Medical Education Online

ISSN: 1087-2981 (Online) Journal homepage: www.tandfonline.com/journals/zmeo20

Generating multiple-choice questions using reverse engineering techniques

Sujin Shin, Jaehwa Choi, Eunmin Hong, Minjae Lee & Miji Lee

To cite this article: Sujin Shin, Jaehwa Choi, Eunmin Hong, Minjae Lee & Miji Lee (2026) Generating multiple-choice questions using reverse engineering techniques, Medical Education Online, 31:1, 2611632, DOI: 10.1080/10872981.2025.2611632

To link to this article: https://doi.org/10.1080/10872981.2025.2611632

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 09 Jan 2026.

Submit your article to this journal

Article views: 1318

View related articles

View Crossmark data

Citing articles: 1 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=zmeo20

MEDICAL EDUCATION ONLINE
2026, VOL. 31, NO. 1, 2611632
https://doi.org/10.1080/10872981.2025.2611632

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Generating multiple-choice questions using reverse engineering techniques

Sujin Shin[a], Jaehwa Choi[b], Eunmin Hong[c], Minjae Lee[a] and Miji Lee[d]

[a]College of Nursing, Ewha Womans University, Seoul, Republic of Korea; [b]Department of Educational Leadership, Graduate School of Education and Human Development, The George Washington University, Washington, DC, USA; [c]College of Nursing, Wonkwang University, Iksan, Republic of Korea; [d]Department of Nursing, Catholic University of Pusan, Busan, Republic of Korea

ABSTRACT

The study is aimed at exploring the potential of using ChatGPT-4 to develop multiple-choice questions that adhere to item development principles through the application of reverse engineering techniques in nursing education. To determine whether ChatGPT-4 could generate multiple-choice questions regarding medical-surgical nursing, the researchers first evaluated ChatGPT-4's level of nursing knowledge. The researchers then developed 20 prompts using reverse engineering techniques; these were subsequently entered into ChatGPT-4, resulting in the generation of 60 multiple-choice questions. All 60 questions were reviewed by five nursing education experts to evaluate compliance with item development principles. ChatGPT-4 demonstrated a high accuracy rate (98.00%) when answering multiple-choice questions about medical-surgical nursing education. AI-generated questions primarily adhered to the principles of item development, with an average compliance score of 7.90/9.00. The least adhered to among the nine-question development principles was 'The question stem should use only essential words or phrases.' Given that the AI-generated questions may serve as preliminary drafts for use in nursing education following expert review, ChatGPT-4 can function as a valuable tool in the design of educational materials. When utilizing ChatGPT-4, it is beneficial for nursing educators to apply reverse engineering techniques that enhance question quality through repeated comparison, as well as to analyze the prompts and corresponding AI-generated questions. Furthermore, including structured outlines and representative sample multiple-choice questions using Markdown syntax in the prompt can enhance the quality of AI-generated questions.

ARTICLE HISTORY
Received 2 July 2025
Revised 17 November 2025
Accepted 22 December 2025

KEYWORDS
Educational measurement; multiple-choice questions; generative artificial intelligence; educational technology; nursing education

Introduction

As artificial intelligence (AI) technology is increasingly applied across diverse fields, numerous studies have evaluated its validity in applications such as automated grading [1] and feedback generation [2]. Research has increasingly focused on developing and validating multiple-choice questions (MCQs) [3–5]. MCQs are widely used in nurse licensure examinations as a form of knowledge-based assessment [6–8]. AI enables time-efficient MCQ generation by facilitating collaboration among item writers [9] and producing MCQs with clear language, a single correct answer, and high-quality distractors [10]. One study found that AI-generated questions were comparable in quality to those written by human experts [11], suggesting that AI- generated MCQ development is a viable approach. Given AI's demonstrated utility in MCQ design, its potential applications in nursing education warrant further exploration [12]. Although such studies are limited, research has begun to examine the feasibility of using AI to generate MCQs for nursing assessment [13], reflecting growing interest in integrating AI into MCQs development.

The development of MCQs requires a sound understanding of the subject matter to construct a compelling MCQs that can accurately assess an examinee's knowledge [14]. Previous studies have shown that ChatGPT can solve MCQs, including the nurse licensure examination, which indicates the presence of nursing knowledge in ChatGPT [15,16]. This suggests the possibility of developing MCQs using ChatGPT that can be used in

CONTACT Miji Lee miji_lee@naver.com Department of Nursing, Catholic University of Pusan, Busan, Republic of Korea

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

nursing education. ChatGPT is an AI model designed to interact conversationally, follow instructions in a prompt, and provide detailed responses [17]. Owing to its foundation in large language models (LLMs), it can respond with text that resembles human language [18]. ChatGPT is a transformer-based model pretrained to predict the next token in text and can perform complex language tasks with enhanced accuracy [19].

OpenAI [20] emphasises the importance of prompt engineering, specifically crafting effective prompts when using ChatGPT. It recommends six strategies for better outcomes: writing clear instructions, providing reference text, breaking complex tasks into smaller components, allowing adequate processing time, integrating external tools, and systematically testing modifications. Recent studies highlight the significance of prompt engineering when utilising LLMs such as ChatGPT in educational settings. In physics education, prompt engineering demonstrated how carefully crafted prompts can enhance ChatGPT's ability to deliver accurate, context-specific responses [21]. Furthermore, a literature review on prompt engineering in developing MCQs in medical education indicated that AI-generated questions have comparable quality to human-written questions and significantly reduce MCQs development time [3]. Despite the importance of prompt engineering, this topic remains underexplored in nursing education [22].

Reverse engineering, a subset of prompt engineering, supports the creation of test items that align closely with reference formats [23]. It involves tailoring prompts to specific outputs, improving LLMs' adaptability through context-specific modelling and prompt refinement [24]. Reverse engineering involves the decomposition and reintegration of particulars by recovering their meaning and comprehensive relationships to gain a deeper understanding of the subject [25,26]. Haynie et al [27]. defined reverse engineering as 'the process of creating a design or blueprint by analysing a final product or system, often via identification of system components and their interrelationships, and creating representations of that product or system in an enhanced form or at a higher level of abstraction.' Reverse engineering is a method for constructing test specification components based on actual tests [28]. Traditional item-writing procedures generally employ a forward design approach, which begins with selecting a specific content domain and then constructing the stem, options, and answer. In contrast, generic reverse engineering utilises a backward design perspective. This approach starts from the 'model answers or expected learning outcomes' rather than the content domain. When applying reverse engineering using AI, such as in the AREA (assessment reverse engineering by AI) framework [25], this concept is automated. An automated item-generation system can input a 'justified exemplar and item schema into the prompt, thereby creating an item-generating template' [23]. This method strengthens the alignment between learning objectives and assessment items, which enhances the content validity of the assessment [29]. Therefore, the aim is to explore the potential of using ChatGPT-4 to develop MCQs that adhere to the principles of item development and can be used in nursing education by applying these reverse engineering techniques.

Materials and methods

Selection of large language models

This study utilised ChatGPT with the GPT-4 model. At the time of our research, ChatGPT was the most rapidly adopted web application in history and the most prominent LLMs in both public use and academic research on MCQ generation [30,31].

Given the nascent stage of this research area, our objective was not a comparative analysis across multiple LLMs but rather an in-depth, foundational investigation into the capabilities of a single, state-of-the-art model. By focusing on GPT-4, we aimed to establish a benchmark for the feasibility and educational potential of LLMs in nursing education assessment. We acknowledge the rapid evolution in this field. Future work should build upon these findings by conducting comparative evaluations with other contemporary and next-generation models (e.g., GPT-5) to provide a broader understanding of their relative strengths and weaknesses.

Study design

This study employed a descriptive design to assess ChatGPT-4's level of knowledge regarding medical-surgical nursing, and ability to generate MCQs that comply with item development principles using prompts developed through reverse engineering techniques.

Study purpose

This study explored the potential use of AI-generated questions for nursing education. The specific research questions were as follows:

1. Does ChatGPT-4 possess adequate nursing knowledge to generate MCQs?
2. Can AI-generated questions follow the principles of item development?

Setting and sample

Using a purposive sampling technique, five nursing education experts in South Korea were recruited for the study. The experts were professors from nursing colleges with >15 years of educational experience and expertise in developing or reviewing items for the Korean Nursing Licensure Examination.

