Received: 13 June 2024 | Accepted: 22 August 2024

DOI: 10.1002/prp2.70004

ORIGINAL ARTICLE

Evaluating the quality of multiple-choice question pilot database: A global educator-created tool for concept-based pharmacology learning

Adeladlew Kassie Netere[1] | Tony Hughes[1] | Anna-Marie Babey[2] | Martin Hawes[3] | Janet Mifsud[4] | John P. Kelly[5] | Willmann Liang[6] | Mark Hernandez[7] | Kelly Karpa[8] | Hesham Al-Sallami[9] | Lynette B. Fernandes[10] | Patrik Aronsson[11] | Carolina Restini[12] | Fabiana Caetano Crowley[13] | Elvan Djouma[14] | Tina Hinton[15] | Johnson J. Liu[16] | Fatima Mraiche[17] | Paul J. White[1]

Correspondence
Paul J. White, Faculty of Pharmacy and Pharmaceutical Sciences, Monash University, Parkville, Vic, Australia.
Email: paul.white@monash.edu

Abstract

The Core Concepts of Pharmacology (CCP) initiative is developing educational resources to transform pharmacology education into a concept-based approach. This study evaluated the quality of global educator-created MCQs in generating items for the pharmacology concept inventory (PCI) instrument and developed as a resource for learning pharmacology fundamental concepts. A panel of 22 global pharmacology experts recruited from the CCP initiative research team participated in the MCQ pilot database design and evaluation. The quality analysis framework of the MCQs in the pilot database included four assessment tools: item writing guidelines (IWGs), Bloom's taxonomy, the CCP, and the MCQ design format. A two-phase evaluation process was involved, including inter-rater agreement on item quality, followed by resolving conflicts that occurred in quality assessment. The chi-square (χ^2) test of independence and Cramer's V correlation tests were utilized to measure the relationship among quality assessment attributes. About 200 MCQs were gathered and 98% underwent expert evaluation. Nearly 80% addressed one or more CCP, with 52% designed using a context-dependent format. However, only 40% addressed higher levels of Bloom's cognitive domain and 10% adhered to all IWGs. A strong positive correlation was observed between the context-based item format and its effectiveness in assessing the higher cognitive domain, the main CCP and improved IWGs adherence. Context-based item construction can assess the higher cognitive skills and fundamental pharmacology concepts, showing potential for rigorous PCI development. The pilot database will store items to create the PCI, aiding the development of a concept-based pharmacology curriculum.

For affiliations refer to page 10.

KEYWORDS
concept-based approach, context-dependent item, multiple-choice question, pharmacology learning, quality evaluation

1 | INTRODUCTION

The International Society for Basic and Clinical Pharmacology Education Section (IUPHAR-Ed) Core Concepts of Pharmacology (CCP) initiative has proposed the development of a concept-based approach to pharmacology learning.[1–3] This approach aims to transform pharmacology learning by creating conceptual assessment instruments and educational materials. These resources are designed to assist students and educators in mastering the fundamental principles of pharmacology. Pharmacology educators have identified, defined, and unpacked the core concepts.[1,3] The next phase encompasses designing instruments for teaching and assessing these concepts. This involves developing and validating concept inventories (CIs) to identify misconceptions and measure conceptual understandings. The CI design includes establishing a multiple-choice question (MCQ) database created by international pharmacology educators to assess student comprehension.[2]

MCQs are promising tools for developing CI instruments, which are used to assess students' learning on established conceptual frameworks.[4] One approach in developing CI items involves a methodological focus on pinpointing student misconceptions by analyzing their responses to MCQs.[5–7] This approach consists of creating MCQ distractors using list of validated misconceptions.[8] MCQs are widely employed in higher education to assess professional competence due to their efficiency in evaluating diverse content,[9,10] ease of grading, and suitability for large-scale assessments.[11,12] Well-written and high-quality MCQs can provide comprehensive assessment, foster critical thinking, and improve decision-making skills.[13] Moreover, they can play a role in broader educational improvements, resulting in improved student performance, and the establishment of validated assessment practices.

Despite their benefits, MCQ fall short in evaluating the depth of understanding, and practical skills, focusing on selecting correct answers rather than expressing reasonings and articulating thoughts in professional settings.[14,15] Also, the quality of MCQs could be compromised due to violations of accepted quality standards, termed item-writing guidelines (IWGs).[9,16,17] This non-compliance results in lower-quality items, adversely affecting student performance and compromies critical decisions.[16] It can hinder cognitive levels and item parameters, allowing low-ability individuals to guess correctly while misleading high-ability test-takers.[18] MCQs often assess lower cognitive levels, reducing their effectiveness,[15,17] particularly in health sciences when learners are expected to analyze complex information for clinical decisions.[15] If assessment tools fail to measure complex cognitive skills, students may not be evaluated on whether they demonstrated higher cognitive levels of knowledge and skills.[17] Thus, quality MCQs are recommended to promote higher-order critical thinking and prepare students for clinical training or research.[16,19]

This study aimed to assess the quality of MCQs in the pilot database submitted by international pharmacology educators and generate potential items for the pharmacology concept inventory (PCI) instrument. A novel methodological approach combined four analysis frameworks to produce a holistic assessment of MCQ development quality and suitability for testing fundamental pharmacology concepts: (1) evaluating instructor-created MCQ quality based on adherence to IWGs; (2) examining the cognitive domain levels embedded in MCQs to assess students' critical thinking skills in their pharmacology knowledge; (3) evaluating of the effectiveness of MCQs in testing the CCP; and (4) analyzing the design structure of MCQs as context-dependent or non-context-dependent to identify the construction format.

2 | QUALITY ASSESSMENT FRAMEWORK

MCQ quality evaluation utilized a blend of four instruments: IWGs, Bloom's taxonomy, CCP and the design nature of MCQ structure. These instruments aim to craft high-quality MCQs able to assess critical thinking and cognitive skills, evaluate fundamental concepts of pharmacology, and design MCQ formats in real scenario vignettes. Educational instruments such as IWGs,[17,20] Bloom's Taxonomy [21] and CCP[1,3] have been developed to address these challenges. These tools serve as a fundamental framework for designing MCQs and assessing their quality in alignment with educational learning objectives, while also providing a classification for the goals of education systems within a cognitive model of complexity.[22] Using integrated instruments to assess items improves quality analysis making the teaching-learning process simpler and more effective. They also enable the monitoring and evaluation of new learning performance and competency.[21,23] Assessing the quality of MCQs to test CCPs according to these frameworks ensures the development of well-designed instruments that effectively measure learning outcomes.[1,9,20,21] High-quality MCQs will be considered further to create a repository of items for PCI instruments.

3 | METHODS AND MATERIALS

The assessment involved multiple stages such as MCQ item pilot database establishment, item categorization according to CCP, and a two-step quality evaluation conducted by a group of international pharmacology educators.

