© IJES 2020

PRINT: ISSN 0975-1122 ONLINE: ISSN 2456-6322

Int J Edu Sci, 31(1-3): 1-11 (2020)

DOI: 10.31901/24566322.202031.1-3.1146

Systems Based Assessment of Multiple Choice Questions (MCQs) for Quality Assurance in Testing

Haidee Pacheco[1], Joannes Paulus Hernandez[1,3], Rene Carsula[2], Richard Dennis Dayrit[1] , Hamdan Mohammad Albaqawi[1] and Eddieson Pasay-an[1]

[1]College of Nursing, University of Hail, Hail City, Saudi Arabia
[2]College of Nursing, King Saud University, Riyadh, Saudi Arabia
[3]Dialysis Nurse II, DaVita Lynbrook, New York, USA

KEYWORDS Artificial Neural Networks. Lexical Density. Multiple-Choice Examinations. Option Affinity. Systems Based. Quality Assurance

ABSTRACT This study aimed to assess the properties of MCQs using various courses in a novel holistic approach employing quantitative-comparative design. The data were collected from two university campuses in Saudi Arabia offering a Bachelor of Science in Nursing (BSN) program between 2018 and 2020. The multiple-choice questions on 'item difficulty' (34.36 ± 17.42), 'item discrimination' (0.32 ± 0.20), and 'option affinity' (0.32 ± 0.16) were reasonably good. The 'lexical density' was "Very Complex" (85.08 ± 13.37) with the Basic Adult Care Nursing final examination (91.54 ± 9.42). The 'readability index' was low (7.65 ± 3.08) in the Fundamentals of Nursing I Theory final examination (6.18 ± 3.13), but "High" in information entropy (4.23 ± 0.10) with the Fundamentals of Nursing I Theory examination (4.25 ± 0.11). Statistical differences (<.001) were noted in option affinity,' 'lexical density,' and 'readability index. Further, the multilayer perceptron (MLP) prediction model shows 'lexical density' (100%) as the most important. Findings indicate the need for quality assurance measures in the form of faculty training.

INTRODUCTION

Assessments in the learning context are most effective when they are knitted together with the curriculum goals (UNESCO 2014), however, most of which deviated from valuing individual learner differences (Schuwirth and Ash 2013). This assessment can be a valuable tool to explore the needs of the students on what to improve in the teaching and learning process. In educational institutions, assessments through examinations were most useful to evaluate the learning of the students. Although assessments are based on purpose, as to what the teachers wanted to assess. For example, the multiple-choice questions (MCQs), which assess the higher-order knowledge of the learners, however, it fails to evaluate the competencies to argue and communicate effectively (Looney 2011). In this context, the teacher will look into another way of assessing the learners when based on its objectives (for example, performance assessment, adaptive assessment).

Various studies in different course offerings of numerous educational institutions have been conducted for evaluating the quality of multiple-choice questions (MCQs) in examinations (Ramah et al. 2020). Statistical approaches that commonly utilize test item parameters such as item difficulty, item discrimination, and distractor efficiency are regarded as sample dependent or based on the ability of the test takers or students (Brown and Abdulnabi 2017). Despite the presence of various literature on writing quality MCQs (Carriveau 2016; Bowkett and Walker 2018), there has been a paucity of studies that focused on the quality of MCQs based on how they were composed by instructors. The traditional manner of measuring the distractor efficiency of MCQs that focuses on the number of non-functional distractors (Hingorjo and Jaleel 2012) takes only the attracting power of the distractors or incorrect options to be selected by a test taker as the correct answer. This approach provides a fragmented evaluation of the quality of MCQs. Such is contrary to a systems approach that would settle for no less than a consideration of all the interacting parts like the standard, the stimulus, the stem, the key or the correct option, and the distractors (Krish 2017).

This present study recognizes the need for a novel, holistic systems approach to evaluating the quality of MCQ examinations. This entails the inclusion of an assessment of the teacher's construction of the examination, specifically the lexical characteristics of the MCQs in terms of 'lexical density,' 'readability', and 'information entropy'. As such, it adds to the assessment of the student's side of the examination that utilizes the traditional statistical approach that considers 'item difficulty,' 'item discrimination' and 'option affinity' (that is, theoretically, the 'attractiveness' of MCQ answer choices among test-takers given a number and order of options regardless whether true or false statements) in lieu of 'distractor efficiency'. While English is a second language to Saudi students, it befitted the investigators to develop a novel and comprehensive quality assurance measure in assessing and comparing MCQ examinations in order to determine specific areas for improvement.

With the advent to improve the quality of learning, this study gathers new information on the progress of the learners in terms of their higher learning skills. It assesses the necessary improvement that shall take place in the teaching and learning process. This study aimed to assess the quality of multiple-choice questions (MCQ) utilizing a novel holistic approach and to determine specific areas of improvement in developing the MCQs through combined traditional assessment such as item difficulty and item discrimination with the systems approach (for example, determining 'option affinity,' in lieu of distractor effinity, and with the teacher-dependent lexical characteristics) in the form of 'lexical density,' 'readability index,' and 'information entropy'.