Instrument

The principles of item development were evaluated using the item checklist included in the Item Development Guide by the Korea Health Personnel Licensing Examination Institute [32]. The checklist comprised 17 items divided into three categories: content (7 items), stems (4 items), and options (6 items).

Prior research by Little et al [33]. has shown that indicators may not consistently contribute to validity across contexts. They found that restricting analyses to relevant subscales can still preserve validity, provided these subscales accurately represent the construct. This study examined whether AI-generated questions in accordance with established item development principles, focusing specifically on the format of stems and options. As a result, only the subscales related to stems and options were used, and the content subscale was omitted.

Similarly, Boateng et al [34]. reported that a scale remains valid if the retained items sufficiently represent the construct, even after some items are removed. In this study, the sample MCQs on the nervous system used a four-option format. Following research team discussions, the remaining items in the option subscale were still confirmed to represent the construct after one item was removed. Therefore, the item 'Are all five options related to the stem?' was excluded as inapplicable.

Data collection/procedure

This study followed a structured process to evaluate ChatGPT-4's capacity to generate MCQs that adhere to item development principles in the field of medical-surgical nursing, with data collection conducted from April to July 2024. To assess ChatGPT-4's nursing knowledge, the researchers entered each of 20 sample MCQs into ChatGPT-4 five times, yielding a total of 100 responses. Subsequently, the researchers developed 20 prompts using reverse engineering techniques to guide the generation of MCQs in medical-surgical nursing. Each prompt was entered into ChatGPT-4 three times, resulting in a total of 60 MCQs. Finally, five nursing education experts independently reviewed the 60 MCQs to assess their adherence to item development principles.

Prompt

Writing efficient prompts necessitates the use of basic and intuitive language to diminish ambiguity and enhance the machine's capability for accurate interpretation [35]. Following this methodology, prompts employing reverse engineering were formulated. Assessment Reverse Engineering by AI (AREA), a method of applying reverse engineering, includes detailed methodologies such as AI-based Lesson Plan for Harmonic Assessment (ALPHA), Base-Item Engineering Transformation by AI (BETA), and Specification and Instruction Guided Modelling by AI (SIGMA) and Ontology Model Emulation by generative AI (OMEGA) [25].

This study employed three methods within the AREA framework: BETA, SIGMA, and OMEGA. OMEGA simulates an ontology-based model for curricular content. In this study, content from the

nervous and respiratory systems was organised into a structured outline file, which served as a reference for ChatGPT-4 during the creation of MCQs. SIGMA relies on specification and instruction, utilising nine item development principles and a standard format of MCQs. This ensured that all AI-generated questions adhered to established guidelines. BETA deconstructs exemplar items (i.e., few-shots) and rebuilds them by varying features. For this, ChatGPT-4 received 20 sample MCQs. These samples helped the LLMs analyse item format, structure, and principles, allowing it to reproduce and expand upon established patterns when creating new MCQs.

Prompts were systematically organised into two main sections: an introduction, and MCQ development. The introduction explains the role of ChatGPT-4 and the tasks that it should perform. In the MCQ development stage, ChatGPT-4 was instructed to generate MCQs by referring to an outline file that structurally organises the content of the nervous and respiratory systems in medical-surgical nursing, the format of MCQs, and the sample MCQs provided by the researchers. Although ChatGPT-4's word input is limited, it can be effectively managed through 'structured outlining.' This method efficiently condenses complex information and facilitates the clear communication of key points. The ability to distil intricate content aids in the rapid grasping of core concepts, thereby providing a solid foundation for text generation and revision [36].

The prompts were written in Markdown, a lightweight markup language selected for its dual utility in being both highly readable for humans and explicitly structured for machines. Its plain-text syntax uses simple symbols (e.g., # for headings, * for list items) to define a clear document hierarchy and represent elements such as emphasis, lists, and block structures [37,38].

This characteristic is pivotal to our methodology. For humans (i.e., subject matter experts), the syntax is intuitive, allowing for the clear and consistent outlining of item development principles and sample MCQs. For the machine (ChatGPT-4), this same syntax provides an unambiguous, parsable structure that functions as a lightweight knowledge model. This aligns with the core principle of structuring knowledge representation well for both human and machine comprehension, i.e., Ontology Model-centred Generation (OMG) [39]. Therefore, Markdown was not merely a choice of convenience but a deliberate decision to ensure our instructions were transmitted to the AI with high fidelity and structural integrity.

The 20 Sample MCQs presented by the researcher in the prompt were developed by a nursing college professor who majored in medical-surgical nursing and had >15 years of nursing education experience (Appendix 1). The 20 sample MCQs comprised 10 questions on the nervous system and 10 questions on the respiratory system for medical-surgical nursing. Three nursing education researchers, one with a master's degree and two with doctoral degrees, all specialising in medical-surgical nursing and with teaching experience in nursing colleges, reviewed the 20 sample MCQs to assess content overlap and compliance with the item development principles. The review confirmed that all 20 sample MCQs were unique in content and adhered to the item development principles, leading to their selection as the study's sample MCQs.

Finally, the item development principles that ChatGPT-4 should follow when generating MCQs were presented. Appendix 2 shows the prompts developed by the researchers to generate MCQs using ChatGPT-4 based on the first example MCQ.

To ensure the standardisation of the MCQs generated by ChatGPT-4, we specified the MCQ development criteria within each prompt. Each prompt was designed to produce a single MCQ, specifying a 4- or 5-option multiple-choice format and listing the nine item development principles to ensure compliance. Additionally, to maintain consistency across all generated MCQs, the prompt structure and content were standardised, guiding ChatGPT-4 to follow a uniform approach in MCQs generation throughout the study.

Data analysis

Data analyses were performed using Microsoft Excel (Microsoft Corporation, Redmond, WA, USA) and R version 4.5.1 (R Foundation for Statistical Computing, Vienna, Austria).

The analysis included an evaluation of ChatGPT-4's knowledge level in medical-surgical nursing by calculating the average accuracy rate across 100 responses. The extent to which the 60 AI-generated questions conformed to item development principles was assessed using a set of predefined indicators: the

(a) average number of item development principles satisfied across the three MCQs generated for each sample MCQ; (b) average number of experts who judged each principle as compliant across the three MCQs generated for each sample MCQ; and (c) ranking of the nine item development principles based on their compliance rates across all 60 MCQs.

Inter-rater reliability among five experts was evaluated using Fleiss' kappa, Percent agreement, and Gwet's AC1. Although kappa statistics are commonly used to assess inter-rater agreement, Fleiss' kappa is sensitive to category prevalence and may yield low values, even when there is high observed agreement [40]. To address this limitation, it is advisable to supplement Fleiss' kappa with additional indices such as Percent agreement and Gwet's AC1, which provide a more comprehensive assessment of inter-rater reliability [40,41].

Results

ChatGPT-4's level of nursing knowledge

Table 1 presents the accuracy results of the 20 sample MCQs to confirm the nursing knowledge level of ChatGPT-4. After ChatGPT-4 answered the 20 sample MCQs five times each, all responses except two were correct, and the average accuracy rate was 98.00%.

Evaluations of nursing education experts regarding whether ChatGPT-4 can generate MCQs according to compliance with the item development principles.

As for inter-rater reliability, Fleiss' kappa was 0.16, indicating slight agreement. The percent agreement was 62.00%, showing that more than half of the items achieved complete consensus among all experts. Gwet's AC1 was 0.77 (95% CI = 0.74–0.81, $p < .001$), corresponding to substantial agreement.

Table 2 and Appendix 3 present the evaluation by the five nursing education experts, who assessed compliance with the nine item development principles for the 60 AI-generated questions, three MCQs per sample, resulting in 60 MCQs. The average results of the five nursing education experts' reviews, which assessed how many of the principles of item development the three MCQs per sample complied with, showed that the MCQs complied with a minimum of 5.90, a maximum of 8.90, and an average of 7.90 principles per MCQ.

Table 3 and Appendix 4 present the evaluations by five nursing education experts, detailing which of the nine item development principles the 60 AI-generated questions adhered to; the average result of the evaluation was 4.50.

Table 1. Accuracy rate for sample MCQs answered by ChatGPT-4.