3.1 | MCQ pilot database construction

Experts in pharmacology education from the IUPHAR-Ed CCP initiative research group were asked to provide MCQs from their question sets. A team of international pharmacology educators submitted the items and the MCQ pilot database was constructed. Some of the MCQs had been crafted specifically to assess the fundamental pharmacology concepts linked to the identified CCP list. These questions were developed over several years and utilized at individual institutions to evaluate students' academic performance and competencies within a pharmacology course. They were intended for a diverse range of undergraduate and graduate students in degree-seeking programs (e.g., health science professions, biomedical science programs, etc.) across universities around the world. Additionally, undergraduate pharmacy students enrolled in Monash University's Faculty of Pharmacy and Pharmaceutical Sciences developed approximately one-quarter of the questions as part of their thesis project. The students underwent training from a senior pharmacology content expert on how to craft MCQs and received guidance on the utilization of IWGs. The items they developed were then reviewed by two research team members and qualified items were included in the MCQ pilot database.

3.2 | MCQ pilot database evaluation

Leveraging their prior engagement in the CCP framework, international pharmacology education content experts were invited via email to evaluate the quality of the MCQ pilot database. These experts were recruited from several countries across four continents based on their deep understanding of pharmacology, extensive educational experience, and expertise in MCQ development. Their pharmacology teaching and research experiences ranged from 6 to over 20 years (Table S1). The experts were then grouped into seven sub-groups, each composed of two experts, which were subsequently formed based on their preferences and CCP expertise. Each group received approximately 30 MCQs to analyze using the analysis framework described below. Item providers participated in the review process following the anonymous coding and categorization of items to prevent any potential bias. This approach ensured that items were assessed by reviewers other than their original author, thereby maintaining impartiality. Subsequently, the research team introduced a one-hour online briefing for the evaluator teams, explaining how the analysis framework was employed in assessing the MCQs pilot database. This evaluation aimed to determine the quality of MCQ writing, assess their ability to address higher cognitive domains, evaluate their effectiveness in testing fundamental pharmacology concepts, examine item design structure, and ultimately use them as a repository for generating an item for the PCI instrument.

3.3 | Analysis framework

This analysis framework employed the aforementioned four instruments to evaluate the quality of MCQs submitted by international educators and the MCQs in the pilot database. This innovative approach includes (1) IWGs, which served as a benchmark for crafting high-quality MCQs and ensuring adherence to the standards; (2) Bloom's taxonomy for cognitive functions, used to assess the efficacy of MCQs in evaluating critical thinking and professional competency[22]; (3) CCP employed to evaluate MCQs for their quality to test the fundamental concepts of pharmacology[1]; and (4) MCQ design formatting, a framework used to examine the design format of items ensuring whether the question is context-dependent or context-independent and its correlation to other quality assessment attributes. This approach facilitated a thorough examination of the MCQs, ensuring their suitability for testing core concepts in a detailed way (Figure 1).

3.3.1 | Item writing guidelines (IWGs)

IWGs offer a structured framework to construct and design standard MCQs, enhancing quality and ensuring precision in evaluating knowledge or skills across various educational contexts.[20,24] Assessing global educator-created MCQs with IWGs may improve clarity, validity, and reliability, supporting context-based learning in pharmacology education. Despite the absence of standardized IWGs and inconsistencies in their utilization, improving the writing quality of MCQs is crucial. Consequently, we compiled a list of IWGs through a comprehensive review of the literature and widely used test-writing guidelines, which encompassed identifying, adopting and compiling IWGs to mitigate potential limitations.[16,17,20,25] The sub-group of the research team identified and refined 23 elements from the work of Haladyna and colleagues,[20] as well as Tarrant and co-workers,[17] revised guidelines incorporating expert feedback. These elements were identified in previous studies as areas of noncompliance[16,17,25] (Table 1).

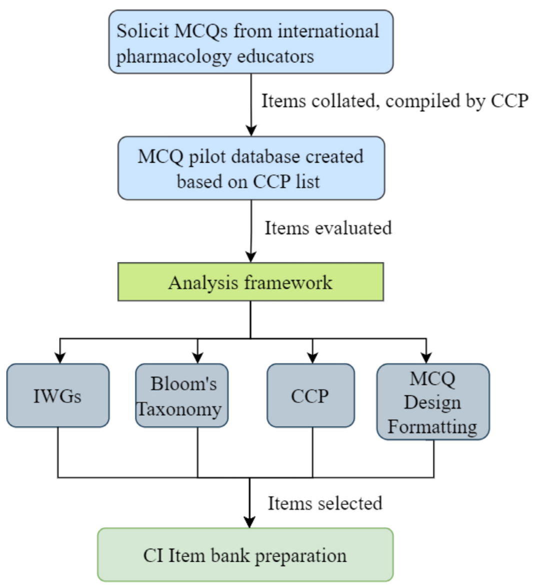

FIGURE 1 MCQ collation and quality evaluation framework.

TABLE 1 Compiled item-writing guidelines (IWGs) assess the MCQ writing quality.

Context categories	Specific guidelines and codes used in this study
Content context	1. Every item should reflect specific content and a single specific mental behaviour. 2. Use novel material to test higher-level learning. 3. Keep the content of each item independent from the content of other items. 4. Avoid over-specific and over-general content when writing MC items. 5. Each MCQ should have a clear and focused question. 6. Questions and all options should be written in clear, unambiguous language.
Formatting concerns	7. Avoid complex or K-type MCQs (a combination of other given statements).
Style concerns	8. All options are grammatically consistent with the stem. 9. Avoid repeating words in the stem and the correct option.
Writing the stem	10. Each MCQ should have the problem in the stem of the question, not in the option. 11. Avoid gratuitous or unnecessary information in the stem or the options. 12. Avoid awkward stem structure (Finish the sentence, fill in the blank, grammatically flawed).
Writing the choices	13. The basic format for MCQs is the single best answer. 14. Arrange MCQ options in alphabetical, chronological, or numerical order. 15. All options should be similar in length and amount of detail. 16. Avoid the use of "none of the above" as the last option. 17. Avoid the use of "all of the above" as the last option. 18. Avoid the use of negatives (e.g., not, except, incorrect). 19. Avoid providing logical cues in the stem and the correct option that helps the student identify the correct option without knowing the material. 20. Avoid convergence cues in options where there are different combinations of multiple components to the answer. 21. Options should be worded to avoid the use of absolute terms (e.g., never, always, only, all). 22. Options should be worded to avoid using vague terms (e.g., frequently, occasionally, rarely, usually, commonly). 23. Make all distractors plausible.