METHODOLOGY

Study Design

The study is a quantitative-comparative design that utilized examination papers in a multiple choice question (MCQ) items. This included the students' examination papers on Fundamentals of Nursing Theory I, Basic Adult Care Nursing Theory, Mental Health Nursing Theory, and Community Health Nursing Theory. The data were collected from two university campuses in Saudi Arabia offering a Bachelor of Science in Nursing Program between 2018 and 2020. Figure 1 is the model of the proposed system in this study.

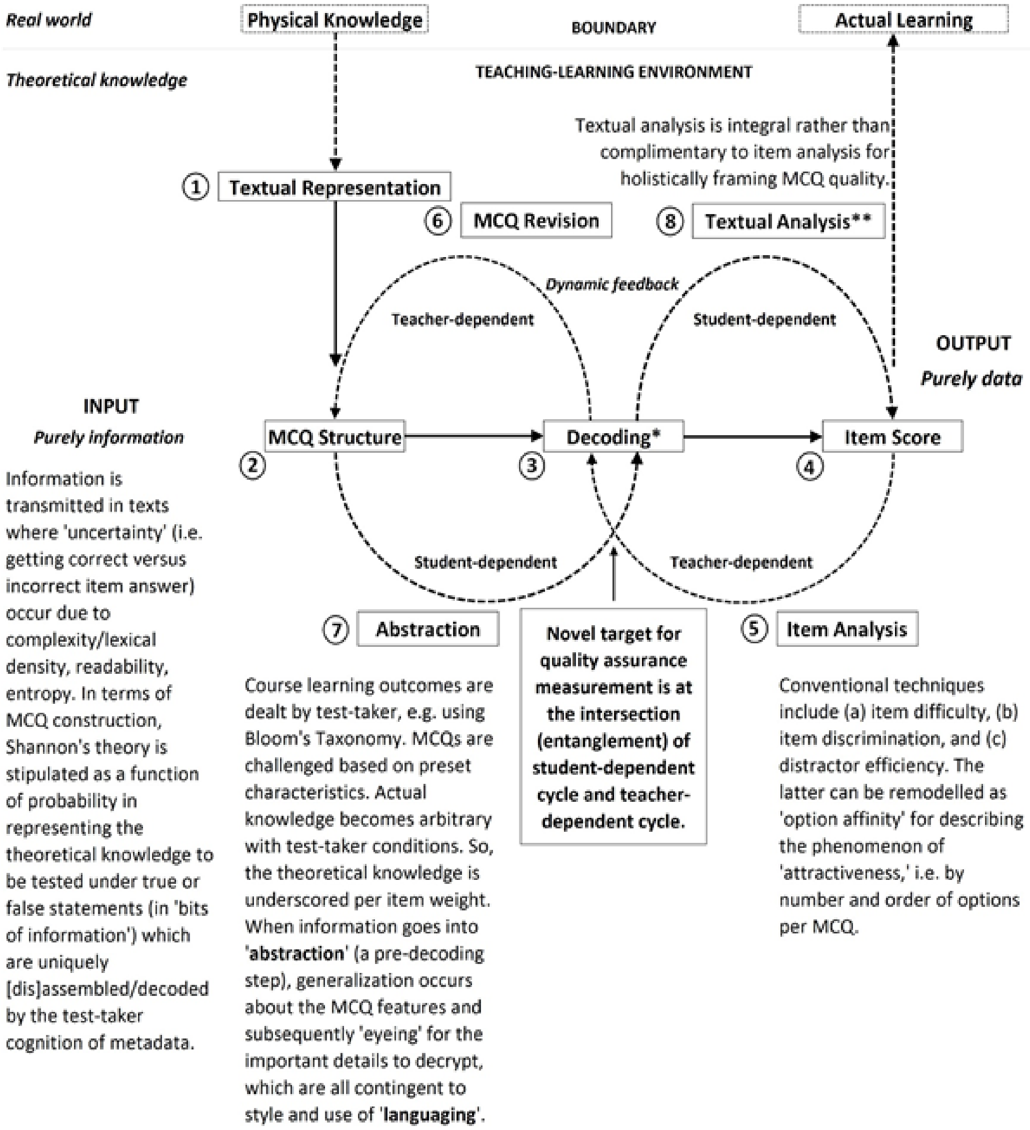
Instruments for Multiple Choice Question Assessment

An MCQ final examination paper on campus A had 36 items for the course Fundamentals of Nursing I Theory during the first semester of the academic year 2019-2020. There were 50 items in the final examination for the course Basic Adult Care Nursing Theory, and 30 items for the course Mental Health Nursing Theory. Conversely, in Campus B, a midterm examination had 40 items and the final examinations had 60 items for the course Community Health Nursing Theory.

The student-dependent characteristics of the MCQs such as 'item difficulty, 'item discrimination,' and 'option affinity' were determined through item analysis. This considers the high-performers' (HP) or top 27 percent and low-performers' (LP) or bottom 27 percent scores from each test (Ebel and Frisbie 1991) and were subsequently presented using descriptive statistics in the form of means and standard deviations/ SD using Microsoft Office 365 Excel.

The term 'item difficulty' pertains to the level of difficulty a MCQ provides a test taker in the process of selecting the correct option from a set of choices. In this study, it described the percentage of students who incorrectly answered the item and was computed by obtaining the percentage of students who got the correct answer in the MCQ and subsequently subtracting this from 100. Expressed as a percentage, 'item difficulty' is verbally interpreted as easy when less than or equal to 30 percent, "Fairly Difficult" when greater than 30 percent but less than or equal to 50 percent, "Difficult" when greater than 50 percent but less than or equal to 70 percent and "Very Difficult" if greater than 70 percent.