	Answer					
Sample MCQ No.	1	2	3	4	5	T (%)
1	T	T	T	T	T	100.00
2	T	T	T	T	T	100.00
3	T	T	T	T	T	100.00
4	T	T	T	T	T	100.00
5	T	T	T	T	T	100.00
6	T	T	T	T	T	100.00
7	T	T	T	T	T	100.00
8	T	T	T	T	T	100.00
9	T	T	T	T	T	100.00
10	T	T	T	T	T	100.00
11	T	T	T	T	T	100.00
12	T	T	T	T	T	100.00
13	T	T	T	T	T	100.00
14	T	T	T	T	T	100.00
15	T	T	T	T	T	100.00
16	T	T	T	T	T	100.00
17	F	T	T	T	F	60.00
18	T	T	T	T	T	100.00
19	T	T	T	T	T	100.00
20	T	T	T	T	T	100.00
Average						98.00
Standard deviation						8.94
Range						60.00-100.00
Distribution of correct answers						100% ($n = 19$); 60% ($n = 1$, Sample MCQ No. 17)

F, false; MCQs, multiple-choice questions; T, true.

Table 2. Evaluation by the five nursing education experts regarding compliance with the nine principles of item development for MCQs generated by ChatGPT-4 ($N = 5$).

MCQ No.		
Sample MCQs	MCQs generated by ChatGPT-4	Average number of principles of item development that complied[a]
1	1–3	7.00
2	4–6	7.00
3	7–9	8.30
4	10–12	8.10
5	13–15	8.70
6	16–18	8.10
7	19–21	5.90
8	22–24	7.00
9	25–27	7.20
10	28–30	8.10
11	31–33	8.60
12	34–36	8.90
13	37–39	8.60
14	40–42	8.10
15	43–45	8.50
16	46–48	8.30
17	49–51	7.40
18	52–54	7.30
19	55–57	7.90
20	58–60	8.40
Average		7.90
Standard deviation		0.77
Range		5.90–8.90

MCQ, multiple-choice question.
[a]Scores range from 0–9.

Table 4 summarises expert evaluations of 60 ChatGPT-generated MCQs based on nine item development principles categorised as stem-related (P1–P4) and option-related (P5–P9). The overall mean score of 4.46 ± 0.32 indicates substantial adherence to established item development standards. Regarding stem-related principles, ChatGPT performed well in negation (P4, 4.80) and clarity (P3, 4.60), indicating that stems were clear and avoided unnecessary negativity. Independence (P1, 4.20) and conciseness (P2, 3.80) scored lower, suggesting that some stems required reading the options for full comprehension or contained redundant wording. For option-related criteria, high performance was observed in homogeneity (P5, 4.70), non-redundancy (P8, 4.70), and mutual independence (P6, 4.60), indicating that ChatGPT produced balanced and plausible options with consistent internal logic. However, ordering (P9, 4.40) and uniqueness (P7, 4.30) were slightly weaker, reflecting occasional inconsistencies in logical sequencing or overlapping options.

Integrating the reverse engineering framework provides crucial academic and practical context for the results in Table 4. The strong performance in clarity (P3) and distractor quality (P5, P6, P8) is likely a direct benefit of this backward design approach, which uses a 'justified exemplar and item schema' to constrain the AI and enhance content validity. This process is central to the AREA (assessment reverse engineering by AI [25]) framework, which emphasises that input facets - such as providing 'few-shot' exemplars or relying on 'zero-shot' reasoning - are visible and controllable. Because these inputs can be seen, their impact on the output can be observed, allowing the developer to iteratively modify them for better results. Conversely, the weaker scores in conciseness (P2) and independence (P1) may also be an artefact of this method, where the AI included extra context to strictly adhere to the exemplar. While these findings confirm the AI-generated MCQs met essential quality criteria, a direct comparison against manually developed items was outside this study's scope, though it remains a critical next step to highlight which aspects of this reverse engineering approach perform better or worse than traditional methods.

Discussion

This study aimed to evaluate the possibility of using AI to develop MCQs in nursing education. When developing MCQs, sufficiently understanding the subject matter is essential to construct a compelling item that can accurately assess an examinee's knowledge [14]. In this study, we confirmed the level of nursing knowledge in ChatGPT-4 by presenting sample MCQs corresponding to the nervous and respiratory system domains of medical-surgical nursing, for which MCQs were to be developed. This study assessed

Table 3. Evaluation by the five nursing education experts regarding the compliance with each principle of item development ($N = 5$).

MCQ No.		Average number of nursing education experts who evaluated that the principle of item development was compliant[a]									
Sample MCQ	MCQs generated by ChatGPT-4	P1	P2	P3	P4	P5	P6	P7	P8	P9	Average
1	1–3	2.00	3.30	4.70	3.00	5.00	4.70	4.70	5.00	4.30	4.10
2	4–6	3.70	3.00	5.00	3.30	3.00	4.70	3.70	4.70	4.00	3.90
3	7–9	5.00	3.00	5.00	5.00	4.70	4.70	5.00	5.00	4.00	4.60
4	10–12	5.00	3.30	5.00	5.00	5.00	4.70	5.00	4.30	3.30	4.50
5	13–15	5.00	4.00	4.70	5.00	4.70	5.00	5.00	5.00	5.00	4.80
6	16–18	5.00	3.30	5.00	5.00	5.00	5.00	4.70	3.70	5.00	4.60
7	19–21	3.00	2.70	4.00	5.00	4.00	2.00	4.00	3.00	2.00	3.30
8	22–24	3.00	5.00	4.30	5.00	5.00	4.30	3.70	4.30	4.30	4.30
9	25–27	3.70	4.70	4.70	5.00	4.70	5.00	2.70	5.00	4.30	4.40
10	28–30	4.00	2.70	5.00	5.00	5.00	5.00	5.00	5.00	4.30	4.60
11	31–33	5.00	4.70	5.00	4.70	5.00	5.00	5.00	5.00	5.00	4.90
12	34–36	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	4.70	5.00
13	37–39	4.00	4.30	5.00	5.00	5.00	5.00	4.30	4.30	5.00	4.70
14	40–42	5.00	5.00	4.00	5.00	4.00	4.00	4.00	5.00	4.00	4.40
15	43–45	4.70	4.00	5.00	5.00	5.00	5.00	4.00	5.00	5.00	4.70
16	46–48	5.00	4.00	4.30	4.30	4.70	4.70	4.70	5.00	4.00	4.50
17	49–51	4.00	3.30	3.00	5.00	5.00	4.00	3.30	5.00	4.70	4.10
18	52–54	4.30	2.30	4.30	5.00	4.70	4.30	3.30	4.00	4.30	4.10
19	55–57	3.30	3.70	4.00	5.00	5.00	5.00	4.00	5.00	4.70	4.40
20	58–60	3.30	5.00	4.00	5.00	5.00	5.00	4.70	5.00	5.00	4.70
Average		4.20	3.80	4.60	4.80	4.70	4.60	4.30	4.70	4.40	4.50
Standard deviation		0.87	0.86	0.53	0.56	0.50	0.68	0.68	0.55	0.71	0.38
Range		2.00-5.00	2.30-5.00	3.00-5.00	3.00-5.00	3.00-5.00	2.00-5.00	2.70-5.00	3.00-5.00	2.00-5.00	3.30-5.00

MCQ, multiple-choice question; P, principle of item development; P1, the question stem should allow for the correct answer to be inferred without looking at the options; P2, the question stem should use only essential words or phrases; P3, the content being asked in the question stem should be easily understood by the examinee; P4, the question stem should not be unnecessarily pessimistic; P5: the options should be homogeneous; P6, the options should not give hints to each other; P7, there should only be one correct answer and multiple answers should not be possible; P8, if words are repeated in the options, they must be necessary; P9, the order of the options should be regular or logical.

[a]Scores range from 0–5.

Table 4. Evaluation by the five nursing education experts regarding the compliance with each principle of item development organised by rank.