3.3.2 | Cognitive domains of Bloom's taxonomy

This framework was used to examine the cognitive domains embedded in the MCQs. It categorizes cognitive function into six domains with the ordering of the two highest domains having been swapped in a revised version.[26] This framework was further categorized into two cognitive levels: K1 or K2.[21,26] K1 (lower cognitive domains) represents knowledge and comprehension, the foundational tier of learning, focusing on the recall or recitation of facts, observations, or definitions. K2 (higher cognitive domains) assesses the higher cognitive domains of application and analysis, which include intellectual abilities and skills that denote a higher level of learning and are defined as "the mental process of organizing and working with materials and problems to achieve a purpose".[21] Therefore, we assessed the extent to which MCQs integrate the six domains of cognitive domains and two categories[27] (see Table S2).

3.3.3 | CCP

This framework was employed to examine the quality of educator-created MCQs in assessing the CCP. These concepts were previously identified and refined by White et al, Guilding et al, and the CCP research team.[1,3] The CCP framework was then used to ensure the items were sufficient to test the core principles and knowledge of the discipline and to ensure a context-based learning and teaching approach. This consolidated framework empowers pharmacology educators to prioritize critical conceptual knowledge recognized by field experts. The MCQ pilot database sourced from international pharmacology educators was created to serve as the foundation for the PCI instrument. To ensure a context-based learning and teaching approach, the quality of the items was evaluated against this CCP framework.

3.3.4 | MCQ design formatting

Pharmacology, as a discipline, employs both quantitative and qualitative models to describe (or explain) concepts.[28,29] The CCP list includes both clinical and mathematical models[1,3] and the pharmacology MCQs design may employ diverse structures. The MCQ construction format is used to assess whether the MCQ was created as a context-dependent or context-independent style in the application of evaluating the learner's comprehension, critical thinking skills, and ability to apply knowledge in real-life or context-specific situations. For this study, we classified MCQ design structures as context-dependent if they were constructed in case scenarios/vignettes, graphs, and tables format. Items or sets of items presented as reading passages or statements were categorized as context-independent.

3.4 | Assessment approach and scoring procedure

Initially, we received single correct response MCQs that were already categorized according to the CCP framework.[1,3] IWGs were utilized with a nominal metric system, rated as "Yes" for those that adhered to the guidelines and "No" for those that did not. This approach aimed to determine whether the MCQ exhibited flaws, specifically focusing on item-writing flaws. This framework comprised 23 elements with five sub-categories to assess the MCQs: content concerns (six elements), formatting concerns (one element), style concerns (two elements), writing the stem (three elements), and writing the choices (11 elements).[20] MCQs were assessed by the cognitive domain framework, which contained six levels: knowledge, comprehension, application, analysis, synthesis, and evaluation. They were further analyzed based on two categories of cognitive levels: K1 (lower cognitive domains), or K2 (higher cognitive domains). Also, MCQs were rated as "effective" if they adequately assessed one sub-concept or the main idea of pharmacology, or as "not effective" if they did not meaningfully assess the concept or only assessed minor aspects, according to the CCP framework. Additionally, the MCQ design structure was rated as context-dependent if they were constructed in case scenarios/vignettes, graphs, and tables format, and context-independent if designed as none of the listed structures or presented as a clear statement.

3.5 | Inter-rater reliability (IRR) for MCQ pilot database

We utilized a two-step process for the evaluation and validation of the MCQ pilot database. The IRR analyses were employed in the "irr" package in R.[30]

3.5.1 | First phase: Item evaluation

Fourteen experts took part in the initial evaluation process, with each item being independently rated by two experts. The degree of agreement, such as percentage agreement and IRR, were assessed. The selection of IRR metrics was based on the assumptions of study design,[31] with statistical tests chosen according to the metric factor. For assessing the nominal (Yes/No) factor of item writing quality, common kappa was analyzed based on the IWGs.[31–33] The cognitive domains of Bloom's taxonomy, consisting of six categories, and the CCP four categories, were assessed using weighted kappa as an ordinal metric system. To enhance understanding of reliability levels, the percentage agreement among raters was combined with kappa reliability tests. While there are no rigid rules and strict guidelines dictating the required level of agreement for evaluation, we employed commonly referenced rule-of-thumb levels for percentage agreement and Cohen's Kappa classifications for IRR. Percentage agreement ranging from 75% to 90% was suggested as acceptable.[34] IRR was interpreted using kappa coefficients (κ) as follows $\kappa \leq 0$ indicates poor, 0.01–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1.00 perfect agreement.[35]

3.5.2 | Second phase: Resolving disagreements between first raters

Items for which there was disagreement between the first two raters in the initial assessment phase were flagged and sent to a third rater for resolution. Each rater evaluated half of the items, identifying writing flaws and assessing how well the items tested CCP and cognitive domain levels. Items that still had rating discrepancies, especially concerning the effectiveness of items in assessing the ordinal categorical levels of CCP and the cognitive domain, were discussed by all raters if unresolved by the third rater. Virtual discussions (via Zoom) among all raters were used to address any remaining rating discrepancies.

3.6 | Correlation analysis of MCQ quality attributes

Pearson's chi-square (χ^2) and Cramer's V correlation coefficient tests were utilized to assess the correlation and strength of the relationship among quality assessment attributes. The chi-square test was employed to investigate the association among categorical variables, while the strength of association was determined using Cramer's V correlation coefficient test.[36]

4 | RESULTS

4.1 | Item provider and evaluator expertise

A panel of 22 experts participated in MCQ provision and evaluation. Among them, 10 were women, eight were men and one preferred not to respond, while background information was unavailable for three experts. Initially, we sent an email request to a group of experts to submit MCQs, 13 of whom provided items from their existing question sets. Then, a panel of 16 experts, seven from the provider group and 10 additional experts evaluated the MCQs. The majority (77%, $n=17$) of experts held PhD degrees, including Doctor of Science and PharmD degrees. Experts were affiliated with universities in eight countries across four continents. Experts possess extensive experience in pharmacology education and research, including teaching disciplines such as basic and clinical pharmacology, systems-based pharmacology, pharmacokinetics, pharmacotherapeutics, drug discovery and development, and pharmacogenomics. Most experts (82%, $n=18$) have experience teaching cohorts ranging from 50 to 250 students enrolled in a variety of undergraduate and graduate degree-seeking programs (e.g., health science professions, biomedical science programs, etc.). The programs included biomedical or medical science, pharmacy, nursing, podiatry, physiology, optometry, dentistry, physiotherapy, and veterinary medicine , including MSc students and PhD candidates. Additionally, some experts have experience teaching paramedicine and non-medical prescribers such as nurses, paramedics, and physiotherapists (team members' expertise is illustrated in Table S1).

4.2 | Quality assessment of the MCQ pilot database

After compiling 200 global pharmacology educators-designed questions, we created a pilot database of MCQs that underwent expert evaluation. Four (2%) of these were excluded since they lacked item structure details and were unable to be evaluated according to the analysis framework. As a result, 98% of the MCQs were deemed suitable for evaluation. These items showed different levels of adherence to the IWGs, covered various cognitive domains, tested the CCP and were designed in various design structures.