On the other hand, 'item discrimination' manifests how strongly a MCQ can distinguish between high- and low-performing test-takers. It was computed using the formula $[2 \times (H – L)] / (N1 + N2)$, where H stands for the number of high-performing students who selected the correct option in that item. The L being the number of low-performing students who selected the correct option in that item, and where N1 and N2 represents the number of students belonging to the high-performing and low-performing groups, respectively. Ranging from -1.00 to 1.00, 'item



Textual analysis is integral rather than complimentary to item analysis for holistically framing MCQ quality.

Information is transmitted in texts where 'uncertainty' (i.e. getting correct versus incorrect item answer) occur due to complexity/lexical density, readability, entropy. In terms of MCQ construction, Shannon's theory is stipulated as a function of probability in representing the theoretical knowledge to be tested under true or false statements (in 'bits of information') which are uniquely [dis]assembled/decoded by the test-taker cognition of metadata.

Course learning outcomes are dealt by test-taker, e.g. using Bloom's Taxonomy. MCQs are challenged based on preset characteristics. Actual knowledge becomes arbitrary with test-taker conditions. So, the theoretical knowledge is underscored per item weight. When information goes into '**abstraction**' (a pre-decoding step), generalization occurs about the MCQ features and subsequently 'eyeing' for the important details to decrypt, which are all contingent to style and use of '**languaging**'.

Novel target for quality assurance measurement is at the intersection (entanglement) of student-dependent cycle and teacher-dependent cycle.

Conventional techniques include (a) item difficulty, (b) item discrimination, and (c) distractor efficiency. The latter can be remodelled as 'option affinity' for describing the phenomenon of 'attractiveness,' i.e. by number and order of options per MCQ.

Note. *The core, recursive, linear pathway is through steps 2, 3, and 4.*

**Refers to extraction of processed quantitative information or appearing as labelled metadata.*

***Under the student-dependent cycle, this can be described as the summative [technical] decryption where critical elements are [presumed] understood among test-takers. On the other hand, quantitative measure of MCQ '**textuality**' and its methodical appreciation are lacking in educational institutions for quality assurance.*

Fig. 1. Framework for quality assurance measurement of MCQs
Source: Authors

discrimination' is verbally interpreted as "Poor" if less than or equal to 0.19, "Marginally Good" when it is from 0.20 to 0.29, "Reasonably Good" from 0.30 to 0.39 and "Very Good" when at 0.40 and above.

Statistically, 'option affinity' of each MCQ refers to the mean of the attracting power of all of the options (A, B, C, and D in four-option MCQs) to be selected by the test taker as the answer. The attracting power of each option was computed using the formula for simplified item analysis: (GD) / (HP + LP). The HP stands for the number of students belonging to the high-performing group that selected a particular option (that is, option A), LP representing the number of students belonging to the low-performing group that selected the same particular option (option A) and GD, or group difference, being equal to HP – LP. Similar to item discrimination, option affinity has a range of values from -1.00 to 1.00 and is verbally interpreted as "Poor" if less than or equal to 0.19, "Marginally Good" when it is from 0.20 to 0.29, "Reasonably Good" from 0.30 to 0.39 and "Very Good" when at 0.40 and above.

Conversely, the teacher-dependent characteristics of the MCQs such as 'complexity/lexical density' and 'readability index' were computed using the Textalyser Software available at http://textalyser.net. Equally, information entropy was computed using the online calculator for Shannon entropy (http://planetcalc.com/2476/.) Data for lexical characteristics were subsequently presented using descriptive statistics in the form of means and standard deviations through Microsoft Excel.

The 'lexical density' is a form of complexity expressed as a percentage and is measured by obtaining the ratio of lexical items (nouns, verbs, adjectives and adverbs) to the total number of words in a statement multiplied by 100 (Ure 1971). It is classified as simple if less than or equal to 30 percent, "Fairly Complex" if greater than 30 percent but less than or equal to 50 percent, "Complex" if greater than 50 percent but less than or equal to 70 percent, and "Very Complex" if greater than 70 percent.

The 'readability index' is determined from the number of words or syllables present in a sentence. The Textalyser Software used in this study generates a readability or Gunning Fog index. This is usually ranging between 0 and 20 computed through the formula: (0.4) x [(total words / total sentences) + 100 (complex words/total words)]. The complex words are classified as those having three or more syllables. The Gunning Fog index provides a grade level score such that readability index of 6 indicates that the text should be readable for sixth graders while text readable by the general public has a score of 8 and a readability score over 17. This indicates that the text should be readable for college graduates (Readability and the Gunning Fog Index, 2020). In this study, a 'readability index' greater than or equal to 13 was classified as "High," while that less than 13 were classified as "Low". Such classification is based on the Gunning-Mueller Clear Writing Institute's assignment of a Gunning Fog index of 13 to 15 for college freshmen, sophomores and juniors and a score of 11 and 12 for junior and senior high school students (Wiley 2019).

