org: 8080/nea/teachers.pdf

Rush, B. R., Rankin, D. C. and White, B. J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education 16:250* DOI 10.1186/s12909-016-0773-3

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education, 42,* 198-206. http://dx.doi.org/10.1111/j.1365-2923.2007.02957x

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). *Measurement and evaluation in psychology and education* (5th ed.). New York, NY: MacMillan.

Violato, C. (1991). Item difficulty and discrimination as a function of stem completeness. *Psychological Reports 69(3 P11):739-743*

(March, 2021). Ghana-Gross enrollment ratio in tertiary education https://knoema.com/atlas/Ghana/topics/Education/Tertiary-Education/Gross-enrolment-ratio-in-tertiary-education

Appendix

DEPARTMENT OF LANGUAGES EDUCATION

MID-SEMESTER EXAMINATION

PAPER TITLE: COMMUNICATION SKILLS DURATION: 25 MINS

PAPER CODE: GPD 111

INDEX NUMBER: …… CLASS: ……...

Answer all the questions on the question paper.

1. Communication is a universal activity because _____ .
   a. it is a credible source of data collection
   b. it creates the right atmosphere of dialogue
   c. it enables people to give out or receive information
   d. it is therapeutic
2. Which one of the following does NOT constitute one of the reasons why we communicate?
   a. To establish relations
   b. To persuade
   c. To share information
   d. To test the efficacy of words
3. Non-verbal communication largely involves the use of _____ .
   a. cues
   b. posters
   c. symbols
   d. vision
4. The concluding paragraph has the following functions EXCEPT _____ .
   a. to introduce new research idea
   b. to refer to cause or effect of issues
   c. to summarize main ideas
   d. to suggest solutions
5. _____ are to move the reader to make a particular choice or to take a particular course of action.
   a. Expository paragraphs
   b. Mainstream paragraphs
   c. Narrative paragraphs
   d. Persuasive paragraphs
6. Paragraphs can be distinguished according to the following EXCEPT _____ .
   a. Function
   b. Length
   c. Position
   d. Unity
7. Which one of the following is not a major drawback in effective communication?

a. Distortion
b. Faking attention
c. Noise
d. Semantic distraction

8. New perspectives are discovered when the author _____ the work.
   a. edits
   b. proofreads
   c. simmers
   d. writes
9. A group of students were given a topic to write for an assignment. The students have to go through the following stages EXCEPT _____ .
   a. drafting
   b. prewriting
   c. revision
   d. submission
10. _____ is a reading technique that aims at understanding and obtaining of a story or text.
    a. Close reading
    b. Scanning
    c. Skimming
    d. Studying
11. The final step in the pre-writing stage of the writing process is _____ .
    a. brainstorming
    b. clustering
    c. editing
    d. outlining
12. The stage in the communication process where the recipient seeks the correct meaning of the message is called _____ .
    a. channelling
    b. decoding
    c. feedback
    d. interpretation
13. The use of siren by the police, fire service or ambulance to suggest urgency of the situation is an example of
    a. Haptics
    b. Kinesics
    c. Objectics
    d. Oculesics
14. Converting an idea into written or spoken form of language is called _____ .
    a. Decoding
    b. Encoding
    c. Ideation
    d. Interpretation

15. Which of the following is NOT a feature of a good note?
    a. Descriptive
    b. Readable
    c. Reflects the source
    d. Understandable
16. All the following are ways of writing notes EXCEPT _____ .
    a. Detailing
    b. Headline
    c. Paraphrasing
    d. Spidergram
17. Which of the following reading techniques is employed at the Survey Stage of the SQ4Rs Method?
    a. Browsing
    b. Drifting
    c. Scanning
    d. Skimming
18. Which of the following is NOT a negative reading habit?
    a. Fixation
    b. Regression
    c. Stress
    d. Vocalization
19. The type of reading that is undertaken for academic and professional purposes is ___ .
    a. Diving
    b. Extensive reading
    c. Faster reading
    d. Intensive reading
20. Any paragraph that lacks Coherence would lack Unity.
    a. TRUE
    b. FALSE