	Principle of item development	
Score[a]	No. (score)	Contents
Above average (≥4.50)	P4 (4.80)	The question stem should not be unnecessarily pessimistic.
	P5 (4.70)	The options should be homogeneous.
	P8 (4.70)	If words are repeated in the options, they must be necessary.
	P3 (4.60)	The content being asked in the question stem should be easily understood by the examinee.
	P6 (4.60)	The options should not give hints to each other.
Below average (<4.50)	P9 (4.40)	The order of the options should be regular or logical.
	P7 (4.30)	There should only be one correct answer, and multiple answers should not be possible.
	P1 (4.20)	The question stem should allow for the correct answer to be inferred without looking at the options.
	P2 (3.80)	The question stem should use only essential words or phrases.
Average	4.46	
Standard deviation	0.32	
Range	3.80-4.80	

P, principle of item development; P1, the question stem should allow for the correct answer to be inferred without looking at the options; P2, the question stem should use only essential words or phrases; P3, the content being asked in the question stem should be easily understood by the examinee; P4, the question stem should not be unnecessarily pessimistic; P5: the options should be homogeneous; P6, the options should not give hints to each other; P7, there should only be one correct answer and multiple answers should not be possible; P8, if words are repeated in the options, they must be necessary; P9, the order of the options should be regular or logical.

[a]Scores range from 0–5; number of nursing education experts who evaluated the principles of item development as compliant.

ChatGPT-4's level of nursing knowledge by answering 20 sample MCQs. Although a small number of MCQs were used, ChatGPT-4 achieved a high accuracy rate (98.00%) after answering the 20 sample MCQs five times. This indicates a promising starting point for the use of AI in the development of MCQs. However, interpretation of this result requires consideration of different types of questions. For instance, Zhu et al [42]. found that questions answered incorrectly in multiple-choice format were correctly answered in open-ended form, with accuracy rates ranging from 92.00% to 96.00%. Therefore, while the high accuracy observed in this study suggests the potential of ChatGPT-4 for nursing education

assessment, these results should be interpreted cautiously and not generalised without further validation using larger and more diverse data sets.

The 98.00% accuracy achieved by ChatGPT-4 is noteworthy, but this result must be interpreted within the context of our methodology and the items' intended purpose. Firstly, the sample MCQs used as exemplars in this study were primarily designed for learning and knowledge reinforcement, rather than for high-stakes summative assessment. Consequently, they targeted a moderate level of difficulty, which likely contributed to the model's high performance. Secondly, this study focused on generating new items by varying 'surface' features (e.g., phrasing, clinical context) from well-established, high-quality exemplars, which likely contributed to the high quality of the final output. The exemplars themselves were primarily designed for learning and knowledge reinforcement, thus targeting a moderate level of difficulty.

This highlights a critical area for future investigation: manipulating 'deep' features that directly influence an item's difficulty and cognitive complexity. It remains an open question whether the high quality observed in this study can be maintained when the generation process moves beyond modifying surface features. Future research should explore methods to deliberately control item difficulty, an emerging field that involves emulating difficulty during generation [39] or using fine-tuned models to predict it [43]. Developing such capabilities is essential for creating dynamic, adaptive learning systems that can tailor challenges to individual student needs [44]. Furthermore, another promising avenue is to employ the LLMs as a self-reviewer, where the AI generates an item and then critically evaluates AI-generated items. Such a process would require a rigorous meta-evaluation [39] to validate the LLM's reviewing capabilities, ensuring its judgments align with those of human experts.

The accuracy confirmed in this study is different from those reported by Su et al [45]. and Taira et al. [16], who confirmed the accuracy of ChatGPT-3.5 on the nurse licensure examination; the overall accuracies were 80.75% and 69.80%, respectively. A reason for this discrepancy may be the difference in the ChatGPT model's performance; ChatGPT-4 surpasses and outperforms ChatGPT-3.5 in its advanced reasoning capabilities [46]. Another reason may be the format of the test questions. The format of the sample MCQs used in this study was of a single type, 'single-answer multiple-choice question'; four or five options were presented, of which only one was the correct or most appropriate answer. However, the questions used in the study of Su et al [45]. and Taira et al [16]. were presented together with the complex format of 'single-answer and multiple-response MCQs.' Their studies reported a higher probability of incorrect responses to single-answer and multiple-response MCQs. A third reason may be the differences between the test subjects. In the study by Su et al. [45], the accuracy was high at 88.70% for general medicine and 80.00% for medical-surgical nursing; the other test subjects scored relatively low at 63.00–70.00%. The present study is limited in that it does not cover the entire medical-surgical nursing field; only the nervous and respiratory systems were evaluated. Additionally, all 20 sample MCQs used in this study were related to medical-surgical nursing. Similarly, in the study by Taira et al. [16], there was a difference in accuracy among the test subjects. Therefore, when selecting the AI model, nursing educators are required to carefully consider the model's performance, the format of the MCQs, and the subject areas to be assessed.

In this study, we examined whether the AI-generated questions complied with the item development principles. Evaluation by nursing education experts revealed that the AI-generated questions complied with an average of 7.90 of the nine item development principles. Among the 60 MCQs, all but three satisfied at least seven principles. The three items generated from Sample MCQ 7 (MCQs No. 19–21) complied with only 5.90 principles, marking a clear deviation from the others. Sample MCQ 7 focused on assessing the Glasgow Coma Scale (GCS) score in a patient with increased intracranial pressure. Although the generated MCQs had stems and options similar to the original, they mainly failed to meet P2 (use of essential words or phrases), P6 (avoidance of inter-option cueing), and P9 (logical ordering of options). For instance, the stems included unnecessary expressions (e.g., 'over the course of treatment' and 'would most accurately indicate the greatest improvement'). Additionally, the options also lacked a logical order, such as ascending or descending total GCS scores. Notably, P2 exhibited the lowest compliance among the nine item development principles across all MCQs. Such issues may mislead students or reduce the discriminative power of the assessment. Notably, P2 had the lowest overall compliance rate across all MCQs. These findings are consistent with Klang et al. [9], who reported that 15% of generated items required correction due to inaccuracies or outdated terminology. Similarly, Kıyak and Emekli [3] highlighted that while ChatGPT accelerates MCQ development, expert review is essential to ensure educational quality. Although AI-generated questions require a revision process, this method is more efficient than the existing method, wherein experts participate in the entire item development process, from item writing to

item review. Henceforth, we propose a new process in which ChatGPT-4 generates draughts of MCQs, and experts revise them.

Polverini and Gregorcic [21] reported on how select prompting approaches can improve the performance of ChatGPT-4. Similarly, we improved the quality of ChatGPT-4 responses in this study by utilising prompt engineering strategies, such as specifying the domain, specifying how to act, and providing comments on the context. However, a single prompt was used; chain-of-thought (CoT) prompting was not used. CoT prompting can enhance the reasoning capabilities of LLMs to solve complex reasoning tasks [47]. In 2025, models such as OpenAI's o3 [47] and DeepSeek's R1 [48] were developed, further emphasising the growing importance of enhancing the reasoning capabilities of LLMs. It would be meaningful for future research to utilise LLMs specialised in reasoning capabilities, or to apply CoT prompting when developing MCQs.

To develop effective test items, one must understand the assessment objectives, be familiar with the types of items, and faithfully follow the principles of item development [49]. In this study, when writing a prompt for MCQ generation, we specifically presented content that can be used as a reference when generating MCQs, including the format of MCQ and principles that must be followed when developing MCQs. Additionally, the prompt was presented with an outline file that structurally organised the content of the nervous and respiratory systems in medical-surgical nursing and the sample MCQs. It is essential to design prompts by providing specific exam styles, reference text, personas, or clear and detailed explanations of what is needed for an item when generating MCQs [3]. The 20 sample MCQs presented in this study were developed by a nursing college professor specialising in adult nursing with > 15 years of experience in nursing education, and reviewed by three researchers; thus, the quality of the items can be considered high. These prompts, which included the outline file and sample MCQs, were developed using a Markdown format by the researchers. Markdown is a markup language with a plain-text formatting syntax that instructs machines to display, describe, or manipulate a document's structure and format for readability, simplicity, and flexibility [50]. Markdown uses simple symbols, such as creating a heading by adding a #symbol in front of a heading to control the shape and structure of the text, or creating a list using numbers followed by a dot and space [51]. In certain tasks using ChatGPT-4, the use of Markdown formatting elicited a higher model performance; additionally, it was found that Markdown can more effectively facilitate CoT reasoning [52]. Furthermore, the researchers employed a reverse engineering approach by repeatedly inputting the prompts, reviewing the outputs, and refining the prompts based on a comparison between the inputs and generated results, thereby contributing to the systematic enhancement of prompt quality. The researchers provided a structured outline and high-quality example MCQs using prompts that incorporate Markdown through reverse engineering; thus, it can be inferred that the MCQs were generated in compliance with the principles of item development.

