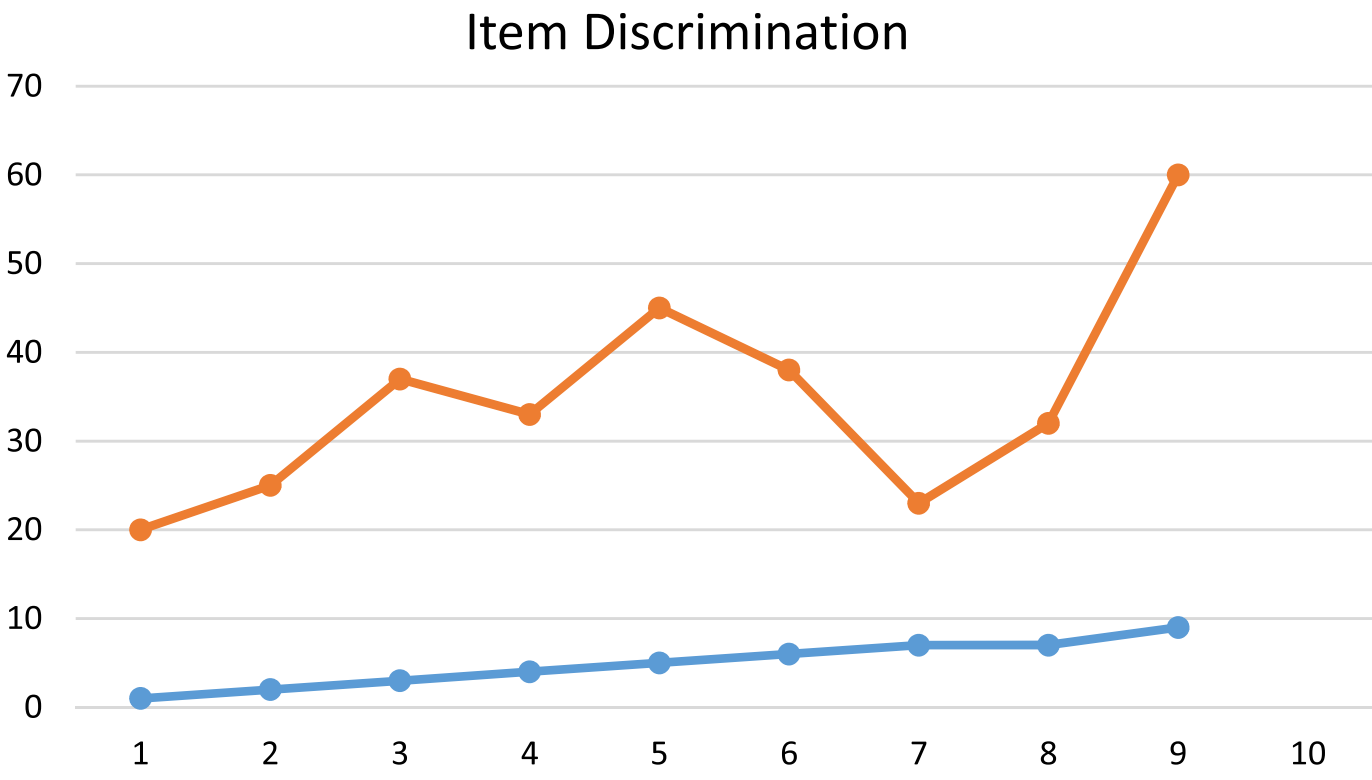
cogent education



EDUCATIONAL ASSESSMENT & EVALUATION | RESEARCH ARTICLE

Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation

Kennedy Quaigrain and Ato Kwamina Arhin

Cogent Education (2017), 4: 1301013

cogent oa

Quaigrain & Arhin, *Cogent Education* (2017), 4: 1301013
http://dx.doi.org/10.1080/2331186X.2017.1301013

cogent education

EDUCATIONAL ASSESSMENT & EVALUATION | RESEARCH ARTICLE

Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation

Kennedy Quaigrain[1] and Ato Kwamina Arhin[2]*

Received: 19 December 2016
Accepted: 23 February 2017
Published: 16 March 2017

*Corresponding author: Ato Kwamina Arhin, Department of Interdisciplinary Studies, University of Education Winneba - Kumasi Campus, P.O. Box 1277 College of Technology Education, Kumasi, Ghana
E-mail: quamyna@gmail.com

