E-ISSN: 2476-9606
Abstract Proceedings International Scholars Conference
Volume 7 Issue 1, October 2019, pp.499-511
https://doi.org/10.35974/isc.v7i1.1128

Analysis of Multiple-Choice Questions (MCQs): Item and Test Statistics from the 2nd Year Nursing Qualifying Exam in a University in Cavite, Philippines

April M. Obon[1], Kristel Anne M. Rey[2]
College of Nursing, Adventist University of the Philippines
AMObon@aup.edu.ph

ABSTRACT

Multiple Choice Questions (MCQs) is used extensively as a test format in nursing education. However, making MCQs still remains a challenge to educators. To avoid issues about its quality, this should undergo item analysis. Thus, the study evaluated item and test quality using difficulty index (DIF) and discrimination indices (DI), with distractor efficiency (DE); determined the reliability using Kuder-Richardson 20 coefficients (KR20); and identified which valid measure was developed. The study was conducted among 41 level two nursing students in the College of Nursing. The qualifying examination comprised of 194 MCQs. Data were entered in Microsoft Excel 2010 and SPSS22 and were analyzed. According to DIF, out of 194 items, 115 (59.3%) had right difficulty and 79 (40.7%) were difficult. Regarding DI, 17 (8.8%) MCQs were considered very good items to discriminate the low and high performer students. While 21 (10.8%), 32 (16.5%), 24 (12.4%), and 100 (51.5%) demonstrated good, fair quality, potentially poor, and potentially very poor items, respectively. On the other hand, the number of items that had 100% distractor effectiveness is 57 (29.4%), as 65 (33.5%), 49 (25.3%), and 23 (11.9%) revealed 66.6%, 33.3% and 0%, respectively. The reliability of the test using KR20 is 0.9, suggesting that the test is highly reliable with considered good internal consistency. After careful analysis of each item, 55 (28.35%) items were retained without revisions. Further, the stem of the 24 (12.37%) items, the distractors of the 66 (34.02%) items and both the stem and distractors of 46 (23.71%) items were modified, and 3 (1.55%) items were removed. The researcher recommends doing an analysis between upper and lower scorers and its relationship to DE. For future study, it will be beneficial to explore other factors like student's ability, quality of instructions, and number of students in relation to quality of MCQs.

Keywords: Difficulty index, Item discrimination, Distractor efficiency, Item analysis

INTRODUCTION

In nursing education multiple choice questions (MCQs) are often used to assess knowledge among nursing students (D'Sa & Visbal-Dionaldo, 2017; Mannion, Hnatsyhyn, O'Rae, Beck & Patel, 2018). Although MCQ's are easy to check and analyze particularly on large number of students, they are often difficult technically and time consuming (Mannion et al., 2018;

Odukoya, Adekeye, Igbinoba, & Afolabi, 2017). Nurse educators are expected to be competent in constructing MCQ's and more than that they should be adept in analyzing whether items are valid and reliable in assessing student's learning (Hijji, 2017) because concerns regarding the quality of tests used for assessment is now increasing (D'Sa & Visbal-Dionaldo, 2017). However, only a few nurse educators have been trained specifically to develop quality MCQs and the skill in performing item analysis (Mannion et al., 2018). Poorly constructed MCQ's without item analysis could jeopardize the integrity of MCQ's (Odukoya et al., 2017; Rehman, Aslam, Hassan, 2018). Further, inaccurate evaluation could impact the grade of the students which if often final and irreversible impeding the career pathway of students (Reichert, 2011). In the study done by Nedeau-Cato, Laughlin, and Rus (2013) in a nursing school it was found that 85% of items have at least one flaw. Another related studies in which exams who did not underwent items analysis resulted to 91.8% of the items have one or more items that are flawed (Hijji, 2017). It seems like conducting item analysis is essential for every exam to reduce errors and improve integrity of each exam. It is imperative that the quality of MCQ's should be evaluated regularly. Item analysis is done by analyzing four components namely: Difficulty Index (DIF), Discrimination Index (DI), Distractor Efficiency (DE), and reliability test using the Kuder-Richardson Formula 20 (KR20). Hence, the study evaluated item and test quality using difficulty index (DIF) and discrimination indices (DI), with distractor efficiency (DE); determined the reliability using Kuder-Richardson 20 coefficients (KR20); and identified which valid measure was developed.