Although the reverse engineering methodology in this study demonstrated that structured prompts systematically improve the quality of AI-generated questions, further validation with actual learners remains necessary. Recent studies have begun to address this gap by evaluating the educational effectiveness of AI-generated questions. Emekli and Karahan [53] found that AI-generated questions showed difficulty and discrimination indices similar to educator-developed MCQs. In contrast, Chauhan et al [54]. found that many MCQs failed to meet acceptable standards of difficulty, discrimination, and distractor effectiveness, thereby underscoring the continued need for expert review. These findings suggest that future empirical research in nursing education may consider conducting item analyses and learner-centred validation on MCQs that have undergone expert review. Such a comprehensive evaluation will allow for a more complete understanding of the educational value of AI-generated MCQs in nursing education.

Limitations

This study has several limitations that also point to important avenues for future research. First, this study utilised 20 MCQs as few-shot exemplars to guide the AI in generating new MCQs. A crucial methodological question is the optimal number of such examples. The ideal size of this 'few-shot' sample is not yet established and is difficult to generalise, as it likely depends on the breadth and complexity of the target knowledge domain. For instance, a broad subject may require more exemplars than a narrowly focused one. Future work should systematically investigate this variable, perhaps through simulation studies or other a priori processes, to determine how the quantity and quality of exemplars affect the generated

MCQs. Second, our exclusive focus on medical-surgical nursing restricts the generalisability of these findings. While this domain provided a robust testing ground, further research is needed to validate this methodology across other nursing specialties and health disciplines to establish its broader utility in educational assessment.

Conclusions

This study highlights the potential of ChatGPT-4 as a supportive tool for generating MCQs in nursing education. Reverse engineering techniques can facilitate the development of AI-generated questions by supplying model answers. ChatGPT-4 demonstrated the ability to generate distractors aligned with item-writing guidelines by leveraging contextual prompts derived from established nursing education frameworks. However, it should be used as a complementary resource, with all AI-generated questions requiring expert review.

Regarding the implications for educators and assessment designers, our findings suggest that generative AI, when guided by a structured framework like reverse engineering, can be a powerful partner in curriculum development rather than a simple replacement for human expertise. To integrate these tools responsibly, educators should adopt a 'human-in-the-loop' model, as emphasised in this study. The AI's role is to generate a high-quality 'first draft' of MCQs, but the educator's expert review remains the critical step for quality control. This expert review is essential to validate the items' alignment with learning objectives, check for 'plausibility of distractors,' and ensure the 'clarity of stem' meets pedagogical standards. Furthermore, this hybrid approach addresses key ethical considerations. By having an expert review and curate all AI-generated content before deployment, educators ensure fairness, mitigate potential AI biases, and maintain final accountability for the validity of their assessments.

As a significant future extension, this study's methodology could be formalised within the CAFA frameworks [55] and platform [56] by creating a specialised, generative agent. For instance, the reverse engineering technique aligns directly with the REPAIR (Reverse Engineering Passage-based Assessment by Intelligent Reconstructor) or OMEGA (Ontology Model Emulated by Generative AI) frameworks. This specialised 'Item Generator' agent could then be integrated into a larger multi-agent system using the MACRO (Multi-Agent Context Routing and Orchestration) protocol. In such a workflow, the generator agent would create the MCQs and pass their Data ID (DID) to a second agent, such as an AGILE (Adaptive, Generative, and Iterative Learning Engine) agent, which would then administer the test adaptively. To gain a more comprehensive understanding of the applicability of AI in nursing education, future research needs to broaden the range of nursing domains assessed and compare the performance of different AI models.

Acknowledgements

None.

Disclosure statement

The authors report there are no competing interests to declare.

Funding

This article is funded by the National Research Foundation of Korea, grant id-2023R1A2C2006838.

ORCID

Sujin Shin 0000-0001-7981-2893
Jaehwa Choi 0000-0003-4225-9582
Eunmin Hong 0000-0003-2917-8067
Minjae Lee 0000-0001-7615-9593
Miji Lee 0000-0001-7946-7006

Data availability statement

The authors confirm that the data supporting the findings of this study are available within the article or its appendices.

Ethical approval

This study was approved by the Institutional Review Board of the Ewha Womans University (Approval no. ewha-202403-0022-01) on March 17, 2024.

Informed consent from participants

Written informed consent to participate in the research and publish the data was obtained from the five nursing education experts.

References

[1] Jukiewicz M. The future of grading programming assignments in education: the role of ChatGPT in automating the assessment and feedback process. Think Skills Creat. 2024;52:101522. doi: 10.1016/j.tsc.2024.101522
[2] Jacobsen LJ, Weber KE. The promises and pitfalls of ChatGPT as a feedback provider in higher education: an exploratory study of prompt engineering and the quality of AI-driven feedback. 2023. [cited 2024 Aug 6]. Available from: https://osf.io/preprints/osf/cr257
[3] Kıyak YS, Emekli E. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. Postgrad Med J. 2024;100:858–865. doi: 10.1093/postmj/qgae065
[4] Mistry NP, Saeed H, Rafique S, et al. Large language models as tools to generate radiology board-style multiple-choice questions. Acad Radiol. 2024;31:3872–3878. doi: 10.1016/j.acra.2024.06.046
[5] Özbay Y. Evaluation of ChatGPT as a multiple-choice question generator in dental traumatology. Med Rec. 2024;6:235–238. doi: 10.37990/medr.1446396
[6] Korea Health Personnel Licensing Examination Institute. Nursing exam information. 2024. [cited 2024 Aug 6]. Available from: https://www.kuksiwon.or.kr/subcnt/c_2005/7/view.do?seq=7
[7] National Council of State Boards of Nursing. NCLEX-RN® test plan. 2023. [cited 2024 Aug 6] Available from: https://www.ncsbn.org/public-files/2023_RN_Test%20Plan_English_FINAL.pdf
[8] Nursing and Midwifery Council. Nursing and Midwifery Council test of competence 2021: Part 1. 2024.[cited 2024 Dec 6]. Available from: https://www.nmc.org.uk/globalassets/sitedocuments/registration/toc-materials-2023/nursing-handbooks/test-of-competence-2021-cbt-information-booklet-for-nurses.pdf
[9] Klang E, Portugez S, Gross R, et al. Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. BMC Med Educ. 2023;23:772. doi: 10.1186/s12909-023-04752-w
[10] Doughty J, Wan Z, Bompelli A, et al. A comparative study of AI-generated (GPT-4) and human-crafted MCQs in programming education In: Proceedings of the 26th Australasian Computing Education Conference. 2024. pp. 114–123. doi: 10.1145/3636243.3636256
[11] Ali FA, Sharif SY, Ata M, et al. The ChatGPT develops multiple choice questions (MCQs) for postgraduate specialty assessment: a reality or a myth? Pak J Neurol Surg. 2024;28:142–149. doi: 10.36552/pjns.v28i1.963
[12] Castonguay A, Farthing P, Davies S, et al. Revolutionizing nursing education through AI integration: a reflection on the disruptive impact of ChatGPT. Nurse Educ Today. 2023;129:105916. doi: 10.1016/j.nedt.2023.105916
[13] Lin Y, Li C, Li H. Comparing AI-generated and traditional textbook multiple-choice questions in nursing education: a prompt engineering-based Delphi study. *Med Teach*. 2025; Epub ahead of print. doi: 10.1080/0142159X.2025.2533401
[14] Zaidi NLB, Grob KL, Monrad SM, et al. Pushing critical thinking skills with multiple-choice questions: does Bloom's taxonomy work? Acad Med. 2018;93:856–859. doi: 10.1097/ACM.0000000000002087
[15] Huang H. Performance of ChatGPT on registered nurse license exam in Taiwan: a descriptive study. Healthcare. 2023;11:2855. doi: 10.3390/healthcare11212855
[16] Taira K, Itaya T, Hanada A. Performance of the large language model ChatGPT on the National Nurse Examinations in Japan: evaluation study. JMIR Nurs. 2023;6:e47305. doi: 10.2196/47305
[17] OpenAI. Introducing ChatGPT [Internet]. 2022.[cited 2024 Aug 7]. Available from: https://openai.com/index/chatgpt/
[18] Sætra HS. Generative AI: here to stay, but for good? Technol Soc. 2023;75:102372. doi: 10.1016/j.techsoc.2023.102372
[19] Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. arXiv [Preprint]. 2023. arXiv:2303.08774.
[20] OpenAI. Prompt engineering [Internet]. 2023. [cited 2024 Aug 7]. Available from: https://platform.openai.com/docs/guides/prompt-engineering/strategy-split-complex-tasks-into-simpler-subtasks
[21] Polverini G, Gregorcic B. How understanding large language models can inform the use of ChatGPT in physics education. Eur J Phys. 2024;45:025701. doi: 10.1088/1361-6404/ad1420