Reviewing editor:
Sammy King Fai Hui, The Education University of Hong Kong, Hong Kong

Additional information is available at the end of the article

Abstract: Item analysis is essential in improving items which will be used again in later tests; it can also be used to eliminate misleading items in a test. The study focused on item and test quality and explored the relationship between difficulty index (*p*-value) and discrimination index (DI) with distractor efficiency (DE). The study was conducted among 247 first-year students pursuing Diploma in Education at Cape Coast Polytechnic. Fifty multiple-choice questions were administered as an end of semester examination in Educational Measurement course. Internal consistency reliability of the test was 0.77 using Kuder–Richardson 20 coefficient (KR-20). The mean score was 29.23 with a standard deviation of 6.36. Mean difficulty index (*p*) value and DI were 58.46% (SD 21.23%) and 0.22 (SD 0.17), respectively. DI was noted to be a maximum at a *p*-value range between 40 and 60%. Mean DE was 55.04% (SD 24.09%). Items having average difficulty and high discriminating power with functional distractors should be integrated into future tests to improve the quality of the assessment. Using DI, it was observed that 30 (60%) of the test items fell into the reasonably good or acceptable value ranges.

ABOUT THE AUTHORS

Kennedy Quaigrain for 8–10 years has been the Ghana executive director for the international education non-government organization, Link Community Development. He is an education practitioner and has been involved in educational development and research in the past 15 years in Ghana. He holds a doctorate degree in Education Assessment from the University of Nottingham and taught at the University of Cape Coast and the University of Education in Ghana.

Quaigrain has facilitated the development and expansion of an innovative method of assessment, monitoring, and evaluation that builds on education accountability and student learning.

Ato Kwamina Arhin teaches at the College of Technology Education, University of Education, Kumasi Campus. Kwamina teaches Educational Assessment. He has a postgraduate degree in Educational Measurement and Evaluation from the University of Cape Coast, Ghana. Kwamina is interested in conducting classroom assessment-related research and promoting quality of education interventions.

PUBLIC INTEREST STATEMENT

The quality of teaching in most schools reflects the value teachers put on their test items. Teachers prepare and administer tests during the school year. Multiple-choice test has a lot of rules governing it construction. Strict adherence to the principles of test construction, test administration and analyses, and reporting is very essential, especially when norm-referenced tests are developed for instructional purposes. Item analysis is the process of collecting, summarizing, and using information from student's responses to assess the quality of test items. Difficulty index and discrimination index are two parameters which help to evaluate the standard of multiple-choice questions. In all, 5 items showed negative discrimination and 15 items had discrimination ranging from 0.1 to 1.9. The remaining 30 items were in the range of 0.2–0.49. That is acceptable to excellent discrimination. This research establishes the importance of item analysis to the classroom teacher.

Subjects: Mathematics & Statistics; Mathematics Education; Humanities

Keywords: difficulty index; discrimination index; distractor efficiency; item analysis; non-functional distractor (NFD)

1. Introduction

Teachers at all levels of education prepare and administer many formal teacher-made tests during the school year. Tests are, therefore, indispensable tools in the educational enterprise. Strict adherence to the principles of test construction, test administration and analyses, and reporting is very essential, especially when norm-referenced tests are developed for instructional purposes. But to what extent are teacher-made test reliable and valid? Anamuah-Mensah and Quaigrain (1998) state:

> In Ghana, the teacher is placed in a sensitive and central role in the testing and evaluation process. This makes it imperative for the classroom teacher to be adequately conversant in the formalized testing techniques to ensure sound facilitation of the herculean task that impinges on his or her profession. (Anamuah-Mensah & Quaigrain, 1998, p. 32)

This makes it imperative for teachers to be well versed in testing techniques to enable them to reliably and validly evaluate student progress. However, some best practices in item and test analysis are too infrequently used in teacher-made classroom tests. The two common test formats are essay and objective tests. Frequently used objective tests are multiple-choice, matching, and short-response items. In Ghana, essay items are becoming infrequently used (even in higher education) due to increasing student numbers. Multiple-choice questions (MCQs) and short-answer items are becoming very frequent.

