Crimson Publishers
Wings to the Research

Review Article

Open-Ended Testing Stays Relevant for Critical Knowledge Assessment

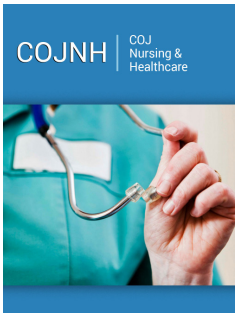
Noémi Fuller[1], Mónika Ferenczy[2], Szilvia Szunomár[3], Orsolya Máté[4], Zsuzsanna Germán[5], Annamária Pakai[6], Sziládiné Katalin Dr Fusz[7], Miklós Zrínyi[8]*, András Oláh[9]

[1]Director, Faculty of Health, University of Pécs, Hungary

[2,3,4,5,6,7]Assistant Professor, Faculty of Health, University of Pécs, Hungary

[9]Dean, Faculty of Health, University of Pécs, Hungary

ISSN: 2577-2007



***Corresponding author:** Miklós Zrínyi, Assistant Professor, Faculty of Health, University of Pécs, Hungary

Submission: October 18, 2019

Published: :November 07, 2019

Volume 5 - Issue 4

How to cite this article: András O, Noémi F, Mónika F, Szilvia S, Orsolya M, Zsuzsanna G, Annamária P, S, Katalin, Miklós Z. Open-Ended Testing Stays Relevant for Critical Knowledge Assessment. COJ Nurse Healthcare.5(4). COJNH.000619.2019. DOI: 10.31031/COJNH.2019.05.000619.

Abstract

Aim: To evaluate the difference/bias rate between open-ended (OE) vs multiple-choice questions (MCQs) for a critical nursing skill.

Method: Three hundred and seventy-six nurses responded from 3 nursing schools to a 20-item MCQs and OE instrument. Questions concerned nursing knowledge of maintaining clear airways and interventions related to trachea suction. Subjects first responded to the OE instrument followed by MCQs. Both tests were paper based. Statistical analyses included paired t-tests and one-way ANOVA.

Result: Outcomes showed significant differences between OE and MCQs, MCQs were scored higher. On MCQs, Licensed Vocational Nurse (LVN) and Associate Degree in Nursing (ADN) group scores did not differ. As for OE, both ADNs and BSNs did significantly better than LVNs. BSNs increased their lead markedly over both groups in OE.

Conclusion: MCQs overestimated knowledge levels when respondents' knowledge base was less confident or professional qualification was lower. When knowledge base was solid, the difference between OE and MCQs disappeared.

Highlights

A. Multiple-choice test scores can be biased upwards compared to open-ended tests

B. BSN nurses outperformed LVN and ADN nurses on open-ended tests

C. Nurses with more/longer experience score equal on both test types

D. Open-ended tests are still advantageous to assess critical skills

Keywords: Open-ended; Multiple-choice; Test; Test bias; Critical skill

Introduction

Valid, reliable and standardized knowledge assessment has always been a challenge for educators and continue to pose new dilemmas. By the adoption of computerized systems and technological innovations, knowledge assessment has gradually shifted towards electronic exams and choice-based response alternatives, in line with expectations of millennial medical students whose preference is for speed and efficacy [1]. Reid & Colleagues [2] also documented that nurses in general thought that computer-based testing was more user friendly and would rather take a multiple-choice question exam. As argued by Morrision & Walsh [3], measuring nursing students' capacity for critical thinking remains a constant challenge in which carefully crafted multiple-choice questions (MCQs) may provide a reliable platform.

The debate, however, whether to use MCQs or open-ended (OE) techniques is ongoing. Friewald et al. [4] pointed out that easy administration, scoring and evaluation make MCQs a tempting option and are widely utilized by faculty. Research, however, confirmed that OE and MCQs measure different characteristics of the thought process [5]. So called 'higher order' MCQs have been questioned whether or not they are able to test clinical reasoning skills of medical students [4]. MCQs were also criticized to expose students to false statements which later they recall as being true [6].