4.3 | Adherence to the IWGs

For the IWG framework, 196 MCQs were evaluated, of which 90% were non-adherent to at least one IWG. Specifically, 20 (10.3%) adhered to the IWGs without any violations, 45 (23.1%) violated one IWG, and 53 (27.2%) items had two or three flaws. The remaining items (n=78, 40.0%) had four or more flaws. The most frequent omitted elements were failure to "arrange MCQ options in alphabetical, chronological, or numerical order" (66%, n=129), construct items with "use novel material to test higher-level learning" (52%, n=101), and "avoid awkward stem structure" (29%, n=56). Guidelines such as "avoid the use of absolute terms", "none of the above", and "content of each item was independent of others" had a high degree of adherence (Figure 2). Acceptable ranges (75%–98%) of percentage agreement were observed between raters on the quality assessments using IWGs. The IRR scores revealed substantial agreement for two guidelines (κ=.72, .79), and moderate agreement for three guidelines (κ=.44, .59, .60). The remaining were categorized as fair ($\kappa \leq$.40) or slight ($\kappa \leq$.20) (Table 2).

4.4 | MCQ cognitive domain category

For this framework out of 195 MCQs, 60% (n=117) addressed lower cognitive levels (38% on knowledge and 22% on comprehension). The remaining 40% (n=78) assessed a higher cognitive level (30% on application and 10% for analysis). None of the MCQs addressed the top two domains such as synthesize or evaluate (Figure 3). The degree of rater agreement was measured by percentage agreement, showing acceptable agreement (43.5%), and weighted kappa indicating moderate reliability on cognitive level measurements (κ=.52; Table 3).

4.5 | Item effectiveness in assessing the CCP

In the first phase of evaluation, there was a 44% agreement between raters, with a fair reliability level (κ=.25) indicating an agreement on the effectiveness of testing fundamental concepts of pharmacology (Table 3). In the second phase of evaluation, after disagreements were resolved, the majority of MCQs (79%, n=154) were rated as effective in addressing one or more sub-concepts of the CCP, while the remainder did not address a concept or only covered a minor aspect.

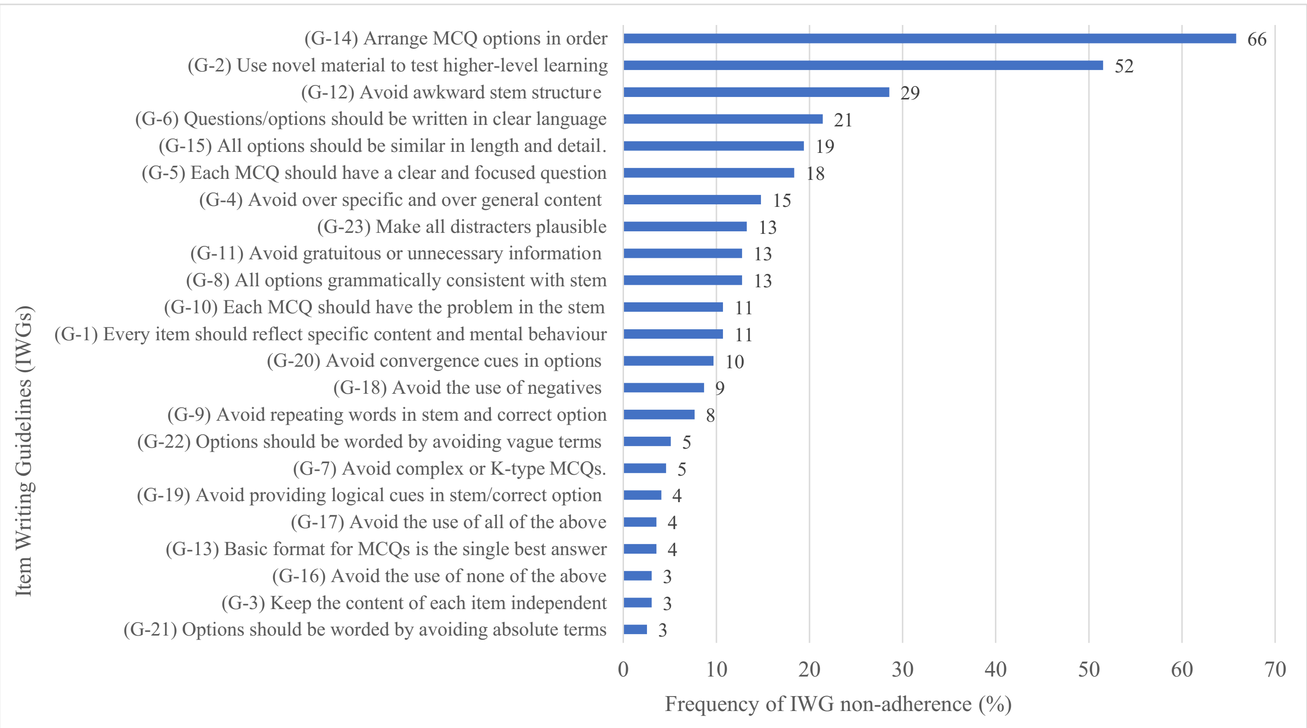

FIGURE 2 Assessment of MCQs compliance to the IWGs by global pharmacology educators.

TABLE 2 The degree of rater agreement on the MCQs pilot database measured with IWGs.

Inter-rater agreement					
	Agreement level	Kappa Statistic IRR Scores			
IWGs	% agreement	Kappa Value	Agreement level	Z-Score	*p*-value
IWG1	79	0.07	Slight	1.04	
IWG2	72	0.44	Moderate	6.19	<.001
IWG3	84	0.03	Slight	0.46	
IWG4	78	0.20	Fair	2.94	.003
IWG5	68	0.12	Slight	1.75	
IWG6	67	0.19	Slight	2.69	.007
IWG7	95	0.59	Moderate	8.29	<.001
IWG8	85	0.39	Fair	5.52	<.001
IWG9	88	0.16	Slight	2.70	.007
IWG10	82	0.33	Fair	4.65	<.001
IWG11	75	0.03	Slight	0.43	
IWG12	71	0.16	Slight	2.29	.022
IWG13	91	0.22	Fair	3.13	.002
IWG14	NA	0.32	Fair	4.28	<.001
IWG15	83	0.37	Fair	5.38	<.001
IWG16	98	0.72	Substantial	10.50	<.001
IWG17	97	0.60	Moderate	8.71	<.001
IWG18	96	0.79	Substantial	11.20	<.001
IWG19	90	0.04	Slight	0.62	
IWG20	84	0.163	Slight	2.40	.017
IWG21	95	0.163	Slight	2.59	.010
IWG22	88	0.106	Slight	2.00	.046
IWG23	79	0.147	Slight	2.20	.028

Note: Agreement level is categorized based on Kappa value.
Abbreviations: IWGs, item writing guidelines; NA, not applicable.