Designing MCQs to assess students' knowledge comprehensively at the end of a semester is a complex and time-consuming process. Tests play an important role in giving feedback to teachers on their educational actions; therefore, the quality of the test is a critical issue. Having administered and scored a test, a teacher needs to know how good the test items are and whether the test items were able to reflect the students' performance in the course in relation to the specific learning objectives taught over the period of time. There are several ways for teachers to use assessment. Popham (2008) reiterates that teachers utilize assessments in order "to understand the student's prowess at the learning outcome, whether it is cognitive, affective, or psychomotor" (p. 256). Adjustments can, and should, be made to the instruction given to students based on assessments. These include modifications based on student needs, pace of instruction, coverage of course material, and developing a more effective and comfortable classroom-learning environment. Xu and Liu (2009) posited that teachers' knowledge in assessment and evaluation is not a static process but rather a complex, dynamic, and ongoing activity. These authors allude to the need for classroom teachers to constantly update their knowledge regarding assessment practices.

As stated earlier, MCQs are used widely in schools to assess students. A typical MCQ consists of a question or an incomplete statement, referred to as the stem, and a set of two or more options that consist of possible answers to the question. The student's task is to select the one option that provides the best answer to the question posed. The best answer is referred to as the key and the remaining options are called distractors. Only one option should be unequivocally correct and distractors should be unequivocally wrong**.** It is relatively difficult to write MCQs, especially creating good distractors. It is obvious that crafting MCQs is not a simple process, but it becomes very intricate searching for distractors that are plausible. In fact, the appropriate quality of MCQ is based on the availability of distractors. A good distractor should be able to discriminate between the informed and the uninformed student.

Ease of scoring can make multiple-choice (MC) testing particularly appealing to teachers who teach courses with large enrollments. Another advantage is that a well-constructed MC test can

yield test scores at least as reliable as those produced by a constructed-response test, while also allowing for a large portion of the topics covered in a course to be assessed in a short period of time (Bacon, 2003). Well-written MC items can serve to assess higher level cognitive processes, although creating such items does require more skill than writing memory-based items (Buckles & Siegfried, 2006; Palmer & Devitt, 2007). Also, a lot of effort and time is required to construct good quality MCQs as compared with essay questions. This is possible if the test constructor follows rigidly the numerous guidelines for writing MCQs.

However, multiple-choice items are often criticized for focusing on what students can remember and do not assess students' abilities to apply and analyze course-related information (Walsh & Seldomridge, 2006). Another criticism is that the format of MCQs let students guess even when they have no substantive knowledge of the topic under consideration (Biggs, 1999). However, Downing (2003) points out that blind guessing is quite uncommon on well-written classroom tests and informed guessing, which is based on a critical consideration of the question and the available options, provides a valid measure of student achievement.

One major concern in the construction of MCQs for examinations is the reliability of the test scores. Usually, good MCQs are those subjected to a rigorous process of item analysis. Item analysis is the process of collecting, summarizing, and using information from students' responses to assess the quality of the test items. Item analysis allows us to observe the characteristics of a particular item and can be used to ensure that items are of an appropriate standard for inclusion in a test, or else that the items need improvement.

Item analysis allows us to observe the item characteristics, and to improve the quality of the test (Gronlund, 1998). According to Lange, Lehmann, and Mehrens (1967) item revision allows to us identify items too difficult or too easy, items not able to differentiate between students who have learned the content and those who have not, or questions that have distractors which are not plausible. In this wise, teachers can remove these non-discrimination items from the pool of items or change the items or modify instruction to correct any misunderstanding about the content or adjust the way they teach. The present study was undertaken with an objective to assess item and test quality and to explore the relationship between difficulty and discrimination indices with distractor efficiency (DE).

