Review Article

Journal of Advances in Medical Education & Professionalism

A Systematic Review and Meta-Analysis on the Impact of Peer-Made MCQ Question-Bank Usage on Summative Assessments in Medical Education

ROBERT HWANG[1*], MSc, MBBS, BSc, DipHE, FHEA; MAX PALEY[2], MSc; DAVID BULL[3], MSc; OSAMA OMRANI[4], MSc; DIEGO VERGARA-JALANDONI[5], MSc; GERALD EGBURY[1], MSc

[1]Anglia Ruskin University, Department of Medical Education, Cambridge, United Kingdom; [2]Tunbridge Wells NHS Trust, United Kingdom; [3]West Middlesex NHS Foundation Trust, United Kingdom; [4]Addenbrookes NHS Foundation Trust, Cambridge, United Kingdom; [5]Ashford and St Peters NHS Trust, United Kingdom

Introduction: This systematic review and meta-analysis investigated the impact of utilising peer-generated multiple-choice question (MCQ) banks on the summative performance of undergraduate students studying medicine and allied subjects. Answering and writing peer-made MCQ questions are hypothesised to enhance learning through achievement of the domains of Bloom's taxonomy and thus summative examination performance.

Methods: A random-effects meta-analysis of correlation coefficients was conducted on six studies (n = 1,571) published between 2011 and 2021, drawn from MEDLINE, Scopus, Web of Science, PubMed, CENTRAL, and ERIC. The studies included undergraduate medical students from four countries. The risk of bias was assessed using the ROBINS-I tool.

Results: A weak positive correlation was found between answering peer-made MCQs and summative performance (Spearman's ρ = 0.22, 95% CI: 0.15 to 0.29, $p < 0.0001$), with a prediction interval of 0.00 to 0.42, indicating that in future studies, the effect of answering peer-made questions is likely beneficial or, at worst, neutral. A similar weak positive correlation was observed for writing peer-made MCQs (Spearman's ρ = 0.21, 95% CI: 0.09 to 0.32, $p < 0.0004$), though the prediction interval (-0.27 to 0.61) cannot exclude negative correlation between writing questions and summative performance in future studies. The findings suggest that answering and creating peer-generated MCQs positively influence exam performance. The modest correlations likely reflect confounding factors, such as prior academic performance and socio-economic background. This complicates isolating the impact of MCQ banks and may understate their true impact.

Conclusion: This study advocates for the integration of peer-generated MCQ banks into medical curricula, highlighting their potential as a cost-effective method to improve summative performance. Future research should focus on large-scale observational studies to better quantify these effects as well as controlling for confounding factors. The study underscores the value of peer engagement in learning and the utility of peer-made MCQ banks as educational tools.

Keywords: Academic performance; Curriculum; Medical; Students; Writing

**Corresponding author:*
Robert Hwang, MSc, MBBS, BSc, DipHE, FHEA;
Anglia Ruskin University, Department of Medical Education,
Cambridge, United Kingdom;
Tel: +44-7957329605
Email: robert.hwang1@nhs.net

Please cite this paper as:
Hwang R, Paley M, Bull D, Omrani O, Vergara-Jalandoni D, Egbury G. A Systematic Review and Meta-Analysis on the Impact of Peer-Made MCQ Question-Bank Usage on Summative Assessments in Medical Education. J Adv Med Educ Prof. 2025;13(4):259-269. DOI: 10.30476/jamp.2025.104780.2084.

Received: 13 November 2025
Accepted: 3 May 2025

Introduction

Medical education is traditionally delivered in the form of lectures, problem-based learning, and practical teaching. The theoretical component is often assessed using single best answer questions (SBAs) (1, 2). SBAs, a type of multiple-choice question, enable educationalists to cover a broad range of material through testing factual recall and, if well-written, can assess higher-order cognitive skills and critical thinking (3). Bloom's taxonomy is commonly applied to ensure high-quality SBAs, categorising questions by different levels of learning (2). This makes SBAs an effective method of assessment which supports learning. SBAs are also cost-effective and easy to mark, providing specific feedback and maintaining internal consistency (2, 4). These advantages could explain their frequent use in high-stakes exams like (5) the US Medical Licensing Examination, undergraduate medical schools, and the UK Royal Colleges' membership examinations (6).