4.6 | MCQ design format

The MCQs were evaluated based on the design format of items and item sets. More than half of the MCQs (52%, $n=101$) were designed with context-dependent structures such as case scenarios/vignettes, graphs, and tables, while the remaining were created as statement-based.

4.7 | Correlation analysis of quality assessment attributes

Quality assessment attributes association and strength of association were revealed by Pearson's chi-square (χ^2) test of independence and Cramer's V correlation coefficients test. MCQs designed with a context-dependent structure showed a strong positive correlation with other quality measuring criteria: item quality for addressing higher cognitive domains, assessing main pharmacology concepts and improving item adherence to IWGs (Cramer's $V=.66$, .54 and 0.41; $p<.001$), respectively. Likewise, a substantial positive correlation was observed between items adhering to IWGs and item quality for addressing higher cognitive domains (Cramer's $V=.43$; $p=.002$) and assessing main pharmacology concepts (Cramer's $V=.42$; $p<.001$). Similarly, MCQs assessing the main pharmacology concepts were correlated with items addressing higher cognitive domains (Cramer's $V=.37$; $p<.001$; Table 4).

5 | DISCUSSION

This study evaluated the quality of global educator-created MCQ writing, assessed their ability to address higher-order cognitive levels, and fundamental pharmacology concepts, and examined item design structure. The items will be used to design resources that advance the concept-based approach to pharmacology education advocated by the IUPHAR-Ed CCP initiative.[2] High-quality MCQs will be added to a pool to create an item bank, thereby facilitating the creation of PCI instruments. The PCI will be used to identify misconceptions, measure

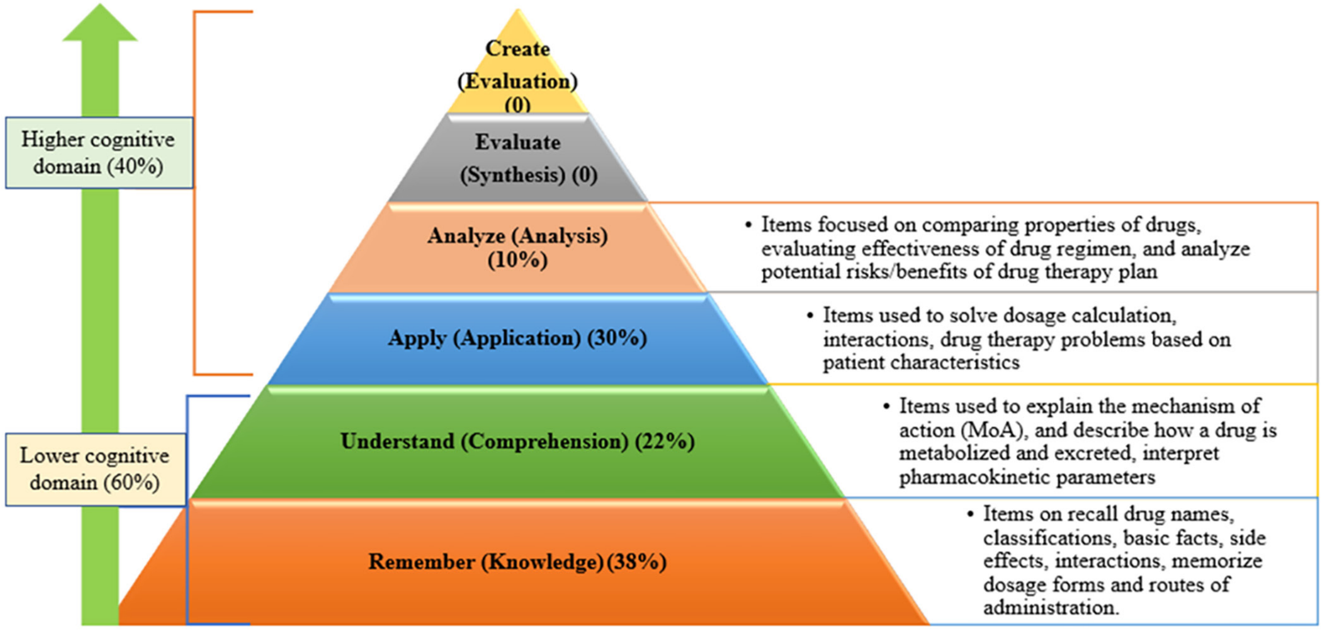

FIGURE 3 Distribution of pharmacology educator-created MCQs for addressing cognitive domains.

TABLE 3 The degree of rater agreement on MCQs pilot database measured with CCP and cognitive domains.

Percentage agreement and Inter-Rater Reliability (IRR) score					
Evaluation Category	**(%) Agreement**	**Weighted Kappa**	**Agreement level**	***Z*-score**	***p*-value**
MCQ quality in assessing the CCP, MCQs by groups and overall quality					
Address the CCP (1–30*)	50	−0.110	Poor	.505	
Address the CCP (31–60)	50	0.276	Fair	.111	
Address the CCP (61–90)	36.67	0.478	Moderate	.007	
Address the CCP (91–120)	30	0.011	Slight	.941	
Address the CCP (121–150)	30	0.174	Slight	1.150	
Address the CCP (151–180)	66.67	0.000	Slight	1.000	
Address the CCP (181–200)	45	0.275	Fair	.028	
Overall quality to assess CCP	44	0.254	Fair	3.61	<.001
MCQ efficiency in addressing cognitive domains, MCQs by groups and overall efficiency					
Cognitive Level (1–30*)	56.67	0.454	Moderate	.001	
Cognitive Level (31–60)	53.33	0.673	Substantial	<.001	
Cognitive Level (61–90)	43.33	0.244	Fair	.174	
Cognitive Level (91–120)	36.67	0.275	Fair	.019	
Cognitive Level (121–150)	20	0.325	Fair	3.300	
Cognitive Level (151–180)	60	0.777	Substantial	<.001	
Cognitive Level (181–200)	30	0.496	Moderate	.004	
Overall quality to address the cognitive domain	43.5	0.516	Moderate	7.53	<.001

Abbreviation: CCP, core concept of pharmacology.
*Number of items provided for evaluation to the first sub-group, same to other groups; Agreement level is categorized based on Kappa value.

conceptual understandings, and facilitate evidence-based instructional strategies. An innovative methodological approach blended four assessment instruments utilized to evaluate the quality of MCQs thoroughly. This initial step in inventory item bank preparation underscores their appropriateness for effectively assessing core concepts. Additionally, recognizing the importance of sharing resources with the global pharmacology education community, we plan to make the item bank available after the initial test and validation.

This study found that 60% of MCQs mostly followed the IWGs, however, 90% did not comply with at least one guideline. Over 20%

TABLE 4 Association and correlation analysis of quality assessment instruments attributes.