2. Methodology

A total of 247 students participated in the test. The test consisting of 50 MCQs was based on Educational Assessment in schools. All respondents were first-year postgraduate students pursuing Diploma in Education program. It was conducted at Cape Coast Polytechnic during the 2016 academic session. All ethical standards were strictly adhered to. Students' responses from the MCQs were analyzed using Microsoft Excel. The MCQs were analyzed for their level of difficulty, measure of difficulty index (*p*-value), power of discrimination as measured by the discrimination index (DI), and distractor analysis for all non-correct options. The data analysis also included Kuder–Richardson formula and point biserial correlations. The results showed the reliability and quality of the test items included in the test. The Kuder–Richardson formula (KR-20) was used to assess internal reliability of the test scores. Most high-stakes tests have internal reliability values of 0.90 or higher, but teacher-made assessments generally have values of 0.80 or lower. According to Rudner and Schafer (2002), a teacher-made assessment needs to demonstrate reliability coefficients of approximately 0.50 or 0.60.

The time given for the test was 65 min. The items had four options, one of them being the correct answer and the other three being distractors. A correct answer was awarded a mark of 1 and there were no negative marks for the incorrect answer. Traditionally, it is recommended to use four or five options per item in order to reduce the effect of guessing. Most classroom achievement tests and international standardized tests (e.g. TOEFL) usually follow the rule four options per item. Thus, the maximum possible score of the test was 50 and minimum of 0.

3. Item analysis procedure

The result of the examinees' performance in a summative test was used to analyze the difficulty index and the DI of each multiple-choice item. The item difficulty index 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 item difficulty index, 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, 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. Which means the higher the difficulty index, the easier the item is understood to be. Those with a p-value between 20 and 90% are considered as good and acceptable. Among these, items with p-value between 40 and 60% are considered excellent, because DI is maximum at this range. Items with p-value (difficulty index) less than 20% (too difficult) and more than 90% (too easy) are not acceptable and need modification. It needs to be conceptualized that a p-value is basically a behavioral measure. Instead of explaining difficulty in terms of some intrinsic characteristic of the item, difficulty 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. Then, the total number of students in the upper 27% who obtained the correct responses and the lower 27% who obtained the correct responses were counted. The DI was calculated using the formula $\mathrm{DI} = \frac{\mathrm{UG}-\mathrm{LG}}{n}$, where UG is the number of students in the upper group who got an item correct and LG is the number of students in the lower group who got an item correct and n is the number of people in the largest of the two groups. The higher the DI the better the test item discriminates between the students with higher test scores and those with lower test scores.

Based on Ebel's (1979) guidelines on classical test theory item analysis items were categorized in their discriminating indices. As a rule of thumb:

(1) If $\mathrm{DI} \geq 0.40$, then the item is functioning satisfactorily.

(2) If $0.30 \leq \mathrm{DI} \leq 0.39$, then little or no revision is required.

(3) If $0.20 \leq \mathrm{DI} \leq 0.29$, then the item is marginal and needs revision.

(4) If $\mathrm{DI} \leq 0.19$, then the item should be eliminated or completely revised.

The DI 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. In the computation of the DI, D, every student's test is scored and the test scores are rank ordered. Subsequently, the 27% of the students at the top and the 27% at the bottom are separated for the analysis. The DI, D, is the number of students in the upper group who answered the item correctly minus the number of students in the lower group who answered the item correctly, divided by the number of people in the largest of the two groups. It is worthy of note that the higher the DI, the better the item. Removing these low discriminating items could seriously impair test validity. In this case, since the typical classroom test measures a variety of instructional objectives, we might expect to find that "low positive indices of discrimination are the rule rather than the exception" (Gronlund, 1985, p. 253).