LITERATURE REVIEW

The quality of test items can be assessed using item analysis. This will help improve the teacher's ability in creating test items. Item analysis can be used to evaluate if the item is difficult or easy (Tracy, 2012). According to Polit and Yang (2015), item analysis is done to evaluate which items to discard, to retain, and needs revision. Therefore, a test needs item analysis to evaluate its performance.

Difficulty index corresponds the proportion of students who correctly answered the item (Mahjabeen et al., 2018; Mukherjee & Lahiri, 2015). The formula used to calculate the DIF is

$$p = \frac{\mathrm{Ru} + \mathrm{Rl}}{\mathrm{T}} x100$$

Where:

RU = the number in the upper group who answered the item correctly.

RL = the number in the lower group who answered the item correctly.

T = the total number who tried the item

The range of DIF is from 0-1 and when multiplied to 100 the p-value is converted to percentage, which the percentage of students who answered the item correctly (Mahjabeen et al., 2018; Mukherjee & Lahiri, 2015). The higher the value, it means that the question is easy. According to Mahjabeen et al. (2018), overall, if the p value is between 20-90%, the question is regarded as good and acceptable and items with p-value between 40-60% are viewed as excellent. Further, items with p - value of less than 20% is too difficult and more than 90% is too easy, which are not acceptable and needs revision. While to Mukherjee & Lahiri (2015), items with DIF with >70% is too easy, between 30-70% is average, between 50-60% is good, and <30% is too difficult. Table 1 will show the range of DIF used in the study from (source).

Table 1. **Range of Difficulty Index Used in the Study**

Range of Difficulty Index	Interpretation	Action
0 – 0.25	Difficult	Revise or Discard
0.26 – 0.75	Right Difficulty	Retain
0.76 – above	Easy	Revise or Discard

Item Discrimination

The DI is used to gauge the effectiveness of an item in the MCQ’s in discriminating the students from high performing students to low performing students (Mukherjee & Lahiri, 2015; Mahjabeen, Alam, Hussan, Zafar, Butt, Konain, & Rizvi, 2018). In getting the DI, the test takers were divided into quartiles. The upper quartile or students who made the highest scores, lower quartile who made the lowest scores, and the students who has average scores or the middle quartile. In calculating the DI, first the DIF must be computed for the upper and lower group, then get the difference of DIF between upper and lower quartiles (Mukherjee & Lahiri, 2015). The formula used was:

$$\text{Index of Discrimination} = DU - DL$$

In item discrimination, the value ranges from -1 to +1. The item is considered to be effective and is discriminating if it has a higher value (Mukherjee & Lahiri, 2015; Musa, Shaheen, Elmardi, & Ahmed, 2018). Mukherjee and Lahiri (2015) explained that if all the test takers in the upper quartile and not in the lower quartile will answer the item correctly, the DI value is 1.00. On the other hand, if the lower group will answer it correctly and none from the upper group, the DI value would be -1.00 (D'Sa & Visbal-Dionaldo, 2017; Mukherjee & Lahiri,

2015) maybe due to item flaws or inefficient distractors (D'Sa & Visbal-Dionaldo, 2017). Therefore, according to Rasiah and Isaiah (as cited in Musa et al., 2018) these “items should be carefully reviewed for the presence of common causes of poor discrimination such as ambiguous wording, grey areas of opinion, wrong keys and areas of controversy” (p. 1478). Musa et al. (2018) added that if the DI value is 1.00, it indicates a perfect discrimination between high and low performing students and if the value is near or less than zero, the item should be removed from the exam. Overall, items with DI value greater than 0.40 are considered as excellent, 0.30-0.39as reasonably good but probably needs improvement, 0.20 to 0.29 are marginal items and should be reviewed, while items with below 0.19 are considered poor and must be removed (Mukherjee & Lahiri, 2015). Mahjabeen et al. (2018) categorized DI as items with value of ≥0.36 are excellent, between 0.25 to 0.35 as good, between 0.21 to 0.24 as acceptable, and items that are ≤0.20 are poor. In this study, the range of item discrimination used is shown in table 2.