A drawback of SBAs is that creating high-quality questions is resource intensive (7). With a limited pool of questions, medical schools often reuse them across years, preventing the release of past exams to preserve assessment validity. Therefore, students turn to commercial question banks for study (8-10). The popularity of these banks may stem from the candidates' need to bridge the gap between theoretical content and exam application.

Commercial question banks like "Passmedicine" or "Pastest" offer instantaneous access to thousands of practice questions, allowing students to efficiently prepare for exams (9, 11, 12). However, these banks may not always align with specific course curricula, and the association between their use and exam performance remains unclear (13, 14).

Alternatively, some students utilise peer-made question banks, where students create and contribute questions that better align with their curricula. This approach fits well with educational theory as creating questions engages the highest level of Bloom's taxonomy, whilst answering them involves analysis, application, understanding, and recall (15).

While research shows a positive correlation between answering peer-made questions and summative exam performance, the strength of this correlation varies (16-20). The impact of creating peer-made questions on exam performance is less clear (16, 20, 21).

In this context, the research question can be set: "What is the impact of both generating and answering peer-generated multiple-choice question banks on the summative assessment performance of undergraduate medical and allied students?"

The importance of this research question cannot be overstated. To date, no meta-analysis has been published investigating the impact of peer-made question banks. The ability to estimate their effect on summative performance could greatly benefit the international community. If a positive correlation is demonstrated, international medical education providers could implement a cost and resource effective intervention to improve the quality of healthcare professionals, with the potential to indirectly improve health outcomes for the international community.

The research question outlined above will be addressed throughout this manuscript, following standard scientific structure. The Introduction section will provide background information and context for the study. The Methods section will detail the research design and approach used to gather data. The Results section will present the findings, followed by a thorough analysis in the Discussion section. Finally, the Conclusion section will summarize the key insights and implications of the research, along with potential directions for future study.

Methods

Eligibility criteria

This systematic review and meta-analysis was not pre-registered but follows PRISMA 2020 guidelines. Studies were included regardless of publication date, randomization, blinding, or sample size. Grey literature and studies not written in English were excluded. The study population comprised undergraduate students studying medicine and allied subjects (e.g., biomedical sciences, pharmacology, veterinary medicine, dentistry).

The outcome measured was the student's final percentage score in summative MCQ examinations.

The inclusion criteria were as follows:

- Studies involving peer-made MCQ question banks – platforms where students author and answer multiple choice questions
- Studies quantifying the intervention by the number of questions answered, authored, or both
- Studies in which the population comprised undergraduate students studying medicine and allied subjects

The exclusion criteria were:

- Grey literature and studies not written in English
- Studies involving commercial question banks or peer-made question banks not in an

MCQ format

- Studies which had measured achievement in non-MCQ examinations, formative assessments, final scores that incorporated coursework
- Whole or parts of studies examining clinical content as this is traditionally assessed in a practical examination
- Post-graduate courses

The objective quantity and subjective quality of questions contained in each bank was intentionally omitted from the inclusion and exclusion criteria. This is because the review focussed on the processes of producing and answering peer-made MCQ and the associated measurable outcomes, rather than assessing the size and quality of the questions banks themselves.

Search strategy and study identification

The following databases were searched between 01/05/24 and 15/05/24:

1. MEDLINE
2. Ovid
3. Scopus
4. Web of Science
5. PubMed
6. Cochrane Central Register of Controlled Trials Issue 5, May 2024
7. ERIC

The following keywords were used to ensure the search strategy remained uniform: "Undergraduate AND Question Bank AND Summative Assessment AND Performance AND Peermade OR SBA OR MCQ". The flowchart of study selection is outlined in Figure 1.