Tweed & Colleagues [7] raised concerns that many students performing well on MCQs have incorrect responses to items that should be considered hazardous and that collecting a sufficient number of correct answers should not offset incorrect responses. Electronic exams, in general, also tend to be biased upwards in terms of scoring compared to paper based examinations [8]. OE methods, on the contrary, require students of higher level of thinking and, instead of simple knowledge recall, knowledge construction. Why OE is less popular has do with faculty time needed to find and score the correct text [9].

When OE was applied to mathematical testing, which requires higher order of thinking, and was compared against MCQ on number of errors and misconceptions, OE was found to produce more favorable results [10]. A similar test was repeated years later to find no significant difference between OE and MCQs [11]. When OE was applied to test whether nurses were able to calculate correct drug dosage (a critical skill that may lead to fatal outcome if performed incorrect), authors found that dosage dilution was answered correctly by 87% of respondents [12]. They also reported a success rate of 69.6% for finding the safe dose as well as 79% for identifying the right ratio/proportion. Had this been performed as an MCQ test, the proportion of correct responses could have been higher. Recall the argument by Tweed et al. [7] who argued that correct answers should not offset incorrect responses for critical skills assessment. Therefore, the aim of our research was to assess a critical care nursing skill (respiratory pathway management) on a large nursing cohort to evaluate the difference between OE vs MCQs.

Hypotheses

A. Subjects will score greater on average on the MCQ instrument.

B. Nurses with BSN degree will score higher than other qualifications on both instruments.

Methods

This investigation utilized a cross-sectional, non-experimental, survey-based research design, respondents being their own control in the study (pairwise comparison). Subjects were recruited from three different locations, the main campus of the University in X [placeholder for institution], and from its two satellite schools X [placeholder for institution]. Participants were all nurses attending continuing education credit classes. They had been approached on the day of attending the class and were asked to participate in a research project that involved knowledge testing. A total of 500 nurses had been contacted. Sampling was convenient in that whoever on the day of testing volunteered to sit for the test had been included. There were no specific exclusion criteria decided for this research. Data was collected over a period of 6 months in Spring and Summer classes of 2017. Participants were asked to sit for two sets of tests after finishing their continuing education classes. The first test administered was an open-ended (OE) instrument followed by a second test taken with a multiple-choice instrument containing identical questions to the OE assessment tool. Each test lasted for 45 minutes.

Participants were seated in a large auditorium leaving enough space between to avoid copying from each other. There was only a small break allowed between two tests, most test takers were asked to stay in place in order to minimize discussion of results. To reduce stress induced testing bias, the principal investigator clarified participants before taking the test that they were part of a research and outcomes of their answers did not influence their continuing education credits by any means. The research was submitted for local IRB approval. Participation was voluntary and anonymized. There was no specific external funding received to support study implementation.

Instruments

The actual research instrument was a paper-based, 20-item, multiple-choice assessment tool developed by a panel of six advanced practice nurses on the topic of securing open airways and nurse management of the trachea. All six nurses worked in critical care units and were considered expert nurses on the topic. As the purpose of the research was not instrument development but testing, we focused on establishing content validity of the instrument by asking another expert panel of eight nurses, coming both from academia and practice, to agree on the set of questions. For each item wording an agreement rate of 90% was accepted to make sure items were clear and represented the topic tested. A final number of 20 items were included in order to allow test takers a sufficient amount of time to complete the task. Sample items on this instrument included "What catheter lubrification do you use before introducing it into the airways to apply trachea suction?