Written examinations are constituted by a matrix of coded information. In this study, it translates into how much (quantitatively represented in bits) and how likely (probability implying the uncertainty of information) MCQs can make sense among test takers. An information-theoretic view (Schmitz 2018: 1) is the approach taken here, where MCQs are assumed arbitrary (relatively interpreted) and contestable (value always queried). Under binary logic, the formulaic expression of Shannon's theory (Shannon 1948) is essential to the faithful decoding of information between true, 1 versus false, 0. The process itself gives rise to 'entropy' (Vajapeyam 2014). Operationally, it becomes the product of the average possible information that can be generated by the MCQ item and the logarithm of the length of information. Higher entropy means more complex information (Kearns 2001: 21) embeds in the MCQ structure.

Data Collection

Data gathering commenced after the approval of the University Ethical Review Committee of each campus. The study investigators accessed answer sheets for various major examinations in the selected courses from both institutions personally. Data collection protocol proceeded to observe strict confidentiality, anonymity, and appropriate safe-keeping measures.

Data Analysis

Data for the dependent variables were coded for analysis with IBM SPSS Statistics 25. Multivariate analysis of variance (MANOVA) was used to determine the significant dependent variables ('item difficulty,' 'item discrimination,' 'option affinity,' 'complexity/lexical density,' 'readability index,' and 'information entropy') across theory course examinations. On the other hand, the Games-Howell post hoc test was performed (instead of Tukey's Test) since homogeneity of variances and equality of sample sizes were not met (Shingala and Rajyaguru 2015) in finding out how theory course examinations differ to each other statistically.

Furthermore, the prediction of the importance of the dependent variables to MCQ item scores was facilitated with the multilayer perceptron/ MLP model, which is a kind of feed forward artificial neural network/ANN running on a supervised machine learning algorithm. Less technically, this is harnessing the robustness in statistical computing with artificial intelligence. This is through simulation of the binary logic processes within the brain's neural network (Abiodun et al. 2018: 4-5) for nonlinear analysis of the interacting variables between student-dependent cycle ('item difficulty,' 'item discrimination,' and 'option affinity') and teacher-dependent cycle ('complexity/lexical density,' 'readability index,' and 'information entropy'). This has been theorized in Figure 1 which is to occur as 'dynamic feedback' and the critical mechanisms to approach quality assurance at their intersection/ entanglement. Moreover, the statistical predictive power of ANNs is better than logistic regression (Dreiseitl and Ohno-Machado 2002: 356), and it provides more accuracy (Abiodun et al. 2018: 3-4). Generally, MLP architecture (https://elogeel.files.wordpress.com/2010/05/050510_1627_multilayerp1.png) has an input layer (independent variables with or without covariates), 2 or more hidden layers, and an output layer (dependent variables). Variables are processed in the equation: output = sum (weights x inputs) + bias (him0000 2018).

RESULTS

The overall MCQ 'item difficulty' was classified as "Fairly Difficult" (34.36 ± 17.42) with the Fundamentals of Nursing I Theory final examination being the most difficult (40.99 ± 18.75). The mean MCQ 'item discrimination' of the major examinations was categorized as "Reasonably Good" (0.32 ± 0.20) with the Basic Adult Care Nursing final examination having the highest discrimination (0.42 ± 0.16). The average MCQ 'option affinity' of the major examinations was likewise determined to be "Reasonably Good" (0.32 ± 0.16) with the Basic Adult Care Nursing final examination showing the best option affinity. Overall MCQ 'lexical density' was rated as "Very Complex" (85.08 ± 13.37) with the Basic Adult Care Nursing final examination displaying the most complexity (91.54 ± 9.42). The mean MCQ 'readability index' was appraised as "Low" (7.65 ± 3.08) with the Fundamentals of Nursing I Theory final examination being least readable (6.18 ± 3.13). The average MCQ item 'information entropy' was evaluated to be high (4.23 ± 0.10) with the Fundamentals of Nursing I Theory examination having the most information entropy (4.25 ± 0.11) (Table 1).

Table 2 revealed a significant difference in terms of (1) 'option affinity' (2) 'lexical density,' and (3) 'readability index' with Pillai's trace = 0.71, F (24,820) = 7.41, p = 0.001). There was a significant difference between option affinity F (4) = 27.14, p = 0.001 which has the largest effect size by partial eta squared = .34 (Prajapati et al. 2010: 1), likewise on 'lexical density' F (4) = 5.58, p = 0.001, and on readability index F (4) = 9.06, p = 0.001. On the contrary, there was no significant difference between theory courses' major examinations on 'item difficulty' F (4) = 2.77, p = 0.28, similarly on item discrimination F (4) = 4.69, p = 0.001, and on 'information entropy' F (4) = 2.96, p = 0.021.

In Table 3, the Games-Howell post hoc test indicated greater difference among other statistically significant groups in the means of the dependent variable 'option affinity' between the following examinations: Basic Adult Care Nursing; Mental Health Nursing; and Community Health Nursing (Final and Midterm).