Table 2. **Range of Item Discrimination Used in the Study**

Range of Discrimination Index	Quality of an Item	Action
≥0.50	Very Good Item	Definitely Retain
0.40 – 0.49	Good Item	Very Usable
0.30 – 0.39	Fair Quality	Usable Item
0.20 – 0.29	Potentially Poor Item	Consider Revising
≤0.20	Potentially Very Poor	Possibly Revise Substantially or Discard

Distractor Efficiency

“Distractor efficiency is the ability of incorrect answers to distract the students” (Mahjabeen et al., 2018, p. 312). There are two types of distractors namely non-functional distractors (NFD) and functional distractors (FD). The options are considered NFD if <5% of the examinees infrequently choose the incorrect answers, and FD if the option is selected by 5% or more students. A DE can be ascertained on the basis of the number of NFD present in an item and the range of DE is 0-100%. If an item has 3 or more NFDs, the DE is considered 0%. However, if an MCQ has two, one or none NFD the DE can be labeled as 33.3%, 66.6%, and 100% respectively (Mukherjee & Lahiri, 2015; Mahjabeen et al., 2018).

Test Reliability

To estimate the internal consistency of reliability of the MCQ's, the formula developed by Kuder-Richardson with two versions Kuder-Richardson 20 was used. The formula KR20 is used to calculate reliability of test items with a range of difficulty, whereas KR21 is used for test items with same difficulty (Riazi, 2016). Therefore, to check the consistency of the MCQs, KR20 was used. According to Polit and Yang (2015) the formula for KR20 is:

$$KR20 = [n/(n-1)] \times [1 - (\sum pq)/Var]$$

Where:

KR20	=	estimated reliability of the full-length test
n	=	number of items
Var	=	variance of the whole test (standard deviation squared)
∑pq	=	sum of the product of pq for n items
p	=	proportion of people passing the item
q	=	proportion of people failing the item (or 1 – p)

The value of reliability can range from zero to 1.00 (Mukherjee & Lahiri, 2015; “Understanding Item Analyses,” 2019) and the numbers closer to 1.00 can suggest greater internal reliability which indicates that the items are all measuring the same thing (Mukherjee & Lahiri, 2015) or the questions tend to “pull together” and low reliability means that the items are unrelated to each other in terms of who answered it correctly. Table number 3 shows the standard to interpret the reliability coefficients for educational tests and measurements. (“Understanding Item Analyses,” 2019).

Table 3. **Guideline to Interpret Reliability Coefficients**

Reliability	Interpretation
.90 and above	Excellent reliability; at the level of the best standardized tests
.80 - .90	Very good for a classroom test
.70 - .80	Good for a classroom test; in the range of most. There are probably a few items which could be improved.
.60 - .70	Somewhat low. This test needs to be supplemented by other measure (e.g., more tests) to determine grades. There are probably some items which could be improved.
.50 - .60	Suggests need for revision of test, unless it is quite short (ten or fewer items). The test definitely needs to be supplemented by other measures (e.g., more test) for grading
.50 or below	Questionable reliability. This test should not contribute heavily to the course grade, and it needs revision.

METHODS

This section discusses the research design, population and sampling technique, instrumentation, data gathering procedures, and statistical treatment.

Research Design

Descriptive validation design was applied in the study that allows the researchers to describe in detail the outcome of the study (Houser, 2018) and examine whether the instrument used is consistent and measures the right thing it is intended to measure (Houser, 2018; Polit & Yang, 2015).

Population and Sampling Technique

The study was conducted in Level II, College of Nursing as a qualifying examination in July 2018. Forty-one (41) second year nursing student candidates for promotion to third year took the qualifying examination and answered the test items. The examination was done to assess how much and how well the students learned during their second year in Nursing for them to be promoted in third year. Thus, purposive sampling was used in the study. Purposive sampling is a non-probability sampling in which the participants are selected based on characteristics of a population and the objective of the study (Crossman, 2018).

Instrumentation

The exam includes NCMN 213 – Care of Mother, Child, and Family, NCMN 224 – Care of Mother, Child, Family, Community, and Population Group at Risk or with Problems, CHNN – Community Health Nursing, PHAN – Pharmacology, and HEED - Health Education. After developing the test items, it was moderated by the 4 clinical instructors all of which have earned their master’s degree in nursing. The exam is originally 200 items, but four items are non MCQs and one has five options. Hence, the total number of items analyzed was 194 MCQs. The paper comprised of 194 multiple choice questions, each having a single stem with four options including one correct answer and three distractors (incorrect answers). Each item was assigned with one mark. The highest possible score was 194 and minimum was zero, with no negative marking.