A. Tap water

B. Distilled water

C. Sterile sodium chloride solution

D. the catheter must not be lubricated" or "How do you choose the size of the catheter for the patient before trachea suction?

a. The diameter of the catheter should not be more than 50% of the inner diameter of the artificial airways

b. The diameter of the catheter should be more than 50% of the inner diameter of the artificial airways

c. Catheter should be selected by the weight of the patient

d. Catheter should be selected by its color

e. The diameter of the catheter should not be more than one-third of the inner diameter of the artificial airways

f. The diameter of the catheter should not be more than 20% of the inner diameter of the artificial airways". Open-ended items included the same questions asking the respondents to describe the answer by their own words.

Each correct answer on the multiple-choice instrument was assigned one (1) point, incorrect answers zero (0) points. Similar coding was used for the OE instrument. When the answer included the correct wording/description, one point was assigned, if the

wording was incorrect zero point was recorded. Tests were checked by the same panel of six advanced practice nurses who developed the instrument. For the OE instrument, in case of ambiguity, two panel members were asked to agree on the final score of the unclear item. Final score for both instruments were calculated by adding all items with a score of 1. The possible range of scores was between 0 and 40. Besides these instruments, a demographic survey was also distributed to record age, gender and highest qualification of nurses along with a few other indicators. The final, English version of the instrument is available from the authors.

Statistical analyses included descriptive statistics of sample characteristics and scores achieved on main tests. One sample Kolmogorov-Smirnov test was used to check the normality assumption of our data. Paired sample t-test as well as Wilcoxon test were used to establish the difference between OE and multiple-choice test results. One-way ANOVA with Bonferroni post-hoc test was used to check whether outcomes of both OE and multiple-choice tests were different across various nurse qualifications. A priori sample size calculations by (level of significance set at 5%, statistical power at 0.85 [15%], and medium effect size [0.25]) showed that a total of 180 subjects had to be recruited to ensure adequate statistical power [13]. All analyses were done by SPSS Windows version 23.0. There was no policy developed for the replacement of missing data. Answers and spaces left blank were considered lack of knowledge in this research.

Result

Our final sample included 376 nurses who decided to take both tests. Average age of the sample was 41.5 (SD 9.85) years, nurses had been working in healthcare for an average of 12.1 (SD 10.19) years. Of the total sample, 22.6% represented RNs (BSN), 43% Associate Degree Nurses (ADN), and the rest of our sample included nurses with a Licensed Vocational Nurse (LVN) degree. As for when nurses had been awarded their degrees, 87% of our sample graduated before 2010. This sample was comprised of 4% male and 96% female nurses.

Table 1: Descriptive Statistics, Test Scores.

	N	Minimum	Maximum	Mean	Std. Deviation
Full Sample					
Multiple-choice	376	2	25	13,38	4,31
Open-ended	376	0	34	6,48	7,24
More Frequently Involved in Respiratory					
Multiple_choice	58	8	23	16,48	3,12
Open-ended	58	2	34	14,48	9,64
Less Frequently Involved in Respiratory					
Multiple-choice	318	2	25	12,81	4,26
Open-ended	318	0	29	5,02	5,62

Table 1 shows descriptive data for both instruments for the full sample and a subsample of nurses who identified themselves as being more or less frequently involved in the clinical management of respiratory pathways. One sample Kolmogorov-Smirnov tests for OE and multiple-choice scales showed significance levels (p) being<0.001, that is, scores were not normally distributed.

Results indicate that the full sample as well as subsamples were skewed to the left, that is, average scores reported on both instruments were below the scale midpoint (below 20 points). This means that nurses on average displayed inferior knowledge of the topic tested. Table 1 however also shows that those with more frequent involvement in respiratory management scored much closer on both instruments than those with less frequent practice.

To test whether total scores achieved on both instruments differed, we employed paired sample t-tests as well as Wilcoxon tests (due to non-normal distribution). Both tests yielded identical differences between the two instruments, therefore we report here results of paired t-tests (Table 2).

Table 2: Paired t-tests.