Table 1: Characteristics of MCQs across theory course examinations

Courses	*Mean ± SD*						
	Total MCQ	*Item difficulty* (% incorrect)*	*Item discrimi-nation**	*Option affinity**	*Complexity/ lexical density (%)*	*Readability Index*	*Information Entropy*
Fundamentals of Nursing	36	40.99 ± 18.75	0.34 ± 0.15	0.31 ± 0.11	83.66 ± 16.54	6.18 ± 3.13	4.25 ± 0.11
Basic Adult Care Nursing	30	33.13 ± 15.51	0.42 ± 0.16	0.39 ± 0.09	91.54 ± 9.42	10.39 ± 2.40	4.22 ± 0.06
Mental Health Nursing	50	33.76 ± 15.76	0.33 ± 0.13	0.32 ± 0.09	88.60 ± 12.91	7.60 ± 3.01	4.24 ± 0.11
Community Health Nursing (Final)	60	30.03 ± 17.64	0.24 ± 0.25	0.31 ± 0.22	80.98 ± 11.93	7.22 ± 2.95	4.20 ± 0.10
Community Health Nursing (Midterm)	40	36.57 ± 17.90	0.32 ± 0.22	0.28 ± 0.17	83.24 ± 13.05	7.62 ± 2.60	4.23 ± 0.06
Overall	216	34.36 ± 17.42	0.32 ± 0.20	0.32 ± 0.16	85.08 ± 13.37	7.65 ± 3.08	4.23 ± 0.10
Verbal Interpretation		*Fairly Difficult*	*Reasonably Good*	*Reasonably Good*	*Very Complex*	*Low*	*High*

*Based on top and bottom 27% scores of examinees

The Games-Howell post hoc test showed the presence of a significant difference with respect to the means of the dependent variable 'complexity/lexical density' between Basic Adult Care Nursing Theory Final Examination and Community Health Nursing Theory Final Examination (Table 4).

The Games-Howell post hoc test revealed the presence of a greater difference among other statistically significant groups with reference to the means of the dependent variable 'readability index' between Fundamentals of Nursing I and Basic Adult Care Nursing and vice versa (Table 5).

Based on the MLP model, the most important among the dependent variables is 'lexical density' (100%). The partitioning of input data included 98 percent on training (batch) optimized through the scaled conjugate gradient, 1.3 percent testing, and 0.7 percent holdout. The input layer had 525 units excluding the bias unit. There were two hidden layers with 20 units in the first layer and 15 units in the second layer. There were likewise two units in the output layer. Adjusted normalized was set at 0.02 in the rescaling method for scale dependents. Hyperbolic tangent was selected as the activation function both input and output. Percent incorrect predictions for multilayer perceptron training, testing, and holdout were 28.1, 0.0, and 0.0, respectively (Table 6).

Table 6: Independent variables importance to item scores using Artificial Neural Networks (Perceptron) Model

Factors	*Importance*	*Normalized importance (%)*
Item discrimination	.170	94.9
Option affinity	.167	93.6
Complexity/lexical density	.179	**100.0**
Readability	.167	93.5
Information entropy	.151	84.6
Item difficulty	.166	92.6

Note. Sample, N=216: batch training, 146; testing, 2; and holdout, 1.

DISCUSSION

The study aimed to assess and compare the quality of MCQ examinations utilizing a novel holistic approach in order to determine specific

Table 2: Multiple analysis of variance between MCQ characteristics for theory course exams

Variables^{b}	*Sum of squares*	*df*	*Mean square*	*F*	*p*	*Partial Eta Squared*
Item Difficulty	3207.52	4	801.88	2.77	.028	.05
Item Discrimination	.67	4	.17	4.69	.001	.08
Option Affinity	3.14	4	.79	27.14	**<.001***	**.34****
Complexity/Lexical Density	3238.28	4	809.57	5.58	**<.001***	.10
Readability Index	288.33	4	72.08	9.06	**<.001***	.15
Information Entropy	.10	4	.03	2.96	.021	.05

*Statistically significant difference: *p< .001*
**Statistically largest in effect size
^{b}Normalized according to Templeton (2011)

Table 3: Games-Howell post hoc test results between groups on option affinity

	Groups	*Mean difference*	*Standard error*	*p*
Fundamentals of Nursing	Basic Adult Care Nursing	.1885	.0349	<.001
Basic Adult Care Nursing	Mental Health Nursing	**.3586***	.0379	<.001
	Community Health Nursing (Midterm)	.3133	.0396	<.001
Mental Health Nursing	Fundamentals of Nursing I	-.1885	.0349	<.001
	Basic Adult Care Nursing	**-.3586***	.0379	<.001
	Community Health Nursing (Final)	-.2223	.0339	<.001
Community Health Nursing (Final)	Basic Adult Care Nursing	**.2223***	.0339	<.001
	Community Health Nursing (Midterm)	.1770	.0359	<.001
Community Health Nursing (Midterm)	Mental Health Nursing	**-.3133***	.0396	<.001
	Community Health Nursing (Final)	-.1770	.0359	<.001

*More statistically significant difference: *p< .001*

Table 4: Games-Howell post hoc test results between groups on complexity

	Groups	*Mean difference*	*Standard error*	*p*
Basic Adult Care Nursing	Community Health Nursing (Final)	10.7797*	2.3304	<.001
Community Health Nursing (Final)	Basic Adult Care Nursing	-10.7797*	2.3304	<.001

*Statistically significant difference: *p< .001*

Table 5: Games-Howell post hoc test results between groups on readability index

	Groups	*Mean difference*	*Standard error*	*p*
Fundamentals of Nursing I	Basic Adult Care Nursing	**-4.1514***	.6788	<.001
Basic Adult Care Nursing	Fundamentals of Nursing I	**4.1514***	.6788	<.001
	Mental Health Nursing	2.6402	.5558	<.001
	Community Health Nursing (Final)	2.9051	.5497	<.001
	Community Health Nursing (Midterm)	2.5535	.5844	<.001
Mental Health Nursing	Basic Adult Care Nursing	-2.6402	.5558	<.001
Community Health Nursing (Final)	Mental Health Nursing	-2.9051	.5497	<.001
Community Health Nursing (Midterm)	Mental Health Nursing	-2.5535	.5844	<.001