Test Attributes	Correlation test strength and directions	Item writing flaws (IWFs) ≤1	2–3	≥4	CCP test level Main/sub-concept	Minor/no concept	Cognitive domain Lower level	Higher level
Context-based format: Yes No	Chi-square Significance Strength/direction	$\chi^2=33.3$, df = 1 $p<.001$ Cramer's $V=.41$			$\chi^2=53.2$, df = 1 $p<.001$ Cramer's $V=.54$		$\chi^2=82.78$, df = 1 $p<.001$ Cramer's $V=.66$	
IWFs: ≤1 2–3 ≥4	Pearson Chi-square Significance Strength/direction				$\chi^2=34.7$, df = 1 $p<.001$ Cramer's $V=0.42$		$\chi^2=36.4$, df = 1 $p=.002$ Cramer's $V=0.43$	
CCP test level: Main/sub-concept Minor/no concept	Pearson Chi-square Significance Strength/direction						$\chi^2=24.9$, df = 1 $p<.001$ Cramer's $V=0.37$	

Abbreviations: CCP, core concept of pharmacology; IWFs, item writing flaws.

had a single non-compliant feature, while about 67% had two or more. Previous studies highlighted similar adherence challenges despite the benefit of IWGs.[16,37] Our non-adherent rates were lower than the 85% reported in the nursing education study,[38] while higher than other studies (e.g., 46%[17] and 76.7%[15]). Differences may stem from our MCQs being designed for various assessments by experts and students, unlike those in other studies, which might have been prepared for other purposes and from pre-existing databases. Furthermore, the varying levels of evidence for each guideline and open interpretation have led to inconsistent application by authors.[17,20] Acknowledging the importance of standardized MCQ quality and inconsistent utilization of existing guidelines, we compiled elements from previous literature to evaluate the educator-created MCQs.[16,17,20,25]

The most common guideline with which MCQ creators were non-compliant is a lack of order in MCQ options. The reason could be that some writers may overlook systematic ordering, opting for random arrangements since the key answer is randomized to prevent pattern guessing.[39,40] Hohensinn and colleagues' analysis confirmed that the location of the key answer has minimal effect on item difficulty.[39] Additionally, some suggested that the sequence does not matter if the correct options are evenly distributed.[40,41] The second most non-compliant guideline was the absence of the use of novel material to test higher-level learning followed by test stem structure awkwardness. Rephrasing teaching materials, like textbooks and course module language or expressions used during instruction while designing is crucial to enhance understanding by avoiding mere memorization.[20] Haladyna and Downing[42] emphasized that most educators prioritize teaching and measuring higher-level thinking, underscoring the importance of crafting content-adherent items while avoiding trivial construction. Additionally, employing IWGs in item construction can help modify awkwardly designed items and improve test quality for higher-order learning, which is widely accepted among testing specialists.[20]

To evaluate MCQ quality, we combined the percentage agreement and IRR. Percentage agreement is straightforward but overlooks chance agreement. IRR accounts for chance, enhancing reliability and robustness.[31,43,44] This integrated approach addresses chance agreement limitations, identifies ambiguous items, improves scoring consistency, and ensures test development integrity.[45–47]

T study revealed that 40% of MCQs addressed higher cognitive domains. This suggests that the MCQs were well-designed to evaluate students' clinical decision-making skills, including solving dosage calculations, assessing and formulating patient-specific drug therapy plans, comparing drug properties, evaluating drug regimen effectiveness, and analyzing the risks and benefits of drug therapy. Our finding is higher than the 7.9% reported for test bank questions from nursing assessments collected over five years[17] and the 28.3% reported for textbook test banks.[15] Designing test items targeting higher cognitive levels can improve the creation of discriminating items, as suggested by Rush.[16] This approach can contribute to strengthening the overall item quality and enriching the item bank for constructing PCI instruments. The remaining items were ranked as a lower cognitive domain focusing on recalling drug names, classifications, basic facts, side effects, and interactions.

Despite addressing various cognitive domains, none of the MCQs were designed to assess the domains of synthesis or evaluation. A known limitation of the MCQ format is its ineffectiveness in evaluating higher-level skills and in articulating ideas and thoughts. This limitation stems from MCQs' inherent focus on selecting correct answers from predetermined options.[14,15] Additionally, the items provided by experts were from available sets, with some prepared by students, meaning they were not specifically written for this study. Masters and colleagues pointed out that the MCQ format typically assesses only the lowest four cognitive domains.[15]

The MCQ pilot database underwent correlation analysis across quality assessment attributes, revealing that context-dependent MCQs significantly showed a strong positive correlation with other quality measuring criteria, such as addressing higher cognitive domains, assessing the main pharmacology concepts, and adhering to IWGs. Considering differing interpretations and applications of correlation coefficient results in scientific research,[36] the context-dependent item construction allows for designing high-quality MCQs. This may facilitate the creation of items capable of assessing higher-level thinking skills, and content-based knowledge, hence assessing the application and transfer of knowledge to real-world situations. This format serves as a valuable tool for evaluating complex clinical-based skills.[20,48,49] Likewise, MCQs are effective in evaluating fundamental pharmacology concepts and show enhanced efficacy in evaluating higher cognitive domains.

Consequently, rectifying MCQ design errors and standardizing item construction quality according to the analysis framework are essential steps to enhance item selection in the item bank repository for PCI instruments. Moreover, employing quality analysis instruments as a guiding framework can help develop high-quality MCQs. This initiative assists in developing resources for evaluating the concept-based learning approach.[1–3]

6 | LIMITATIONS AND FUTURE RESEARCH

This preliminary MCQ pilot database investigation relied on items submitted by a subset of international pharmacology educators, with additional contributions from undergraduate pharmacy students. This raises concerns about the representativeness of the sample. The submitted MCQs were designed for various purposes tailored to students across different degree-seeking programs, suggesting potential variations in the quality of item design, rather than being purpose-built to address pharmacology core concepts.

Consequently, the next step will involve both the creation of dedicated questions that focus on the core concepts, as well as revising, modifying, and conducting expert-based validation of top-scoring MCQs from this project according to the analysis framework. Items that adhere to guidelines, address higher cognitive domains, assess the CCP and are designed with a context-dependent structure will be prioritized for inclusion in the PCI instrument. Likewise, items that can be potentially revised and modified with expert reviews will also be considered. Experts will be extensively involved in item content validation for potential inclusion in the PCI test. This approach will assist in retaining MCQs that can identify student misconceptions, assess conceptual understanding, and create resources for evidence-based pharmacology education.

7 | CONCLUSION

The MCQ quality evaluation employed an analysis framework that blended four assessment instruments. Through this approach, global pharmacology content experts assessed the items, revealing variability in MCQ quality. While a majority of items tested the CCP, less than half effectively addressed higher cognitive levels, and often featured context-dependent structures. The framework successfully identified flawed MCQ structures, addressed common errors, and highlighted item development integrity. Additionally, items with context-dependent structures correlated with higher cognitive level testing, addressing CCP and improving adherence to guidelines.