		Mean	N	Std. Deviation	Std. Error Mean	Sig.
Full Sample	Multiple-choice	13,38	376	4,31	0,22	< 0.001
	Open-ended	6,48	376	7,24	0,37	
Less Frequently Involved in Respiratory	Multiple-choice	12,81	318	4,26	0,23	< 0.001
	Open-ended	5,02	318	5,62	0,31	
More Frequently Involved in Respiratory	Multiple-choice	164,828	58	3,12	0,41	0,118
	Open-ended	144,828	58	9,64	1,26	

Outcomes of the analyses showed significant differences between OE and multiple-choice total scores, favoring multiple-choice, for the full sample as well as for the less frequently involved subsample. However, there was no statistical difference between OE and multiple-choice test scores for nurses who reported more frequent involvement in respiratory management.

Finally, we looked at whether there were underlying differences across nurse qualifications in terms of both OE and multiple-choice total scores. Due to non-normal data distribution, we ran both one-way ANOVA and Kruskal-Wallis tests. Since both tests yielded significant results, we report here the ANOVA outcome. Differences for both models were significant (F_{MCQs} = 10.04; p < 0.001 and F_{OE} = 50.49; p < 0.001). Evidently, for the multiple-choice instrument, LVN and ADN groups did not differ on total test scores whereas BSN nurses achieved significantly better scores than both groups (Table 3). As for the OE instrument, however, both ADNs and BSNs did significantly better compared to LVNs, but BSN nurses increased their lead markedly over both groups when compared to multiple-choice testing.

Table 3: ANOVA Post hoc Multiple comparisons.

Dependent Variable	(I) 3 group	(J) 3 group	Mean Difference (I-J)	Std. Error	Sig.
Multiple-choice	LVN	ADN	-0,163	0,5	1,000
		BSN	-2,506*	0,608	0,000
	ADN	LVN	0,163	0,5	1,000
		BSN	-2,3	0,584	0,000
	BSN	LVN	2,506*	0,608	0,000
		AND	2,343*	0,584	0,000
Open-ended	LVN	ADN	-2,081*	0,763	0,020
		BSN	-9,135*	0,927	0,000
	ADN	LVN	2,081*	0,763	0,020
		BSN	-7,054*	0,891	0,000
	BSN	LVN	9,135*	0,927	0,000
		ADN	7,054*	0,891	0,000

*Mean difference is significant at the 0.05 level.

Discussion

This research aimed to investigate whether there were any differences in knowledge assessment of nurses when open-ended vs multiple-choice instruments had been used. In general, the sample achieved an average total score on both scales lower (13.38 multiple choice and 6.48 open-ended) than was expected. This may have been due to a few reasons. First, test takers were probably less motivated in their test performance (since they knew they had been part of an experiment) to complete as many correct answers as possible as opposed to real life testing.

However, missing data (that would have lowered average scores) on all instruments was very low, suggesting that nurses completed the task as if taking a real-life test. Another potential explanation may be that the majority of respondents (81%) earned their nursing degree before 2010 (Hence relevant knowledge had not been not refreshed) and did not carry out everyday nursing tasks involving respiratory management skills. Had this been a real-life pass/fail test, 23.3% of our sample would have passed on the OE, 45.2% on the MCQ test. Our success rate was much lower than that of Ozyazicioglu et al. [12].

While our test was also generic, however, was more complex than reported by Ozyazicioglu et al. [12], Hence less easy to answer to. Note that when examining maximum scores achieved (Table 1), OE responses outperformed MCQ on all accounts.

As for the difference between OE and multiple-choice test scores, our results conflicted with Stepankova & Emanovsky [11] who reported no differences between the two test modalities in their study, and with Birenbaum & Tatsuoka [10], who found OE more favorable. Our results showed MCQs return greater average scores compared to OE, however, OE maximum scores were higher than MCQs. Importantly enough, when respondents' knowledge base was strongly rooted, we saw no significant difference between the two test modalities. We also confirm an upward scoring bias for MCQs as argued by Washburn et al. [8] despite that both tests were administered paper-based.