Data extraction

From the studies that met the inclusion criteria, two reviewers (RH and MP) extracted the data independently with all discrepancies resolved by a third one (DB). Spearman's coefficient was directly extracted from all the literature except two – the result from Bottomley, et al. (2018) (16) was converted to Spearman's using provided data, and Guilding, et al. (2021) (17) kindly provided the raw data to allow for calculation of Spearman's without adjustment for prior performance.

Measurement and standardisation

The A Der Simonian and Laird random-effects meta-analysis without Hartung – Knapp adjustments was performed on the results of the systematic review using the metacor function in R statistical software version 4.4.1. A random-effects model was chosen as there was an anticipated heterogeneity between studies, leading to the probability that there is not one true effect size, but a distribution of effect sizes influenced by variance in intervention, population and methodology.

Data synthesis

Cochran's Q test, I^2 Statistic and Tau^2

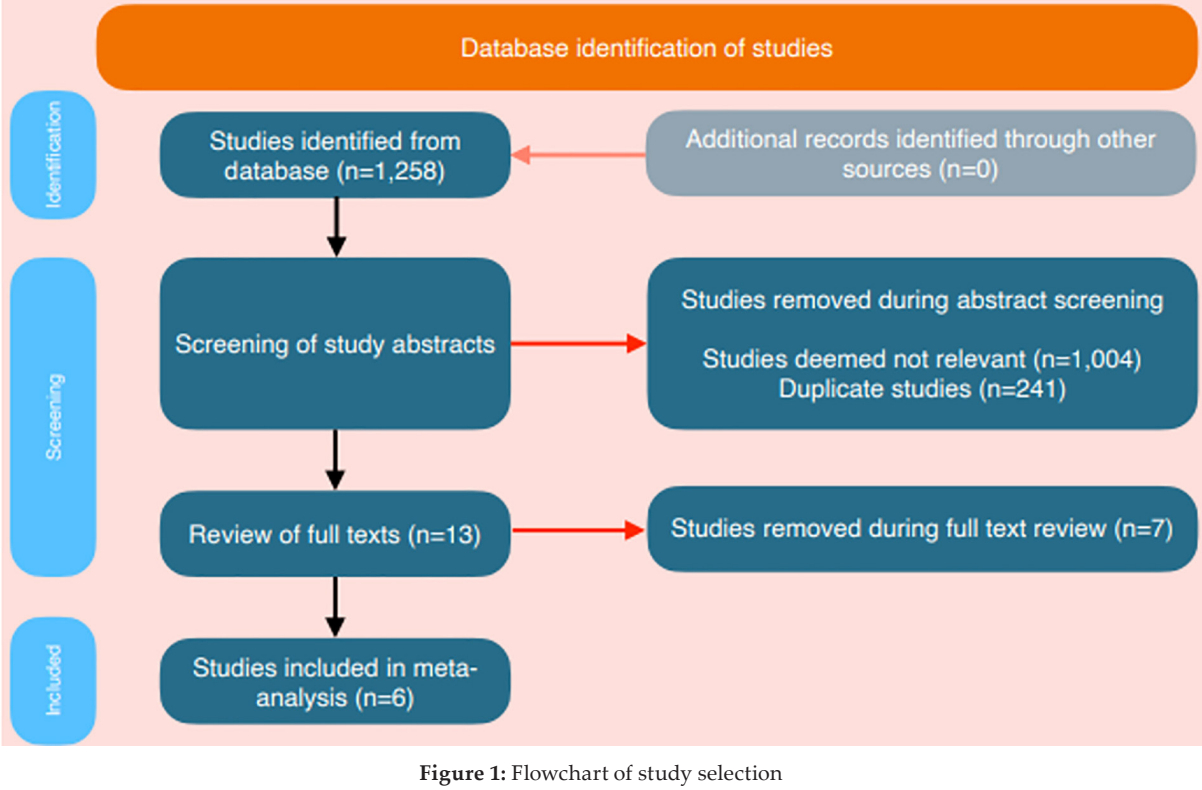

Figure 1: Flowchart of study selection

were calculated and interpreted to assess for heterogeneity and inconsistency. Funnel plots were generated to detect publication bias. Model robustness was checked through sensitivity analysis by performing influence diagnostics, including standardised residuals and Cook's distance, and leave-one-out analysis. Subgroup analysis or meta-regression, including Egger's test, was not performed as originally intended because $K < 10$, potentially culminating in misleading results (22, 23). The R code and data supporting this study are available from the corresponding author on reasonable request.

The risk of bias in the included studies was assessed using the Risk of Bias in Non-randomized Studies of Interventions (ROBINS-I) tool. This process involved evaluating potential sources of bias across seven domains: confounding, selection of participants, classification of interventions, deviations from intended interventions, missing data, measurement of outcomes, and selection of reported results. Each domain was rated as having a "low," "moderate," "serious," or "critical" risk of bias.

To ensure consistency and reliability, assessments were conducted independently by two reviewers, with discrepancies resolved through discussion or consultation with a third reviewer. Relevant information regarding study design, data sources, and potential biases was extracted to provide a comprehensive evaluation. The overall risk of bias for each study was determined based on the highest level of bias identified in any domain. This assessment informed the interpretation of the findings and the strength of the conclusions drawn from the included studies.

Results

Characteristics of the included studies