[22] O'Connor S, Peltonen LM, Topaz M, et al. Prompt engineering when using generative AI in nursing education. Nurse Educ Pract. 2024;74:103825. doi: 10.1016/j.nepr.2023.103825
[23] Lorié WA. An application of reverse engineering to automatic item generation: a proof of concept using automatically generated figures. Appleton, WI: Questar Assessment, Inc; 2013 [cited 2024 Aug 8]. Available from: https://files.eric.ed.gov/fulltext/ED545460.pdf
[24] Femepid S, Hatherleigh L, Kensington W. Gradual improvement of contextual understanding in large language models via reverse prompt engineering. Hoboken, NJ: Authorea Preprints; 2024. [cited 2024 Aug 8]. Available from: https://d197for5662m48.cloudfront.net/documents/publicationstatus/217541/preprint_pdf/00c35f30f8009cd3188de015f6fc20cc.pdf
[25] Choi J. AREA: assessment reverse engineering by AI [CAFA training module]. Clarksville: CAFA Lab, Inc; 2024.
[26] Polanyi M. The tacit dimension In: Knowledge in organisations. London: Routledge; 2009. pp. 135–146.
[27] Haynie KC, Haertel GD, Lash AA, et al. Reverse engineering the NAEP floating pencil task using the PADI design system. Menlo Park, CA: SRI International; 2006.
[28] Foghahaee Z. A review of standardizing an English second language test-item through reverse engineering. Lang Teach Res Q. 2019;11:31–42.
[29] Petrovic-Dzerdz M. Reverse engineering a multiple-choice test blueprint to improve course alignment. Collect Essays Learn Teach. 2024;15(1):e7856. doi: 10.22329/celt.v15i1.7856
[30] Hu K. ChatGPT sets record for fastest-growing user base – analyst note. London, UK: Reuters; 2023. [cited 2024 Aug 7]. Available from: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/
[31] Henry MM, Heryanto NA, Isnan M, et al. Automatic multiple choice question generation: A systematic literature review In: 2024 9th International Conference on Information Technology and Digital Applications (ICITDA). IEEE; 2024. pp. 1–7. doi: 10.1109/ICITDA64560.2024.10810102
[32] Korea Health Personnel Licensing Examination Institute Guidelines for developing questions for national health personnel exams. Seoul: Korea Health Personnel Licensing Examination Institute; 2013.
[33] Little TD, Lindenberger U, Nesselroade JR. On selecting indicators for multivariate measurement and modeling with latent variables: when "good" indicators are bad and "bad" indicators are good. Psychol Methods. 1999;4(2):192–211. doi: 10.1037/1082-989X.4.2.192
[34] Boateng GO, Neilands TB, Frongillo EA, et al. Best practices for developing and validating scales for health, social, and behavioral research: a primer. Front Public Health. 2018;6:149. doi: 10.3389/fpubh.2018.00149
[35] Choi J. Promplish [CAFA training module]. Clarksville: CAFA Lab, Inc; 2024.
[36] Choi J. Structured outliner [CAFA training module]. Clarksville: CAFA Lab, Inc; 2024.
[37] Shieber SM. Why scholars should write in Markdown. New York: ACM; 2014.
[38] IBM How do I use Markdown? [Internet]. Armonk, NY: IBM Documentation; [date unknown] [cited 2024 Aug 7]. Available from: https://www.ibm.com/docs/en/SSYKAV?topic=train-how-do-use-markdown
[39] Choi J. Item difficulty emulation via AI (IDEA) In: Collective AI on the foundation AI (CAFA): The pathway of digital transformation of intelligence. Clarksville: CAFA Lab, Inc; 2025.
[40] McHugh ML. Interrater reliability: the kappa statistic. Biochem Med. 2012;22(3):276–282. doi: 10.11613/BM.2012.031
[41] Wongpakaran N, Wongpakaran T, Wedding D, et al. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol. 2013;13:61. doi: 10.1186/1471-2288-13-61
[42] Zhu L, Chen K, Li Z, et al. Can ChatGPT pass the American Heart Association exams? Open-ended questions outperform multiple-choice format. Resuscitation. 2023;188:109783. doi: 10.1016/j.resuscitation.2023.109783
[43] Li M, Jiao H, Zhou T, et al. Item difficulty modeling using fine-tuned small and large language models. Educ Psychol Meas. 2025;85:1065–1090. doi: 10.1177/00131644251344973
[44] Pandarova I, Schmidt T, Hartig J, et al. Predicting the difficulty of exercise items for dynamic difficulty adaptation in adaptive language tutoring. Int J Artif Intell Educ. 2019;29(3):342–367. doi: 10.1007/s40593-019-00180-4
[45] Su MC, Lin LE, Lin LH, et al. Assessing question characteristic influences on ChatGPT's performance and response-explanation consistency: insights from Taiwan's nursing licensing exam. Int J Nurs Stud. 2024;153:104717. doi: 10.1016/j.ijnurstu.2024.104717
[46] OpenAI. GPT-4 is OpenAI's most advanced system, producing safer and more useful responses [Internet]. 2023. [cited 2024 Aug 7]. Available from:https://openai.com/index/gpt-4/
[47] OpenAI. Introducing OpenAI o3 and o4-mini [Internet]. OpenAI. 2025. [cited 2025 May 16]. Available from: https://OpenAI.com/index/introducing-o3-and-o4-mini/
[48] DeepSeek-AI, Guo D, Yang D, et al DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning [Internet]. arXiv [Preprint]. 2025. [cited 2025 May 16]. Available from: https://arxiv.org/abs/2501.12948
[49] Korea Health Personnel Licensing Examination Institute Guidelines for developing questions for national health personnel exams. Seoul: Korea Health Personnel Licensing Examination Institute; 2018.
[50] Choi J. Markup languages in prompt engineering [CAFA training module]. Clarksville: CAFA Lab, Inc; 2024.

[51] Rafique KA, Khaliq N, Grimm C. HyMark: application of hybrid AI for markdown syntax generation In: Proceedings of the 2024 IEEE Conference on Artificial Intelligence (CAI). Piscataway, NJ: IEEE Publications; 2024. pp. 687–694. doi: 10.1109/CAI59869.2024.00134
[52] He J, Rungta M, Koleczek D, et al. Does prompt formatting have any impact on LLM performance? arXiv [Preprint]. 2024. [cited 2025 May 17]. Available from: https://arxiv.org/abs/2411.10541
[53] Emekli E, Karahan BN. AI in radiography education: Evaluating multiple-choice questions difficulty and discrimination. J Med Imaging Radiat Sci. 2025;56(4):101896. doi: 10.1016/j.jmir.2025.101896
[54] Chauhan A, Khaliq F, Nayak KR. Assessing quality of scenario-based multiple-choice questions in physiology: faculty-generated vsdot ChatGPT-generated questions among Phase I medical students. Int J Artif Intell Educ. 2025;35:1–30. doi: 10.1007/s40593-025-00471-z
[55] Choi J. Collective AI on the Foundation AI (CAFA): The pathway of digital transformation of intelligence. Clarksville: CAFA Lab Inc; 2025.
[56] Choi J, Kim S, Yoon K. Collective AI on the Foundation AI (CAFA) Platform. Clarksville: CAFA Lab Inc; 2012–present Available from: https://ai.cafalab.com

Appendices

Appendix 1. Sample of MCQs.