Item discrimination indices must always be interpreted in the context of the type of test which is being analyzed. Items with low discrimination indices are often ambiguously worded and should be examined. Items with negative indices should be examined to determine why a negative value was obtained. According to Mehrens and Lehman (1991), there are a variety of reasons items may have low discriminating power:

(1) The more difficult or easy the item, the lower its discriminating power—but we often need such items to have adequate and representative sampling of the course content and objectives; and

(2) the purpose of the item in relation to the total test will influence the magnitude of its discriminating power (p. 888).

4. Results

The test consisted of 50 items. The scores of 247 students ranged from 11 to 42 (out of 50). The mean test score was 29.23 and the standard deviation was 6.36. The median score was 30 and the inter-quartile range value was 9. The median score is slightly greater than the mean score. The skewness and kurtosis values for the scores were −0.370 and −0.404, respectively. The values for asymmetry and kurtosis between −2 and +2 are considered acceptable in order to prove normal univariate distribution (George & Mallery, 2010). Mean scores according to groups were: lower: 20.62; middle: 29.28; upper: 36.55. The reliability measured by KR-20 was 0.77. The mean difficulty index was 58% that is $p = 0.58$.

A non-functioning distracter is defined as an option with either a response frequency of <5% or a positive discriminating power. Also, non-functioning distractors have a positive correlation with the total score. Gajjar, Sharma, Kumar, and Rana (2014) recommend that DE is determined for each item on the basis of the number of NFDs in it and ranges from 0 to 100%. If an item contains three or two or one or nil non-functional distractor (NFD) then DE will be 0, 33.3, 66.6, and 100%, respectively. Table 1 shows the summary of test statistics.

Table 2 shows the distribution of difficulty indices of the items. Majority of the items 47 (94%) were of acceptable level of difficulty with *p*-value within the range of 20–90%, while 22 (44%) items among them had excellent *p*-value (40–60%). Three items (6%) were identified to be too difficult (*p*-value < 20%).

Table 3 shows the frequency distribution of DI of the items. The item with the highest DI was item 25 (0.46) and the lowest DI was item 45 (−0.22). Eighteen of the items (36%) had good to excellent discrimination indices (DI ≥ 0.3). According to the analyses and as presented in Table 3, 20 (40%) of the items should be completely revised as they have very low discriminating power. A combination of the two indices (item difficulty and DI) show that 17 (34%) of the items could be called 'ideal' having p-values ranging from 20 to 80% as well as DI ≥ 0.3. However, if only the items with excellent *p*-value (40–64%) and excellent DI (≥0.4) are considered, there are 5 (10%) items which could be labeled as "excellent."

Figure 1 shows the relationship between difficulty index and DI. The graph indicates that as *p*-values increase, DI also increases. However, this increase occurs for *p*-values between 40 and 60% when DI reaches a maximum. Over the range, 40–60%, the DI is more than 0.4. When *p* is more than 60%, DI decreases. On the other hand, when *p* assumes values between less than 0.1 and 0.43 the DI values were between −0.11 and −0.21. Items with *p*-values closer to 0.50 are considered more useful in differentiating between individuals (Kline, 2005).

Table 1. Summary of test parameters

Parameter	Mean	SD
Difficulty index (p)	58.46	21.23
Discrimination index (DI)	0.22	0.17
Distractor efficiency (DE)	55.04	24.09

Table 2. Frequencies of difficulty indices

Difficulty indices	Frequency	Percentage
<10	1	2
10–19	2	4
20–29	4	8
30–39	3	6
40–49	5	10
50–59	6	12
60–69	11	22
70–79	11	22
80–89	5	10
90–99	2	4
Total	50	100

Table 3. Distribution of discrimination indices

Discrimination indices	Frequency	Percentage
<0	5	10
0.1–0.19	15	30
0.2–0.29	12	24
0.30–0.39	13	26
0.40–0.49	5	10
Total	50	100