This systematic review and meta-analysis included 6 studies ($k = 6$, $n = 1,571$) from four countries (Australia, Malaysia, New Zealand and the United Kingdom). The studies were published between 2011 and 2021. The population comprised undergraduate students studying medicine, veterinary sciences, dentistry and biomedical sciences. The detailed characteristics of the included studies are summarised in Table 1. The risk of bias analysis revealed that one study was rated as having a low risk, while four had a moderate risk, and the remaining study was classified as having a high risk. The results of the ROBINS-I can be found in Figure 2.

Table 1: Characteristics of included studies

Study	Methods	Participants	Intervention	Outcomes
Bottomley, 2011	Mixed methods case study; n=107 participants were ranked according to academic performance and submitted surveys	2nd year Human Biology, Molecular Genetics, Biotechnology, and Laboratory Medicine students (Curtin University, New Zealand)	Requirement to write and review PeerWise questions for 10% of the semester mark (4% based on PeerWise score)	Analyzing the relationship between PeerWise activity and academic performance; participants' perceptions of the platform
Rhind and Pettigrew, 2012	Mixed methods educational study; assessing summative performance and participant questionnaires (n=442)	2nd and 3rd year veterinary medicine students	Introduction of PeerWise; n=264 were incentivized with a 2% course mark, n=178 received no incentive	Analyzing PeerWise engagement and examination performance; participants' views on the platform
Pathak and Mon, 2015	Quantitative educational study; assessing weekly academic performance (n=79)	1st year medical students (SEGi University College, Malaysia)	Quantitative grading of MCQs submitted to PeerWise	Comparing academic performance and MCQ quantitative grade
Walsh, 2018	Mixed methods educational study; analyzing summative performance and focus groups (n=603)	1st year medical students (Cardiff University)	A 1-hour session introducing students to PeerWise and asking them to write 1 question each	Comparing question writing frequency and summative performance; participants' perceptions of PeerWise
Nguyen et al., 2020	Mixed-methods educational study; analyzing summative performance and satisfaction surveys (n=254)	2nd year dental students (University of Sydney, Australia) studying neurology as a life science	Writing MCQs for peer review (n=174) every 2 weeks versus writing no MCQ questions (n=80)	Comparing summative performance and satisfaction between the two participant groups
Guilding, 2021	Mixed methods quantitative analysis; assessing summative performance and surveys (n=1,693)	4th year clinical sciences and pharmacology students (Newcastle University Medicine Malaysia)	Non-compulsory use of the PeerWise platform	Comparing summative performance and PeerWise use; participants' perceived benefits of PeerWise