1. An unconscious patient with increasing ICP is on ventilatory support. the nurse notifies the health care provides when arterial blood gas (ABG) measurement results reveal what?
   ① pH of 7.43
   ② PaO_2 of 70 mmHg
   ③ SaO_2 of 94%
   ④ $PaCO_2$ of 35 mmHg

2. A patient has ICP monitoring with an intraventricular catheter. What is a priority nursing intervention for the patient?
   ① Aseptic technique to prevent infection
   ② Consult monitoring of ICP waveform
   ③ Removal of CSF to maintain normal ICP
   ④ Sampling CSF to determine abnormality

3. What drug treatment helps to decrease ICP by expanding plasma and the osmotic effect to move fluid?
   ① Oxygen administrstion
   ② Pentobarbital (Nembutal)
   ③ Mannitol (Osmitro) 25%
   ④ Dexamethasone (Decadron)

4. A patient has a nursing diagnosis of risk for ineffective cerebral tissue perfusion related to cerebral oedema. What is an appropriate nursing intervention for the patient?
   ① Avoid positioning the patient with neck and hip flexion.
   ② Maintain hyperventilation to a PaCO2 of 15 to 20mmHg.
   ③ Cluster nursing activities to provide periods of uninterrupted test.
   ④ Routinely suction to prevent accumulation of respiratory secretion.

5. A patient with a head injury has bloody drainage from ear. what should the nurse do to determine if CSF is present in the drainage?
   ① Examine the tympanic membrane for a tear.
   ② Test the fluid for a halo sign on a white dressing.
   ③ Test the fluid with glucose-identifying strip or stick.
   ④ Collect 5mL of fluid in a test tube and send it to the laboratory for analysis.

6. A nurse monitoring a client who has sustained a head injury would determine that the intracranial pressure (ICP) is rising if which of the following vital sign trends is noted during the course of the work shift?
   ① Increased temperature, decreased pulse, increased respirations, decreased BP
   ② Decreased temperature, increased pulse, decreased respirations, increased BP
   ③Decreased temperature, increased pulse, increased respirations, decreased BP
   ④ Increased temperature, decreased pulse, decreased respirations, increased BP

7. The nurse is admitting a client from the emergency department following a fall that resulted in increased intracranial pressure (ICP). The nurse interprets that the client's Glasgow Coma Scale score has improved the most after making which of the following latest assessments?

① Best eye opening response 5, best motor response 4, best verbal response 8
② Best eye opening response 4, best motor response 6, best verbal response 5
③ Best eye opening response 6, best motor response 5, best verbal response 4
④ Best eye opening response 3, best motor response 8, best verbal response 6

8. The client has been intubated and placed on a mechanical ventilator to reduce intracranial pressure (ICP) by decreasing the carbon dioxide level. Which of the following carbon dioxide values would indicate that the optimal amount of hyperventilation has been achieved?
① $PaCO_2$ 18 mmHg
② PaCO2 29mmHg
③ PaCO2 38mmHg
④ PaCO2 46mmHg

9. When assessing the body functions of a patient with increased ICP, what should the nurse assess first?
① Corneal reflex testing
② Pupillary reaction to light
③ Extremity strength testing
④ Circulatory and respiratory status

10. Which drug treatment helps to decrease ICP by expanding plasma and the osmotic effect to move fluid?
① Oxygen administration
② Mannitol (Osmitrol) (25%)
③ Pentobarbital (Nembutal)
④ Dexamethasone (Decadron)

11. Which of the following is an appropriate nursing intervention for a patient with chronic obstructive pulmonary disease (COPD) experiencing dyspnoea?
① Consume a high amount of grains due to increased energy expenditure.
② Inhale through the mouth and exhale slowly through the nose.
③ Keep the air in the house warm and maintain low humidity.
④ Recommend fluid intake between meals.
⑤ Maintain absolute rest in the supine position.

12. What is the most common causative agent of community-acquired pneumonia?
① Streptococcus pneumoniae
② Staphylococcus aureus
③ Haemophilus influenzae
④ Mycoplasma
⑤ Legionella

13. Which of the following physical examination findings is correct for a patient diagnosed with pneumonia?
① Decreased tactile fremitus upon palpation
② Resonance upon percussion
③ Bronchial breath sounds on auscultation
④ Wheezing on auscultation
⑤ Stridor on auscultation

14. What test is performed to select an appropriate antibiotic for a patient with pneumonia who has a high fever?
① Sputum culture
② Chest X-ray
③ White blood cell count
④ Blood culture
⑤ ESR/CRP test

15. The arterial blood gas (ABG) results of a patient hospitalised with pneumonia are as follows: pH 7.550, PCO2 27 mmHg, PO2 68 mmHg, HCO3- 25mEq/L, SaO2 91%. What is the most appropriate nursing intervention for this patient?
① Connect to a mechanical ventilator.
② Transfuse concentrated red blood cells.
③ Limit endotracheal suctioning if possible.
④ Position in semi-Fowler's and encourage hyperventilation.
⑤ Administer oxygen at 40% FiO2 using a venturi mask.

16. A 75-year-old male with chronic obstructive pulmonary disease (COPD) has fallen into acute respiratory failure due to a respiratory infection. The attending physician immediately inserted an endotracheal tube and applied a mechanical ventilator. If air leakage sounds are heard from the bronchus whenever the patient breathes, what should be checked first?
① Whether food has been aspirated
② The ability to expel sputum
③ Whether the ventilator tube is twisted
④ The inflation status of the endotracheal tube cuff
⑤ Whether the patient can speak

17. A patient with chronic obstructive pulmonary disease (COPD) complained of difficulty breathing and underwent an arterial blood gas analysis (ABGA) with the following results: pH 7.23, PaO2 67 mmHg, PaCO2 62 mmHg, HCO3- 25mEq/L. What is the appropriate nursing intervention?
① Advise rest.
② Instruct to breathe through pursed lips.
③ Provide a plastic bag for breathing.
④ Educate to breathe deeply and quickly due to lack of oxygen.
⑤ Use an oxygen mask to continuously supply 10 L of oxygen.

18. A patient with chronic obstructive pulmonary disease (COPD) has a forced expiratory volume in one second (FEV1) of 30%. What level of dyspnoea can be expected?
① No dyspnoea.
② Dyspnoea on exertion.
③ Dyspnoea at rest.
④ Dyspnoea during exercise.
⑤ Dyspnoea when climbing stairs.

19. A nurse working in the intensive care unit (ICU) of an internal medicine department is caring for a pneumonia patient on a mechanical ventilator. The high-pressure alarm on the ventilator goes off, and after silencing the alarm, a physical examination reveals crackle sounds in both lower lobes of the lungs. What is the most appropriate nursing intervention?
① Measure vital signs.
② Perform endotracheal suctioning.
③ Check if the patient's breathing has returned.
④ Examine all the tubes connected to the mechanical ventilator.
⑤ Check if the endotracheal tube balloon is properly inflated.

20. What nursing intervention is aimed at improving an ineffective breathing pattern due to pneumonia?
① Use an incentive spirometer for inhalation training.
② Frequently recommend postural drainage to facilitate sputum discharge.
③ Advise suppressing coughs to alleviate chest pain.
④ Apply an oral pharyngeal airway to make breathing more comfortable.
⑤ For the relief of crackles, it is recommended to limit fluid intake to within 1,000 mL.

Appendix 2. Prompt for MCQs development.

1. Introduction
1.1. You are a nursing educator specialising in medical-surgical nursing.
1.2. Based on the following contents, you will create Multiple-Choice Questions (MCQs) targeting nursing students who have completed the neurology system in medical-surgical nursing.
1.3. Take your time and think step-by-step, then create the questions afterward.
2. Stages of developing MCQs
2.1. Refer to the 'PDF file' for the scope of MCQs.
2.2. Refer to the '2.9. Example of the MCQs' for developing isomorphic MCQs.
2.3. Write down the chapter number referenced from the PDF file.
2.4. Write down the answer and the rationale for your chosen answer in one line for each item.
2.6. Refer to sections 2.7. to 2.10. for information on developing MCQs.
2.7. Scope of MCQs formulation
2.7.1. Topic: Intracranial pressure (ICP) within the context of the nervous system in medical-surgical nursing
2.7.2. Refer to the 'PDF file' attached.
2.8. Format of MCQs

2.8.1. Number of items: 1

2.8.2. Number of options per item: 4

2.9. Example of MCQs

2.9.1. Item: An unconscious patient with increasing ICP is on ventilatory support. the nurse notifies the health care provides when arterial blood gas (ABG) measurement results reveal what?