Data Gathering Procedure

After finalizing the MCQs, the second-year students who were qualified took the examination on their scheduled date. The exam was divided into two parts. The first part was taken in the morning and the second part in the afternoon. When everyone was done taking the exam, the answer sheets were checked using a scantron machine Students' answers in each item were encoded in Microsoft Excel and SPSS 22 for analysis of DIF, DI, DE, and KR20. Items that were non MCQs and with five options were not included. Item analysis was done and the result of all papers was ranked from highest to lowest scores. Then the result was divided into quartiles. Upper quartile or higher scored (n = 11) and lower quartile or low scored (n=11) groups were included into the analysis, while averaged scores or middle quartiles (n=19), but they were excluded in the study. Each item was analyzed using DIF, DI, and DE. Also, reliability of the test was assessed by estimating the Kuder-Richardson Formula 20 (KR20) coefficients. When the results were ready, each item were moderated by four clinical instructors two of which earned master's degree in nursing and the other two has doctoral degree. Moderation was done to evaluate which items are valid. After careful analysis, there were items that were retained, revised, and discarded and replaced.

Statistical Treatment

Post validation of the paper was done by item analysis. Each item was encoded to Microsoft Excel and SPSS22 and was analyzed using item statistics: Difficulty Index (DIF), Discrimination Index (DI), Distractor Efficiency (DE). Also, reliability of the test was assessed by estimating the Kuder-Richardson Formula 20 (KR20) coefficients.

RESULTS

A total of 194 MCQs were analyzed. Score of 42 students ranged from 71 to 151. Table 4 shows the classification MCQs according to Difficulty index (DIF) and actions proposed. Table 4 shows that out of 194 items, 115 (59.52%) has right difficulty and 79 (40.7%) were difficult. The results of all parameters per item can be seen in Appendix 1.

Table 4. **Classification of MCQs according to difficulty index and actions proposed**

DIF	No. of Items (%)	Interpretation	Proposed Action
0 – 0.25	79 (59.3%)	Difficult	Revise or Discard
0.26 – 0.75	115 (40.7%)	Right Difficulty	Retain

Table 5 shows the classification MCQs according to Item Discrimination (DI) and actions proposed. Table 5 reveals that out of 194 items, Regarding DI, 17 (8.8%) MCQs were considered very good items to discriminate the low and high performer students. While 21 (10.8%) items demonstrated good, 32 (16.5%) items are fair quality, 24 (12.4%) are potentially poor items, and 100 (51.5%) potentially very poor items. The results of all parameters per item can be seen in Appendix 1.

Table 5. **Classification of MCQs according to item discrimination and actions proposed**

DI	No. of Items (%)	Interpretation	Proposed Action
≥0.50	17 (8.8%)	Very Good Item	Definitely Retain
0.40 – 0.49	21 (10.8%)	Good Item	Very Usable
0.30 – 0.39	32 (16.5%)	Fair Quality	Usable Item
0.20 – 0.29	24 (12.4%)	Potentially Poor Item	Consider Revising
≤0.20	100 (51.5%)	Potentially Very Poor	Possibly Revise Substantially or Discard

Table 6 reflects the distractor analysis of MCQs. It shows that out of 776 distractors, 234 (30.1%) were non-functional indicating that these distractors were chosen by less that 5% and 542 (69.9%) items were chosen by 5% or more which are considered to be functional. On the other hand, table 7 reveals the percentage non-functional distractors of MCQs according to distractor efficiency. The number of items that had 100% distractor effectiveness is 57 (29.4%), as 65 (33.5%) items had 66.6% DE, 49 (25.3%) items had 33.3%, and 23 (11.9%) had 0%.

Table 6. **Number of distractors and categorization of MCQs**

DE	No. of Items (%)	Interpretation	Proposed Action
<5%	234 (30.1%)	Non-functional Distractors	Revise or Discard
5% or more	543 (69.9%)	Functional Distractors	Retain

Table 7. **Percentage of non-functional distractors of MCQs according to distractor efficiency**

Number of Non-Functional Distractors	No. of Items (%)	Distractor Efficiency
0 NFD	57 (29.4%)	100%
1 NFD	65 (33.5%)	66.6%
2 NFD	49 (25.3%)	33.3%
3 NFD	23 (11.9%)	0%

Table 8. **Reliability of the MCQs using KR20**

KR20 Reliability	Result	Interpretation
0.8 - .90	0.90	Very good for a classroom test

Table 8 reflects the reliability of the test using KR20. The reliability of the test is 0.90, suggesting that the test is highly reliable with considered good internal consistency. Furthermore, after careful analysis of each item table 9 shows the actions done in each item. Fifty-five (28.35%) items were retained without revisions, the stem of the 24 (12.37%) items, the distractors of the 66 (34.02%) items and both the stem and distractors of 46 (23.71%) items were modified, and 3 (1.55%) items were removed.