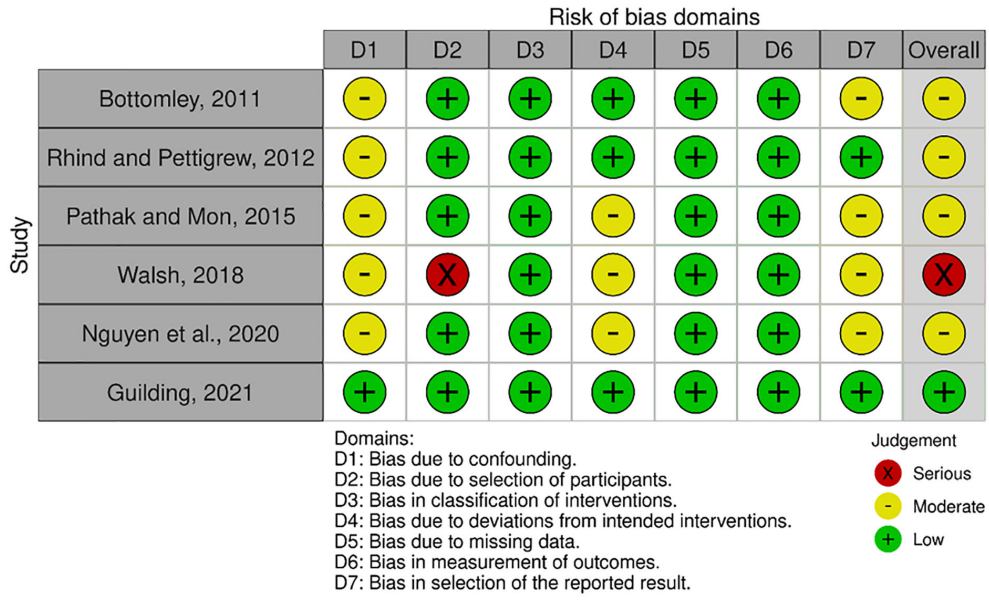

Figure 2: The output from the ROBINS-I tool

The impact of answering peer-made questions on summative performance

A total of k=5 studies analysed the impact of answering peer-made question banks on summative performance. The pooled correlation coefficient was 0.22 (95% CI: 0.15 to 0.29, p<0.0001), as shown in Figure 3. The Q-test revealed moderate to low heterogeneity of the outcomes (Q(4)=7.59, p<0.1078, Tau²=0.0034, I2=47.3%). A 95% prediction interval for the actual outcomes was 0.00 to 0.42. Calculation of the standardised residuals showed that Walsh, et al. (2018) had a value more significant than 2.00 and might be an outlier in this model. Cook's distances, including Walsh, et al. (2018) (20) (Cook's distance=0.5), demonstrated that none of the studies had a significant impact. A sensitivity analysis was performed excluding this study. The pooled correlation coefficient was 0.25 (95% CI: 0.19 to 0.31,p<0.0001), as shown in Figure 4. The Q-test revealed low heterogeneity of the outcomes (Q(3)=2.2, p<0.53, Tau²=0, I2=0%). A 95% prediction interval for the actual outcomes was 0.11 to 0.38. The sensitivity analysis, excluding Walsh, et al. (2018) (20) confirms the robustness of the findings and provides a more precise estimate, showing a slight increase in the pooled effect size and elimination of heterogeneity. Visual inspection of the funnel plot did not show asymmetry that would suggest publication bias (see Figure 5).