The blended analysis revealed that MCQs effectively addressed higher cognitive domains and conveyed the core pharmacology course contents. Furthermore, the framework provides a systematic approach to develop clear, valid, and reliable test items, thereby contributing to the design of assessment tools for evaluating concept-based education. This structured format holds promise to create PCI test items, transforming pharmacology education by assessing content knowledge in realistic scenarios. Additionally, the analysis framework establishes the groundwork for constructing items for PCI development and concept-based curriculum resources, potentially enhancing the quality of MCQs, to assess critical thinking and practical knowledge application in pharmacology education.

AUTHOR CONTRIBUTIONS

Adeladlew Kassie Netere: Conceptualization; methodology; data curation; formal analysis; writing-original draft; writing-review and editing and project administration. **Tony Hughes:** Investigation; visualization; writing-review; project administration. **Anna-Marie Babey:** Investigation; writing-review; editing. **Martin Hawes:** Investigation; writing-review; editing. **Janet Mifsud:** Investigation; writing-review; editing. **John P Kelly:** Investigation; writing-review; editing. **Willmann Liang:** Investigation; writing-review; editing. **Mark Hernandez:** Investigation; writing-review; editing. **Kelly Karpa:** Investigation; writing-review; editing. **Hesham Al-Sallami:** Investigation; writing-review; editing. **Lynette B. Fernandes:** Investigation; writing-review; editing. **Patrik Aronsson:** Investigation; writing-review; editing. **Carolina Restini:** Investigation; writing-review; editing. **Fabiana Caetano Crowley:** Investigation; writing-review; editing. **Elvan Djouma:** Investigation; writing-review; editing. **Tina Hinton:** Investigation; writing-review; editing. **Johnson Liu:** Investigation; writing-review; editing. **Fatima Mraiche:** Investigation; writing-review; editing. **Paul J. White:** Conceptualization; methodology; visualization; writing-original draft; writing-review and editing, and project administration.

AFFILIATIONS

[1]Faculty of Pharmacy and Pharmaceutical Sciences, Monash University, Parkville, Victoria, Australia
[2]School of Science & Technology, University of New England, Armidale, New South Wales, Australia
[3]Department of Comparative Biomedical Sciences, School of Veterinary Medicine, University of Surrey, UK
[4]Department of Clinical Pharmacology and Therapeutics, Faculty of Medicine and Surgery, University of Malta, Msida, Malta
[5]Discipline of Pharmacology and Therapeutics, School of Medicine, University of Galway
[6]Department of Pharmacology and Pharmacy, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong
[7]East Tennessee State University, USA
[8]Department of Medical Education, East Tennessee State University Quillen College of Medicine
[9]School of Pharmacy, University of Otago, Dunedin, New Zealand
[10]School of Biomedical Sciences, University of Western Australia, Perth, WA, Australia
[11]Department of Pharmacology, Institute of Neuroscience and Physiology, The Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[12]Department of Pharmacology and Toxicology, College of Osteopathic Medicine, Michigan State University, USA
[13]Department of Physiology & Pharmacology, Schulich School of Medicine & Dentistry, Western University, London, Ontario, Canada
[14]Department of Microbiology, Anatomy, Physiology and Pharmacology, School of Agriculture, Biomedicine and Environment, La Trobe University, Melbourne, Victoria, Australia

[15]Sydney Pharmacy School, Faculty of Medicine and Health, The University of Sydney, New South Wales, Australia
[16]Department of Pharmacology, School of Biomedical Sciences, Faculty of Medicine and Health, University of New South Wales, Sydney, New South Wales, Australia
[17]Department of Pharmacology, Faculty of Medicine and Dentistry, College of Health Sciences, University of Alberta, Edmonton, Alberta, Canada

ACKNOWLEDGMENTS

The authors acknowledge the expert group members providing their expertise for creating and evaluating the items. Additionally, we acknowledge the undergraduate pharmacy students enrolled in the Faculty of Pharmacy and Pharmaceutical Sciences of Monash University for their contribution in providing items. Open access publishing facilitated by Monash University, as part of the Wiley - Monash University agreement via the Council of Australian University Librarians.

FUNDING INFORMATION

This research received no specific funding from any agency, organization, or institution.

CONFLICT OF INTEREST STATEMENT

The authors affirm that there are no competing interests that could affect the objectivity, impartiality, or accuracy of the research presented in this manuscript.

DATA AVAILABILITY STATEMENT

The data and materials supporting the findings of this study are available from the corresponding author upon reasonable request.

ETHICS STATEMENT

This study was approved by the Monash University Human Research Ethics Committee (MUHREC, approved protocol ID. 31379).

ORCID

Adeladlew Kassie Netere https://orcid.org/0000-0003-0519-509X
Anna-Marie Babey https://orcid.org/0000-0003-3454-5204
Martin Hawes https://orcid.org/0000-0003-4748-5482
Janet Mifsud https://orcid.org/0000-0002-7047-6827
Willmann Liang https://orcid.org/0000-0002-3260-8470
Hesham Al-Sallami https://orcid.org/0000-0002-0685-327X
Carolina Restini https://orcid.org/0000-0001-5158-485X
Fabiana Caetano Crowley https://orcid.org/0009-0001-9468-3361
Fatima Mraiche https://orcid.org/0000-0001-6188-1145
Paul J. White https://orcid.org/0000-0002-7524-1808