2.9.2. Options: ① pH of 7.43 ② PaO2 of 70 mmHg ③ SaO2 of 94% ④ PaCO2 of 35 mmHg

2.10. Criteria for developing MCQs

2.10.1. Stems of MCQs

2.10.1.1. The question stem should allow for the correct answer to be inferred without looking at the options.

2.10.1.2. The question stem should use only essential words or phrases.

2.10.1.3. The content being asked in the question stem should be easily understood by the examinee.

2.10.1.4. The question stem should not be unnecessarily pessimistic.

2.10.2. Options of MCQs

2.10.2.1. The options should be homogeneous.

2.10.2.2. The options should not give hints to each other.

2.10.2.3. There should only be one correct answer, and multiple answers should not be possible.

2.10.2.4. If words are repeated in the options, they must be necessary.

2.10.2.5. The order of the options should be regular or logical.

Appendix 3. Evaluation of five nursing education experts about the compliance with the nine principles of item development of MCQs generated by ChatGPT-4 (N = 5).

MCQ No.		Number of principles of item development that complied*					
Sample MCQ	MCQs generated by ChatGPT-4	Expert1	Expert2	Expert3	Expert4	Expert5	Average
1	1	7	8	6	7	6	6.8
	2	8	8	8	7	7	7.6
	3	8	9	5	4	7	6.6
2	4	8	9	5	4	7	6.6
	5	8	9	6	7	7	7.4
	6	9	9	5	5	7	7.0
3	7	8	9	9	7	5	7.6
	8	9	9	9	8	8	8.6
	9	9	9	9	8	8	8.6
4	10	9	9	7	7	7	7.8
	11	9	9	8	9	9	8.8
	12	8	9	7	6	9	7.8
5	13	9	9	8	8	7	8.2
	14	9	9	9	9	9	9.0
	15	9	9	9	9	8	8.8
6	16	9	9	7	8	8	8.2
	17	9	9	7	8	9	8.4
	18	7	9	7	8	7	7.6
7	19	4	8	4	5	8	5.8
	20	4	8	4	6	9	6.2
	21	4	8	4	5	8	5.8
8	22	7	8	5	7	8	7.0
	23	7	8	5	7	6	6.6
	24	8	8	7	7	7	7.4
9	25	8	8	7	7	7	7.4
	26	6	8	7	7	8	7.2
	27	8	7	7	5	8	7.0
10	28	8	8	8	8	8	8.0
	29	9	8	8	7	8	8.0
	30	9	8	9	8	8	8.4
11	31	9	8	8	8	8	8.2
	32	9	8	9	9	9	8.8
	33	9	9	8	9	9	8.8
12	34	9	9	9	9	9	9.0
	35	9	9	9	9	9	9.0
	36	9	9	8	9	9	8.8
13	37	9	9	9	9	9	9.0
	38	8	8	9	9	9	8.6
	39	8	8	8	8	9	8.2
14	40	8	8	9	8	9	8.4
	41	9	9	4	9	9	8.0
	42	9	9	4	9	9	8.0
15	43	9	9	4	9	9	8.0
	44	8	9	8	9	9	8.6
	45	9	9	8	9	9	8.8
16	46	8	8	8	8	9	8.2
	47	9	9	7	8	9	8.4
	48	9	9	5	9	9	8.2
17	49	9	9	4	8	9	7.8
	50	5	9	6	8	8	7.2
	51	5	9	6	7	9	7.2
18	52	9	9	6	7	8	7.8
	53	3	9	6	5	8	6.2
	54	8	9	6	9	8	8.0
19	55	5	9	9	7	9	7.8
	56	5	9	9	9	9	8.2
	57	5	8	9	8	9	7.8
20	58	9	9	7	9	9	8.6
	59	9	8	6	9	9	8.2
	60	9	8	7	9	9	8.4
Average*							7.9

*Scores range from 0–9.
MCQ, multiple-choice question.

Appendix 4. Evaluation of five nursing education experts about the compliance with each principle of item development (N = 5).

MCQ No.		Number of nursing education experts who responded that the principle of item development was complied with*									
Sample MCQ	MCQs generated by ChatGPT-4	P1	P2	P3	P4	P5	P6	P7	P8	P9	Average
1	1	2	2	4	3	5	5	4	5	4	4.4
	2	2	3	5	3	5	5	5	5	5	4.4
	3	2	5	5	3	5	4	5	5	4	4.3
2	4	3	2	5	3	3	5	4	4	4	4.4
	5	4	5	5	3	2	5	4	5	4	4.4
	6	4	2	5	4	4	4	3	5	4	4.4
3	7	5	2	5	5	4	4	5	5	3	4.4
	8	5	3	5	5	5	5	5	5	5	4.4
	9	5	4	5	5	5	5	5	5	4	4.4
4	10	5	2	5	5	5	5	5	4	3	4.4
	11	5	5	5	5	5	5	5	5	4	4.4
	12	5	3	5	5	5	4	5	4	3	4.4
5	13	5	2	4	5	5	5	5	5	5	4.5
	14	5	5	5	5	5	5	5	5	5	4.4
	15	5	5	5	5	4	5	5	5	5	4.4
6	16	5	2	5	5	5	5	5	4	5	4.4
	17	5	3	5	5	5	5	5	4	5	4.4
	18	5	5	5	5	5	5	4	3	5	4.4
7	19	3	2	4	5	4	2	4	3	2	4.4
	20	3	4	4	5	4	2	4	3	2	4.4
	21	3	2	4	5	4	2	4	3	2	4.5
8	22	3	5	5	5	5	4	4	4	4	4.6
	23	3	5	4	5	5	4	3	4	4	4.6
	24	3	5	4	5	5	5	4	5	5	4.6
9	25	3	5	4	5	5	5	3	5	5	4.6
	26	4	5	5	5	4	5	2	5	4	4.6
	27	4	4	5	5	5	5	3	5	4	4.6
10	28	4	2	5	5	5	5	5	5	4	4.6
	29	4	4	5	5	5	5	5	5	4	4.6
	30	4	2	5	5	5	5	5	5	5	4.6
11	31	5	5	5	4	5	5	5	5	5	4.6
	32	5	4	5	5	5	5	5	5	5	4.6
	33	5	5	5	5	5	5	5	5	5	4.6
12	34	5	5	5	5	5	5	5	5	5	4.6
	35	5	5	5	5	5	5	5	5	4	4.6
	36	5	5	5	5	5	5	5	5	5	4.6
13	37	4	5	5	5	5	5	4	5	5	4.6
	38	4	3	5	5	5	5	4	5	5	4.6
	39	4	5	5	5	5	5	5	3	5	4.5
14	40	5	5	4	5	4	4	4	5	4	4.5
	41	5	5	4	5	4	4	4	5	4	4.5
	42	5	5	4	5	4	4	4	5	4	4.6
15	43	5	5	5	5	5	5	3	5	5	4.6
	44	5	4	5	5	5	5	5	5	5	4.6
	45	4	3	5	5	5	5	4	5	5	4.6
16	46	5	4	5	5	5	5	5	5	3	4.5
	47	5	4	4	4	5	5	4	5	5	4.6
	48	5	4	4	4	4	4	5	5	4	4.6
17	49	4	5	3	5	5	4	4	5	5	4.6
	50	4	2	3	5	5	4	3	5	5	4.6
	51	4	3	3	5	5	4	3	5	4	4.6
18	52	5	2	5	5	5	5	4	4	4	4.6
	53	3	2	3	5	4	3	3	4	4	4.7
	54	5	3	5	5	5	5	3	4	5	4.8
19	55	4	3	4	5	5	5	4	5	4	4.8
	56	4	4	4	5	5	5	4	5	5	4.9
	57	2	4	4	5	5	5	4	5	5	4.9
20	58	4	5	4	5	5	5	5	5	5	4.9
	59	3	5	4	5	5	5	4	5	5	4.8
	60	3	5	4	5	5	5	5	5	5	4.7
Average		*4.2	3.8	4.6	4.8	4.7	4.6	4.3	4.7	4.4	4.5

*Scores range from 0–5.

MCQ, multiple-choice question; *P*, principle of item development; P1, the question stem should allow for the correct answer to be inferred without looking at the options; P2, the question stem should use only essential words or phrases; P3, the content being asked in the question stem should be easily understood by the examinee; P4, the question stem should not be unnecessarily pessimistic; P5: the options should be homogeneous; P6, the options should not give hints to each other; P7, there should only be one correct answer and multiple answers should not be possible; P8, if words are repeated in the options, they must be necessary; P9, the order of the options should be regular or logical.