*More statistically significant difference: *p< .001* areas of improvement. More specifically, it combined the traditional assessment of the student-dependent psychometric properties of MCQ examinations such as ‘item difficulty’ and ‘item discrimination’ with the systems approach of determining ‘option affinity,’ in lieu of ‘distractor efficiency,’ and with the teacher-dependent lexical characteristics of MCQ examinations in the form of ‘lexical density,’ ‘readability index’ and ‘information entropy.’ The MCQs from major examinations were assessed to be “Fairly Difficult” in the item difficulty. This indicates that the difficulty level of the MCQs in these examinations is within the acceptable range. The results were consistent with the evaluation of an anatomy examination among the undergraduate nursing students in Saudi Arabia (D’Sa and Visbal-Dionaldo 2017). Also, a pathology examination in a medical and dental college in Pakistan (Mahjabeen et al. 2017) of two preliminary examinations of undergraduate medical students (Sahoo and Singh 2017) and of five pre-university physiology examinations of medical students in India (Upadya et al. 2019).

In terms of ‘item discrimination,’ the current study revealed that the MCQs from major examinations are “Reasonably Good.” The findings demonstrated that the mean discriminating capacity of MCQs in these examinations was within the admissible categories. The results were in line with the assessment of an anatomy examination among undergraduate medical students (Patil et al. 2016a), of an epidemiology examination in a medical college (Patil et al. 2016b) in India, of a pediatric examination in Bahrain (Kheyami et al. 2018) and of a research methodology examination in Xochicalco University (Licona-Chavez et al. 2020).

The MCQs from major examinations were evaluated to be “Reasonably Good” concerning ‘option affinity’. This implies that all options in the MCQs had a good level of homogeneity in terms of attracting students to select them as the correct answer. In addition, there was a significant difference between the MCQ examinations from groups or pairs of courses. Although to the best knowledge of the investigators, no work to assess the quality of MCQs using option analysis had previously been done. For example, such a significant difference may be explained by observed variability in student behaviors while taking an MCQ examination. Specifically, this is in terms of the high percentage of students who change at least one answer in the course of completing the examination (McNulty et al. 2007) and by the observed propensity of MCQ examination test takers to guess when they are uncertain about the correct answer (Dodeen 2009). Relative to ‘lexical density,’ the results of the study showed that the MCQs from major examinations were “Very Complex”. This suggests that the MCQs may have tested students’ understanding of knowledge inherent in standardized nursing terminologies (SNTs). As a profession that focuses on the provision of care, nursing is considered as a complex process of learning with much critical thinking. As such, it utilizes SNT or language common to and readily understood by all nurses in the process of delivering care to patients (Keenan 1999). Further, there is a significant difference between the Basic Adult Care Nursing and the Community Nursing Theory final examination in terms of ‘lexical density’. The result can be explained by the specificity innate in the different SNTs (Rutherford 2008). These are mostly covered in a course such as Basic Adult Care Nursing Theory that deals with clients with both acute and chronic conditions.

Concerning the ‘readability index,’ the MCQs from major examinations were classified as low and this is in recognition that English is a second language to Saudi students. This signifies that theory instructors from both higher education institutions have made the MCQs in these examinations relatively short, despite the need to create problem-based scenario type MCQs. Moreover, the significant difference between the Basic Adult Care Nursing Theory Final Examination and Fundamentals of Nursing I Theory Examination suggests a diversity of nurse educators. This includes diversity in terms of variables that were not factored in for this study such as nationality, years of working experience in the academe, and most importantly in terms of participation in faculty development programs. This program has rendered support toward the improvement of the writing quality of MCQs in examinations (Abdulghani et al. 2015).

In terms of ‘information entropy,’ the MCQs from major examinations were assessed to be “High”. This finding supports the indication of complex information (Kearns 2001) with test construction. Specifically, MCQs bear so many cognitive interpretations since they appear verbose or too technical. This may also suggest a high degree of disorder, randomness, and imbrication/overlapping meaning. There is no literature in higher education which has underscored the implication of entropy in writing examinations and ultimately within the nursing curriculum nor has addressed any consensus for this as a remarkable metric. However, the results described above can pioneer the call for a new standard.

For the predictive component of the study, the use of the MLP model revealed that the most important among the dependent variables considered relative to MCQ examination item scores are 'complexity/lexical density'. This result underscores the need for quality assurance measures in the form of training that enhance faculty members' skills in composing good quality MCQs (Webb et al. 2015). More importantly, these training should address not only the measurement of psychometric properties of MCQs and their adherence to intended learning outcomes but should focus on the brevity of MCQ stems and options. Numerous guidelines on writing effective MCQs have emphasized the need for clear and concise stems that contain only the necessary information to present the problem (Malamed 2019). In terms of the number of options, there is evidence that suggests the superiority of four-option MCQs relative to the five-option type (Fozzard et al. 2018). Moreover, studies highlighted the advantages of three-option MCQs over the four-option type (Vegada et al. 2016; Rahma et al. 2017).