REFERENCES

1. White PJ, Guilding C, Angelo T, et al. Identifying the core concepts of pharmacology education: a global initiative. *Br J Pharmacol*. 2023;180(9):1197-1209.
2. Guilding C, Kelly-Laubscher R, Netere A, et al. Developing an international concept-based curriculum for pharmacology education: The promise of core concepts and concept inventories. *Br J Pharmacol*. 2024:1-9. doi:10.1111/bcp.15985
3. Guilding C, White PJ, Cunningham M, et al. Defining and unpacking the core concepts of pharmacology: a global initiative. *Br J Pharmacol*. 2024;181(3):375-392.
4. Herman GL, Loui MC, Zilles C. Creating the digital logic concept inventory. *Proceedings of the 41st ACM Technical Symposium on Computer Science Education*. 2010;102-106.
5. Smith MK, Wood WB, Knight JK. The genetics concept assessment: a new concept inventory for gauging student understanding of genetics. *CBE—Life Sciences Education*. 2008;7(4):422-430.
6. Furrow RE, Hsu JL. Concept inventories as a resource for teaching evolution. *Evol: Educ Outreach*. 2019;12:1-11.
7. Wright T, Hamilton S. Assessing student understanding in the molecular life sciences using a concept inventory. *ATN Assessment*. 2008;8:216-224.
8. Caceffo R, Frank-Bolton P, Souza R, Azevedo R. Identifying and validating Java misconceptions toward a CS1 concept inventory. *Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education* (Aberdeen, Scotland UK) (ITiCSE '19). Association for Computing Machinery. 2019;23-29. https://doi.org/10.1145/3304221.3319771
9. Brown GT, Abdulnabi HH. Evaluating the quality of higher education instructor-constructed multiple-choice tests: impact on student grades. *Frontiers in Education*. 2017;2:24.
10. Mehmood B, Sultan S, Azhar M, Shakoor A, Khan NB, Niazi Z. Quality analysis of multiple choice questions. *Pakistan Oral & Dent J*. 2020;40(4):236-239.
11. Bardar EM, Prather EE, Brecher K, Slater TF. Development and validation of the light and spectroscopy concept inventory. *Astron Educ Rev*. 2007;5(2):103-113.
12. Vonderheide A, Sunny C, Koenig K. Development of a concept inventory for the nursing general, organic and biochemistry course. *J STEM Education: Innov Res*. 2022;23(2):15-22.
13. Kerkman DD, Johnson AT. Challenging multiple-choice questions to engage critical thinking. *InSight: A J Scholarly Teach*. 2014;9:92-97.
14. Lopes A, Babo L, Azevedo J, Torres C. Multiple-Choice Tests-A Tool in the Assessing Knowledge. In: *Proceedings of INTED 2010 - 4th International Technology, Education and Development Conference*. 2010;256-265.
15. Masters JC, Hulsmeyer BS, Pike ME, Leichty K, Miller MT, Verst AL. Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. *J Nurs Educ*. 2001;40(1):25-32.
16. Rush BR, Rankin DC, White BJ. The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Med Educ*. 2016;16:1-10.
17. Tarrant M, Knierim A, Hayes SK, Ware J. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Educ Today*. 2006;26(8):662-671.
18. Reynolds CR, Livingston RB, Wilson V. *Measurement and Assessment in Education*. Upper Saddle River; 2009.
19. Kim M-K, Patel RA, Uchizono JA, Beck L. Incorporation of Bloom's taxonomy into multiple-choice examination questions for a pharmacotherapeutics course. *Am J Pharm Educ*. 2012;76(6):114.
20. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Meas Educ*. 2002;15(3):309-333.
21. Bloom B. *Taxonomy of Educational Objectives. Handbook 1 Cognitive Domain, White Plains, NY*. Longman; 1956.
22. Krathwohl DR. A revision of Bloom's taxonomy: an overview. *Theory Pract*. 2002;41(4):212-218.
23. Anderson LW, Krathwohl DR, Airasian PW, et al. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman; 2001.
24. Przymuszała P, Piotrowska K, Lipski D, Marciniak R, Cerbin-Koczorowska M. Guidelines on writing multiple choice questions:

a well-received and effective faculty development intervention. *SAGE Open*. 2020;10(3):2158244020947432.
25. Balaha MH, El-Ibiary MT, El-Dorf AA, El-Shewaikh SL, Balaha HM. Construction and writing flaws of the multiple-choice questions in the published test banks of obstetrics and gynecology: adoption, caution, or mitigation? *Avicenna Journal of Medicine*. 2022;12(3):138-147.
26. Wilson LO. Anderson and Krathwohl Bloom's taxonomy revised understanding the new version of Bloom's Taxonomy. *The Second Principle*. 2016;1(1):1-8.
27. Bonaci CG, Mustata RV, Ienciu A. Revisiting Bloom's taxonomy of educational objectives. *The Macrotheme Review A Multidisciplinary Journal of Global Macro Trends*. 2013;2(2):1-9.
28. Derbalah A, Al-Sallami H, Hasegawa C, Gulati A, Duffull SB. A framework for simplification of quantitative systems pharmacology models in clinical pharmacology. *Br J Clin Pharmacol*. 2022;88(4):1430-1440.
29. Lagunin AA, Romanova MA, Zadorozhny AD, et al. Comparison of quantitative and qualitative (Q) SAR models created for the prediction of Ki and IC50 values of antitarget inhibitors. *Front Pharmacol*. 2018;9:1136.
30. Gamer M, Lemon J, Gamer MM, Robinson A. Package 'irr', Various coefficients of interrater reliability and agreement, Version: 0.84.1. 2012;22:1–32.
31. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor Quant Methods Psychol*. 2012;8(1):23-34.
32. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med*. 2012;22(3):276-282.
33. Norman GR, Streiner DL. *Biostatistics: the Bare Essentials*. PMPH USA (BC Decker); 2008.
34. Graham M, Milanowski A, Miller J. Measuring and promoting interrater agreement of teacher and principal performance ratings. *Center for Educator Compensation Reform Online Submission*. 2012;1-33. https://files.eric.ed.gov/fulltext/ED532068.pdf
35. Figueroa A, Ghosh S, Aragon C. *Generalized Cohen's Kappa: A Novel Inter-Rater Reliability Metric for Non-mutually Exclusive Categories*. Springer Nature Switzerland; 2023:19-34.
36. Akoglu H. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*. 2018;18(3):91-93.
37. Downing SM. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Adv Health Sci Educ*. 2005;10(2):133-143. doi:10.1007/s10459-004-4019-5
38. Nedeau-Cayo R, Laughlin D, Rus L, Hall J. Assessment of item-writing flaws in multiple-choice questions. *J Nurses Prof Dev*. 2013;29(2):52-57. doi:10.1097/NND.0b013e318286c2f1
39. Hohensinn C, Baghaei P. Does the position of response options in multiple-choice tests matter? *Psicológica*. 2017;38(1):93-109.
40. Attali Y, Bar-Hillel M. Guess where: the position of correct answers in multiple-choice test items as a psychometric variable. *J Educ Meas*. 2003;40(2):109-128.
41. Jovanovska J. Designing effective multiple-choice questions for assessing learning outcomes. *Inf Dent*. 2018;18(1):25-42.
42. Haladyna TM, Downing SM. The validity of a taxonomy of multiple-choice item-writing rules. *Appl Meas Educ*. 1989b;1:51-78.
43. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20(1):37-46.
44. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971;76(5):378-382.
45. American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). *Standards for Educational and Psychological Testing*. American Educational Research Association; 2014 https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf
46. Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas*. 2013;50(1):1-73.
47. Messick S, Linn R. *Educational Measurement*. American Council on Education; 1989.
48. Haladyna TM. Context-dependent item sets. *Educ Meas Issues Pract*. 1992;11(1):21-25.
49. Haladyna TM. The effectiveness of several multiple-choice formats. *Appl Meas Educ*. 1992;5(1):73-88.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Netere AK, Hughes T, Babey A-M, et al. Evaluating the quality of multiple-choice question pilot database: A global educator-created tool for concept-based pharmacology learning. *Pharmacol Res Perspect*. 2024;12:e70004. doi:10.1002/prp2.70004