We know that more advanced nurse qualifications are related to higher order of knowledge which was clearly supported by the ANOVA analysis. BSN nurses successively outperformed ADNs and LVN nurses both in OE and MCQs. However, the distance between BSNs and the other two cadres increased markedly when OE scores had been juxtaposed. We also demonstrated the differentiating power of OE vs MCQs for ADNs and LVNs as well. While these two groups achieved statistically identical results for MCQs, ADNs emerged ahead of LVNs in the OE testing phase. Therefore, both null hypotheses were rejected in this research.

Authors of this paper, therefore, conclude that MCQs, in general, have a tendency to overrate knowledge levels when respondents' knowledge base is less confident or professional qualification levels are closer. When knowledge base was strong, the difference

between OE and MCQs disappeared. However, we likewise observed that OE responses outperformed MCQs when maximum achievable scores were considered. Why that happened is up to speculation at this point. We recommend future research to investigate the cause of such difference. Authors, therefore, argue that both assessment methods seem valid to distinguish various knowledge and professional levels, and should be used perhaps interchangeably, or best in combination within the same test, when assessing critical knowledge and skills.

Limitation

Authors acknowledge that knowledge tested in this paper was part of standard nursing education, however, subjects did not frequently refresh and utilize this particular set of knowledge in practice, possibly skewing results. Authors are also aware that exposing subjects to the information of being investigated may have changed their behaviors to testing. However, this was a condition to observe in the ethical approval.

References

1. Pettit RK, McCoy L, Kinney M (2017) What millennial medical students say about flipped learning. Adv Med Educ Pract 20(8): 487-497.
2. Reid J, Robinson D, Lewis C (2016) Student response system versus computer-based testing for undertaking multiple choice question assessment in undergraduate nursing education. Pediatrics and Neonatal Nursing 3(1): 10-14.
3. Morrison S, Free KW (2001) Writing multiple-choice test items that promote and measure critical thinking. J Nurs Educ 40(1): 17-24.
4. Freiwald T, Salimi M, Khaljani E, Harendza S (2014) Pattern recognition as a concept for multiple-choice questions in a national licensing exam. BMC Med Educ 14(14): 232.
5. Ozuru Y, Briner S, Kurby CA, McNamara DS (2013) Comparing comprehension measured by multiple-choice and open-ended questions. Canadian Journal of Experimental Psychology 67(3): 215-227.
6. Xu X, Kauer S, Tupy S (2016) Multiple-choice questions: Tips for optimizing assessment in-seat and online. Scholarship of Teaching and Learning in Psychology 2(2): 147-158.
7. Tweed MJ, Stein S, Wilkinson TJ, Purdie G, Smith J (2017) Certainty and safe consequence responses provide additional information from multiple choice question assessments. BMC Med Educ 17(1): 106.
8. Washburn S, Herman J, Stewart R (2017) Evaluation of performance and perceptions of electronic vs. paper multiple-choice exams. Adv Physiol Educ 41(4): 548-555.
9. Melovitz VCA, DeFouw DO, Holland BK, Vasan NS (2018) Analysis of testing with multiple choice versus open-ended questions: Outcome-based observations in an anatomy course. Anat Sci Educ 11(3): 254-261.
10. Birenbaum M, Tatsuoka KK (1987) Open-ended versus multiple-choice response formats--it does make a difference for diagnostic purposes. Applied Psychological Measurement 11: 385-395.
11. Štěpánková B, Emanovský P (2011) On open-ended and closed-ended questions in didactic tests of mathematics. Problems of Education in the 21st Century 28: 114-122.
12. Özyazıcıoğlu N, Aydın Aİ, Sürenler S, Çinar HG, Yılmaz D, et al. (2018) Evaluation of students' knowledge about paediatric dosage calculations. Nurse Educ Pract 28: 34-39.
13. Power G (2019) Statistical Power Analyses for Windows and Mac.