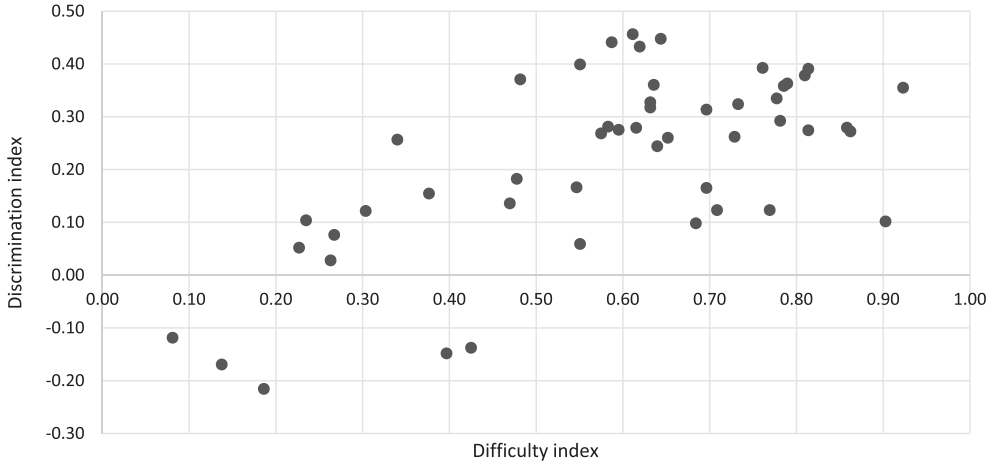


Figure 1. Scatter plot showing relationship between difficulty index and DI of items.

Five items had negative discrimination indices. Also, MCQ 18 can be considered as very easy (p = 92%). However, four items were relatively difficult (p = 23%). Reasons for negative DI can be wrong key, or ambiguous framing of the question. Matlock-Hetzel (1997) posits that items with negative DI are not only useless, but they also reduce the validity of the test.

From Figure 1 is it can be observed that as the items get easy (p = 75% above) the level of DI decreases. Also, as the discrimination indices turn negative then the difficulty index increases.

Table 4. Frequency distribution of functioning distractors

	Frequency	Percentage
Number of items	50	
Number of distractors	150	
Distractors with frequency > 5%	38	25.3
Distractors with discrimination ≥ 5	112	74.7
Functioning distractors per item		
0	24	48
1	14	28
2	11	22
3	1	2

Kehoe (1995) gave the following suggestions for improving the discrimination power of items;

(1) Items that correlate less than 0.15 with total test score should probably be restructured. Perhaps, such items do not measure the same skill or ability as does the test on the whole or that they are confusing or misleading to examinees.

(2) Distractors that are not chosen by any examinees should be replaced or eliminated. They are not contributing to the test's ability to discriminate the good students from the poor students. One should be suspicious about the correctness of any item in which a single distractor is chosen more often than all other options, including the answer, and, especially, so if that distractor's correlation with the total score is positive.

(3) Items that virtually everyone gets right are useless for discriminating among students and should be replaced by more difficult items. Although some very easy early items may be useful to get students settled and more confident about the test. May also be necessary for validity.

4.1. Analysis of distractors

In all 150 distractors were assessed, 38 distractors (25.3%) had a choice frequency of <5%. On the other hand 112 distractors (74.7%) had a frequency choice of ≥5. Items that had positive discrimination indices were 45 (90%). The essence of the distracter analysis was to identify non-functioning and functioning options. A non-functioning option was defined as one that was chosen by fewer than 5% of examinees (Table 4).

The functionality of the distracters serves as an independent indicator of item functioning. Distracters which are chosen by one or more examinees are called functioning distracters and those not chosen by anyone are called non-functioning distracters. Designing options with equal plausibility is a difficult task, especially in an end of semester examination. The functionality of distracters, the flaws in item writing, and the optimum number of options are interrelated and affect the item quality, item performance, and the test results. Tarrant, Ware, and Mohammed (2009) concluded that items with two functioning distracters were more difficult than items with three functioning distracters.