LIMITATIONS

The current study delved into quantitative measurements of psychometric properties and lexical characteristics of MCQ examinations. A determination of other parameters such as 'lexical diversity,' 'option homogeneity' and 'option plausibility' may add to the comprehensiveness of the assessment. Furthermore, a follow-up study with a qualitative component that identifies the difficulties encountered by students in answering MCQ examinations may be suggested to validate the quantitative findings. Demographical profile of the MCQ test item writers and test-takers may be considered for any significant interactions with the study variables. Thus, yielding additional rigor to test the discourses about teaching-learning outcomes with MCQs and quality measurement.

CONCLUSION

The traditional assessment of the student-dependent psychometric properties of these multiple-choice examinations such as item difficulty and item discrimination and the systems approach of determining option affinity revealed that they were all within acceptable ranges. Assessment of the teacher-dependent lexical characteristics of these multiple-choice examinations showed that they had very complex lexical density, low readability indices, and high information entropy. Differences in these multiple-choice examinations in terms of 'option affinity,' 'lexical density,' and 'readability index' were statistically significant. The lexical density is found important in the artificial neural networks (perceptron) model.

RECOMMENDATIONS

While this study indicates the need for quality assurance measures, it highly recommended that faculty members' skills in composing good quality MCQs are reinforced through training and continuing professional development. Moreover, a follow-up study can be conducted with the other courses to validate and compare the results of this study.

REFERENCES

Abdulghani H, Ahmad F, Irshad M, Khalil M, Al-Shaikh G, Syed S, Aldrees A, Alrowais N, Haque S 2015. Faculty development programs improve the quality of Multiple Choice Questions items' writing. *Scientific Reports*, 5: 9556. https://doi.org/10.1038/srep09556

Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H 2018. State-of-the-art in artificial neural network applications: A survey. *Heliyon*, 4(11): e00938. https://doi.org/10.1016/j.heliyon. 2018.e 00938

Bowkett E, Walker S 2018. Tips for Educators: How to Write Multiple-Choice Questions. From<https://www.adinstruments.com/blog/tips-educators howwrite-multiple-choice-questions> (Retrieved on 1 March 2020).

Brown G, Abdulnabi H 2017.Evaluating the quality of higher-education instructorconstructed multiple-choice tests: impact on student grades. *Frontiers in Education*, 2(24): 1-12. https://doi.org/10.3389/feduc. 2017. 00024.

Carriveau R 2016. *Connecting the Dots: Developing Student Learning Outcomes and Outcomes-Based Assessment.* 2nd Edition. Sterling, VA, United States: Stylus Publishing.

Dreiseitl S, Ohno-Machado L 2002. Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, 35(5-6): 352-359. https://doi.org/10.1016/ S1532-0464 (03)00034-0.

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

Dodeen H 2009. Test-related characteristics of UAEU students: Test-anxiety, test-taking skills, guessing, attitudes toward tests, and cheating. *Journal of Faculty of Education*, 26: 31-66.

Ebel R, Frisbie D 1991. *Essentials of Educational Measurement.* 5th Edition. New Jersey, USA: Prentice Hall, Englewood Cliffs.

Fozzard N, Pearson A, du Toit E, Naug H, Wen W, Peak I 2018. Analysis of MCQ and distractor use in a large first year Health Faculty Foundation Program: assessing the effects of changing from five to four options. *BMC Med Educ*, 18: 252.

Hingorjo M, Jaleel F 2012. Analysis of One-Best MCQs: The Difficulty Index, Discrimination Index and Distractor Efficiency. The Journal of Pakistan Medical Association, 62(2): 142-147. From <https://jpma. org.pk/article>details/3255?article_id=3255> (Retrieved on 24 March 2020).

Kearns J 2001. A Mechanism For Richer Representation of Videos For Children: Calibrating Calculated Entropy To Perceived Entropy. Doctoral Dissertation. USA: University of North Texas. From <https://www. research gate.net/publication/34995295_A_mechanism_for_ Richer_reprsentation_of_videos_ for_children_ microform_ calibrating_calculated_ entropy_ to_ peceived_ entropy> (Retrieved on 13 June 2020).

Keenan G 1999. Use of standardized nursing language will make nursing visible. *Michigan Nurse*, 72(2): 12-13.

Kheyami D, Jaradat A, Al-Shibani T, Ali FA 2018. Item analysis of multiple choice questions at the department of paediatrics, Arabian Gulf University, Manama, Bahrain. *Sultan Qaboos University Medical Journal*, 18(1): e68. https://doi.org/10.18295/squmj. 2018. 18.01.011.

Krish N 2017. The Anatomy of Assessment: 5 Elements of a Quality Multiple Choice Question. From <https://www. teachthought.com/pedagogy/5-elements quality-multiple-choice-question/> (Retrieved on 18 July 2020).

Licona-Chavez A, Montiel Boehringer P, Velazques-Liano L 2020. Quality assessment of multiple choice test through psychometric properties. *Med Ed Publish*, 9(1): 91.https://doi.org/10.15694/mep. 2020. 000091.1.

Looney J 2011. Alignment in Complex Education Systems: Achieving Balance and Coherence. *OECD Education Working Papers*, No. 64, OECD Publishing, Paris. https://doi.org/10.1787/5kg3vg5lx8r8-en.