Table 9. **Actions performed after analysis of MCQs**

No. of Items (%)	Actions Done
55 (28.35%)	Both Stem and distractors were retained without revisions
24 (12.37%)	Stem were modified Distractors were retained without revisions
66 (34.02%)	Distractors were modified Stem were retained without revisions
3 (1.55%)	Removed

DISCUSSION

It is recommended that an effective method to increase the validity of examination is to develop MCQs with right difficulty, high discrimination power with increase distractor efficiency (Mahjabeen et al., 2017; Mehta & Mokhasi, 2014). In the current study, most of the items DIF has right difficulty and has poor discrimination, the average functional distractor per item is 3 out four. The result of the current study is different from most recent related study in which the item difficulty is considered in an acceptable level with a high discrimination level (Mahjabeen et al. 2017; Mukherjee & Lahiri, 2015). The main differences of the previous studies from the current study were that the number of items were ranging from 30 to 65 items comparing to the current which is 194 items. As for the number of the students, there are not much differences in the number of students who have taken the MCQs ranging from 40-247 participants. Oermann and Gaberson (2014) said that item discrimination power does not indicate item validity. Moreover, DIF and DI are constantly changing per administration because it is influenced by other factors such as student's ability level, quality of instructions, and the size of the group tested. Hence, teacher should also consider assessing the efficiency of the distractors of each item. Finally, According to Miller et al (as cited in Oermann and Gaberson, 2014), educators must be careful in deleting items with poor results in DIF and DI because it could negatively impact the validity of the exam due to fewer sample content. Apparently, although the average DI and DIF results are not desirable, the DE of the exam in totality is considered "good" meaning that items should be moderated carefully on deciding whether an item be included or not in the test bank even the DIF and DIF are not "ideal".

Item analysis will not be complete without the analysis of the distractors because the presence of distractors itself enhances the measurement properties of each item. The result of the present study is comparable with the previous study conducted by Mahjabeen et. al (2017) where in there are more functional distractors (72%) compared to non-functional distractors (28 %). Salkind (2010) said that a successful distractor seems to attract lower scoring group as an answer, while distractors answered by the higher scoring group mean that either there are two possible items or the right answer should be rechecked (Oerman & Gaberson, 2014; Salkind, 2010). On the other hand, if there are none or few students have answered a certain option, it should be replaced or revised because it does not positively contributing to the measurement of the test item unless the item is mastered by the class to which that particular distractor belongs (Salkind, 2010). It seems that the result of study particularly in the distractor is considered ideal in standardizing an MCQ exam.

KR 20 was particularly used to determine the internal consistency of the exam because it is but fitting due to varying difficulties of items included in the test (Riazi, 2016). Salkind (2010) said that 0.7 is acceptable to short test with less than 50 items but with more than 50 item-test, a KR20 of 0.8 is ideal. The present study with a 194 items MCQs has a reliability of 0.9 suggesting that the test is highly reliable with considered good internal consistency. Meaning the test measures reflect the underlying construct which are maternal and child nursing 1 and 2, Health Education and community health nursing. Hence, making the scores consistent or correlated each time the test is administered.

After careful analysis each item, there were actions done to improve the examination before it is given again to another set of students. According to Burud, Nagandla, and Agarwal (2019), one of the important features of quality assurance of an examination is by implementing item analysis. The decision to revise the stem and distractors of each item must be based on difficulty index, discrimination index, and distractor efficiency.

Conclusion

This study is set out to evaluate item and test quality using index (p-value) and discrimination indices (DI), with distractor efficiency (DE) and KR 20 or 21. The results revealed that the average DIF (0.12) and DI (0.22) are considered not ideal. On the other hand, the average DE is 60.1 % meaning that most of the items have three functional and only 1 non-functional distractors. Further, upon determining the KR20 to test the reliability of the result, it showed that the test has high internal consistency with value of 0.90.

The results of the study give an evidence-based insight on deciding whether items should be moderated, included, or excluded in the test bank for the qualifying exam of level two nursing students. To improve the quality of the exam it is recommended to moderate the existing exams especially items which are difficult has low DI and few functional distractors. It also important that test construction and measurement should be included as one of the topics in faculty development program to increase quality of MCQs. For the analysis of the test, the researcher recommends doing an analysis between upper and lower scorers and its relationship to DE. For future study, it will be beneficial to explore other factors like student's ability, quality of instructions and number of students in relation to quality of MCQs.