5. Discussion

To achieve instructional validity, classroom instruction must synchronize with the test items. This requires developing good test items and also analyzing the items. Mozaffer and Farhan (2012) emphasized the importance of teachers understanding and using statistical analysis of test materials in order to improve their teaching strategies and test construction. The difficulty and discrimination indices are among the tools teachers can use to check whether the MCQs are well constructed or not. Another tool used for further analysis is the DE which analyses the quality of distractors. In the present study, the mean *p*-value was 58% which is within the range of excellent level of difficulty (p = 40–60%). The mean DI found in this study was 0.22 which is considered reasonably good.

The distractors are analyzed to determine their relative usefulness in each item. If students consistently fail to select certain multiple choice options it may be that those options are perhaps implausible and, therefore, of little use as foils in multiple-choice items. Therefore, designing of plausible distractors and reducing the NFDs is important aspect for framing quality MCQs.

Distractors are important components of an item, as they show a relationship between the total test score and the distractor chosen by the student. Student's performance depends on upon how the distractors are designed (Dufresne, Leonard, & Gerace, 2002).

According to Tarrant et al. (2009) DE is one such tool that tells whether the item was well constructed or failed to perform its purpose. Any distractor that has been selected by less than 5% of the students is considered to be a non-functioning distractor (NFD).

Gronlund and Linn (1990) observed that low-scoring students, who have not grasped the subject content, should choose the distractors more often, whereas, high scorers should reject them more often while choosing the correct option. Classroom teachers can use this powerful technique, to help them modify or remove specific items from the test. When this is properly done the modified items can be used for subsequent exams.

Tarrant et al. (2009) asserted that flawed MCQ items affect the performance of high-achieving students more than borderline students. Constructing balanced MCQs, therefore, addresses the concerns of the students of getting an acceptable average grade, (Carroll, 1993). Rodriguez (2005) says that numbers of NFDs also affect the discriminative power of an item. It is seen that reducing the number of distractors from four to three decreases the difficulty index while increasing the DI and the reliability.

The current study shows that items having one NFD had good discriminating ability (DI = 0.26) as compared to items with all three functioning distracters (DI = 0.17). This compares well with other studies favoring better discrimination by three distracters as compared to four (Tarrant et al., 2009). This could be true because writing items with four distractors is a difficult task and writing a good fourth distractor is usually difficult. It is like just trying to fill the gap and the fourth distractor has a high propensity to become the weakest distractor. From the present study, DI also increases. However, this increase occurs for *p*-values between 40 and 60% where DI reaches a maximum.

6. Conclusion and recommendation

Constructing multi-choice test items for an end of semester examination requires time and carefully selecting content that will produce the desired test results. In conclusion, item analysis provides valuable information for further item modification and future test development.

Quality control is important for test development. It is therefore, important for teachers to perform item analysis or seek assistance where they feel inadequate. According to Bonnel and Boureau (1985), performance in an examination should reflect only the proficiency in the target construct and no other irrelevant ones. Items should be modified if students consistently fail to select certain multiple-choice alternatives. This means that the test should be unidimensional.

Items with negative discrimination indices must be deleted or replaced. Classroom teachers must rewrite all items with zero discrimination indices. Also, teachers must replace or rewrite all items with low positive discrimination indices. When all these are done the reliability of the test will be increased.

Funding
The authors received no direct funding for this research.

Author details
Kennedy Quaigrain[1]
E-mail: kennedyquaigrain@gmail.com
Ato Kwamina Arhin[2]
E-mail: quamyna@gmail.com
ORCID ID: http://orcid.org/0000-0002-7702-5869
[1] Department of Educational Studies, University of Education Winneba, Kasoa, Ghana.
[2] Department of Interdisciplinary Studies, University of Education Winneba - Kumasi Campus, P.O. Box 1277 College of Technology Education, Kumasi, Ghana.

Citation information
Cite this article as: Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation, Kennedy Quaigrain & Ato Kwamina Arhin, *Cogent Education* (2017), 4: 1301013.