Mahjabeen W, Alam S, Hassan U, Zafar T, Butt R, Konain S, Rizvi M 2017. Difficulty Index, Discrimination Index and Distractor Efficiency In Multiple Choice Questions. Annals of PIMS, 13(4): 310-315. From <https://apims.net/index.php/apims/article/download/ 9/10> (Retrieved on April 3, 2020).

Malamed C 2019. 10 Rules for Writing Multiple Choice Questions. From <http://theelearningcoach.com/ elearning_design/rules-for-multiple-choice questions> (Retrieved on 23 May 2020).

McNulty J, Sonntag B, Sinacore J 2007. Test-taking behaviors on a multiple-choice exam are associated with performance on the exam and with learning style. *Medical Science Educator*, 17(1): 52-57.

Patil P, Dhobale M, Mudiraj N 2016a. Item analysis of MCQs – myths and realities when assessing them as an assessment tool for medical students. *International Journal of Current Research and Review*, 8(13): 12-16.

Patil R, Palve S, Vell K, Boratne A 2016b. Evaluation of multiple choice questions by item analysis in a medical college in Pondicherry, India. *International Journal of Community Medicine and Public Health*, 3(6): 1612-1616. http://dx.doi.org/10.18203/2394-6040.ijcmph20161638.

Prajapati B, Dunne M, Armstrong R 2010. Sample Size Estimation and Statistical Power Analyses. Optometry Today, 16(7): 10-18. From <http://www. floppybunny.org/robin/web/virtualclassroom/stats/ basics/articles/gpower/Gpower_tutorial_ Prajapati_ 2010-.pdf.> (Retrieved on 12 April 2020).

Rahma N, Shamad M, Idris M, Elfakey W, Salih K 2017. Comparison in the quality of distracters in three and four options type of multiple choice questions. *Advances in Medical Education and Practice*, 8: 287-291. https://doi.org/10.2147/AMEP. S128318.

Ramah N, Yuzrizal A, Syukri M 2020. Analysis of Multiple Choice Questions of Physics Final Examination in Senior High School. Journal of Physics: Conf. Series 1460 (2020) 0121431460, The 1st Annual International Conference on Mathematics, Science and Technology Education. Kota Banda Aceh, Indonesia, 14-15 September 2018. From <https://doi.org/10.1088/17426596/ 1460/1/012143> (Retrieved on 2 April 2020).

Rutherford M 2008. Standardized nursing language: What does it mean for nursing practice? *OJIN: The Online Journal of Issues in Nursing*, 13(1). https://doi.org/10.3912/OJIN.Vol13No01PPT05.

Sahoo D, Singh R 2017. Item and distracter analysis of multiple choice questions from a preliminary examination of undergraduate medical students. *International Journal of Research in Medical Sciences*, 5(12): 5351-5355.http://dx.doi.org/10.18203/2320-6012.

Shannon C 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3): 379-423. doi:10.1002/j.1538-7305.1948.tb01338x.

Schmitz GJ 2018. Entropy and geometric objects. *Entropy*, 20(6): 453. https://doi.org/10.3390/e20060453.

Schuwirth L, Ash J 2013. Assessing tomorrow's learners: in competency-based education only a radically different holistic method of assessment will work. Sixthings we could forget. *Medical Teacher*, 35(7): 555-559.http://dx.doi.org/10.3109/0142159X.2013. 787 140.

Shingala MC, Rajyaguru A 2015. Comparison of Post Hoc Tests for Unequal Variance. International Journal of New Technologies in Science and Engineering, 2(5): 22.33. From <https://www.ijntse.com/upload/1447070311130.pdf> (Retrieved on 24 January 2020).

United Nations Educational, Scientific and Cultural Organisation (UNESCO) 2014.Teaching and learning: Achieving quality for all. Education for All GlobalMonitoring Report. From<https://unesdocunesco-org/ark:/48223/pf0000225660> (Retrieved on 23 February 2020).

Upadyah A, Maheria P, Patel J 2019. Analysis of one-best MCQs in five pre-university physiology examinations. *International Journal of Physiology*, 7(4): 10-15.https://doi.org/10.5958/2320-608X.2019. 00129.X.

Ure J 1971. Lexical density and register differentiation. *Applications of Linguistics*, 443-452.

Vajapeyam S 2014. Understanding Shannon's Entropy Metric for Information. ArXiv preprint arXiv:1405. 2061. From <https://arxiv.org/pdf/1405. 2061> (Retrieved on 10 May 2020).

Vegada B, Shukla A, Khilnani A, Charan J, Desai C 2016. Comparison between three option, four option and five option multiple choice question tests for qualityparameters: A randomized study. *Indian Journal of Pharmacology*, 48(5): 571-575. https://doi.org/10.4103/0253-7613.190757.

Webb EM, Phuong JS, Naeger DM 2015. Does educator training or experienc eaffect the quality of multiple-choice questions? *Academic Radiology*, 22(10): 1317-1322. https://doi.org/10.1016/j.acra.2015. 06. 012.

Wiley A 2019. Measure Clarity with Readability Formulas. From <https://www.linkedin.com/pulse/measure-clarity-readability-formulas-ann-wylie> (Retrieved on 18 May 2020).

Paper received for publication in September, 2020
Paper accepted for publication in September, 2020