REFERENCES

Barud, I., Nagandla, K., & Agarwal, P. (2019). Impact of distractors in item analysis of multiple choice questions. *International Journal of Research in Medical Sciences*, *7*(4), 1136-1139. doi: http://dx.doi.org/10.18203/2320-6012.ijrms20191313

Crossman, A. (2018). *What you need to understand about purposive sampling*. Retrieved from https://www.thoughtco.com/purposive-sampling-3026727

D'Sa, J. L. & Visbal-Dionaldo, M. L. (2017). Analysis of multiple choice questions: Item difficulty, discrimination index and distractor efficiency. *International Journal of Nursing Education*, *9*(3), 109-114. doi:10.5958/0974-9357.2017.00079.4.

Hijji, B. M. (2017). Flaws of multiple choice questions in teacher-constructed nursing examinations: A pilot descriptive study. *Journal of Nursing Education*, *56*(8), 490-496. doi: 10.3928/01484834-20170712-08

Houser, J. (2018). *Nursing research: Reading, using, and creating* (4th ed.). Burlington, MA: Jones & Bartlett Learning.

Mahjabeen, W., Alam, S., Hussan, U., Zafar, T., Butt, R. Konain, S., & Rizvi, M. (2018). Difficulty index, discrimination index, and distractor efficiency in multiple choice questions. *Annals of Pakistan Institute of Medical Sciences, 4,* 310-315. Retrieved from https://www.researchgate.net/publication/323705126_Difficulty_Index_Discrimination_Index_and_Distractor_Efficiency_in_Multiple_Choice_Questions

Mannion, C. A., Hnatyshyn, T., O'Rae, A., Beck, A. J., & Patel, S. (2018). Nurse Educators and Multiple-Choice Examination Practices. Retrieved from http://hdl.handle.net/1880/108887

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

Mukherjee, P. & Lahiri, S. K. (2015). Analysis of multiple-choice questions (MCQs): Item and Test Statistics from an assessment in a medical college of Kolkata, West Bengal. *Journal of Dental and Medical Sciences*, *14*(12), 47-52. www.iosrjournals.org

Musa, A., Shaheen, S., Elmardi, A., & Ahmed, A. (2018). Item difficulty & item discrimination as quality indicators of physiology MCQ examinations at the Faculty of Medicine, Khartoum University. *Khartoum Medical Journal*, *11*(02), 1477-1486. Retrieved from https://www.researchgate.net/publication/328583573_Item_difficulty_item_discrimination_as_quality_indicators_of_physiology_MCQ_examinations_at_the_Faculty_of_Medicine_Khartoum_University

Nedeau-Cayo, R. (2013). Assessment of item-writing flaws in multiple-choice questions. *Journal for Nurses in Professional Development*, *29*(2), 52-57. doi:10.1097/NND.0b013e318286c2f1

Odukoya, J.A., Adekeye, O., Igbinoba, A.O., and Afolabi, A. (2017). Item analysis of university-wide multiple-choice objective examinations: the experience of a Nigerian private university. *European Journal of Methodology*, *52*(3), 983–997. doi: https://doi.org/10.3928/01484834-20170712-08

Oermann, M.H. & Gaberson K.B., (2014). *Evaluation and testing in nursing education* (4th ed). New York, NY: Springer Publishing Company.

Polit, D. F. & Yang, F. M. (2015). *Measurement and the measurement of change*. Philadelphia: Wolters Kluwer.

Rehman, A., Aslam, A. & Hassan, S. H. (2018). Item analysis of multiple choice questions. *Pakistan Oral and Dental Journal*, *38*(2), 291-293. Retrieved from https://www.podj.com.pk/index.php/podj/article/view/245

Reichert, T. G. (2011). *Assessing the use of high quality multiple-choice exam questions in undergraduate nursing education: Are educators making the grade?* Retrieved from Sophia, the St. Catherine University repository. https://sophia.stkate.edu/ma_nursing/15

Riazi, A. M. (2016). *The routledge encyclopedia of research methods in applied linguistics: Quantitative, qualitative, and mixed-methods research*. London: Routledge Taylor and Francis Group.

Salkind, N. J, (Eds.). (2010). *Encyclopedia of research design*. Thousand Oaks, California: Sage Publication.

Tracy, D. A., (2012). *School improvement: Revitalize your school with strategic planning.* USA: Xlibris Corporation.

Understanding item analyses. (2019). *Office of Educational Assessment*. Retrieved from http://www.washington.edu/assessment/scanning-scoring/scoring/reports/item-analysis/