Cover image
Source: Author

References

Anamuah-Mensah, J., & Quaigrain, K. A. (1998). Teacher competence in the use of essay test. *The Oguaa Educator University of Cape Coast, 12*, 31–42.
Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education, 25*, 31–36. http://dx.doi.org/10.1177/0273475302250570
Biggs, J. (1999). *Teaching for quality learning at university*. Buckingham: Society for Research into Higher Education and Open University Press.
Bonnel, A. M., & Boureau, F. (1985). Labor pain assessment: Validity of a behavioral index. *Pain, 22*, 81–90.
Buckles, S., & Siegfried, J. J. (2006). Using multiple-choice questions to evaluate in-depth learning of economics. *The Journal of Economic Education, 37*, 48–57. http://dx.doi.org/10.3200/JECE.37.1.48-57
Carroll, R. G. (1993). Evaluation of vignette-type examination items for testing medical physiology. *American Journal of Physiology, 264*, 11–5.
Downing, S. (2003). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in media education. *Advanced in Health Sciences Education Theory and Practice, 10*, 133–143.
Dufresne, R. J., Leonard, W. J., & Gerace, W. J. (2002). Marking sense of students' answers to multiple-choice questions. *The Physics Teacher, 40*, 174–180. http://dx.doi.org/10.1119/1.1466554
Ebel, R. L. (1979). *Essentials of educational measurement* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.
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 Ahmedabad, Gujarat. *Indian Journal of Community Medicine, 39*, 17–20.
George, D., & Mallery, M. (2010). *SPSS for windows step by step: A simple guide and reference* (10a ed.). Boston, MA: Pearson.
Gronlund, N. E. (1985). *Measurement and Evaluation in Teaching*. New York, NY: Macmillan.
Gronlund, N. E. (1998). *Assessment of student achievement* (6th ed.). Boston, MA: Allyn and Bacon.
Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). New York, NY: Macmillan.
Hotiu, A. (2006). *The relationship between item difficulty and discrimination indices in multiple-choice tests in a Physical science course* (MSc thesis). Florida Atlantic University, Boca Raton, FL. Retrieved June 14, 2015 from www.physics.Au/eAdu/research/education/A.Hotiu_thesis.pdf
Kehoe, J. (1995). Basic item analysis for multiple-choice tests. *Practical Assessment, Research & Evaluation, 4*(10). Retrieved from http://PAREonline.net/getvn.asp?v=4&n=10
Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, CA: Sage.
Lange, A., Lehmann, I. J., & Mehrens, W. A. (1967). Using item analysis to improve tests. *Journal of Educational Measurement, 4*, 65–68. Retrieved from http://www.jstor.org/stable/1434299 http://dx.doi.org/10.1111/jedm.1967.4.issue-2
Matlock-Hetzel, S. (1997). *Basic concept in item and test analysis*. A paper presented at annual meeting of the Southwest Educational Research Association. Retrieved from www.ericae.net/ft/tamu/espy.htm
Mehrens, W. A., & Lehman, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). Belmont, CA: Wadsworth/Thomson Learning.
Mozaffer, R. H., & Farhan, J. (2012). Analysis of one-best MCQs: The Difficulty index, discrimination index and distractor efficiency. *Journal of Pakistan Medical Association, 62*, 142–147. Retrieved from http://jpma.org.pk/PdfDownload/3255.pdf
Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: Modified essay or multiple choice questions? Research paper *BMC Medical Education, 7*, 329. doi:10.1186/1472-6920-7-49
Popham, J. W. (2008). *Classroom assessment: What teachers need to know* (5th ed.) Boston, MA: Pearson Education, Inc. Research into Higher Education and Open University Press.
Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*, 3–13. http://dx.doi.org/10.1111/j.1745-3992.2005.00006.x
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
Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BMC Medical Education, 9*, 861. http://dx.doi.org/10.1186/1472-6920-9-40
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.
Walsh, C. M., & Seldomridge, L. A. (2006). Critical thinking: Back to square two. *Nursing Education, 45*, 212–219.
Xu, Y., & Liu, Y. (2009). Teacher assessment knowledge and practice: A narrative inquiry of a Chinese College EFL. *TESOL Quarterly, 43*, 493–513.

Cogent Education **(ISSN: 2331-186X) is published by Cogent OA, part of Taylor & Francis Group.**