The impact of writing peer-made questions on summative performance

A total of k=4 studies analysed the impact of writing peer-made questions on summative performance. The pooled correlation coefficient was 0.21 (95% CI: 0.09 to 0.32, p<0.0004), as shown in Figure 6.

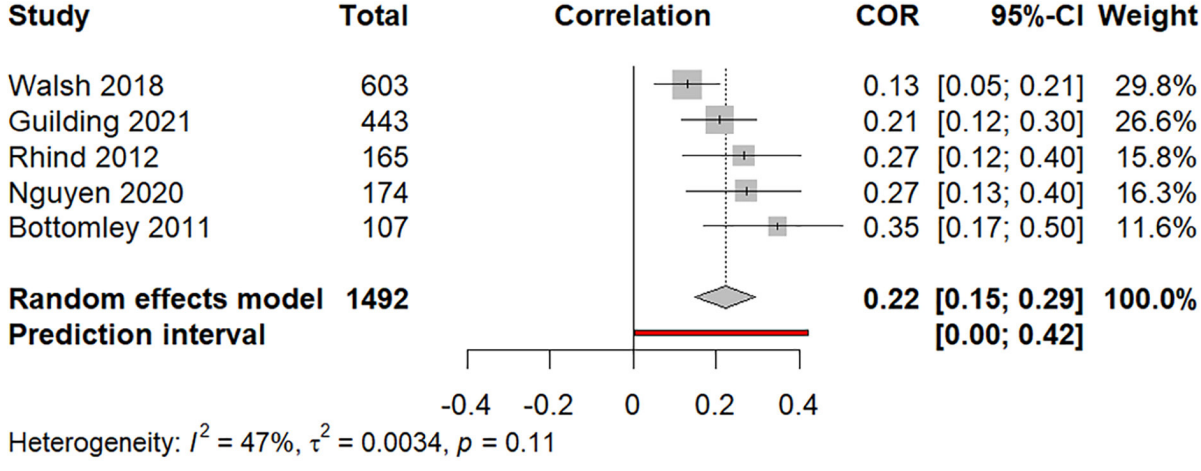

Figure 3: The forest plot showing meta-correlation of answering peer-made questions on summative performance

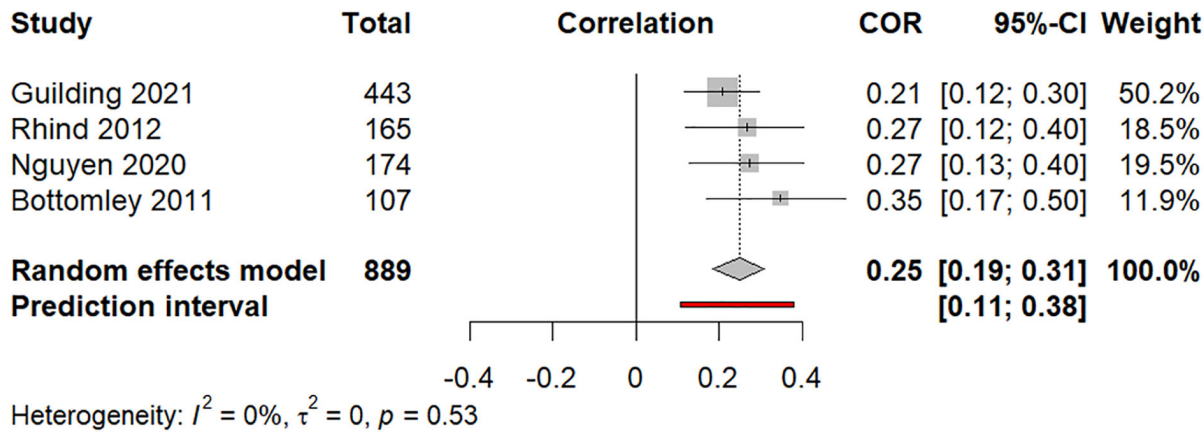

Figure 4: The forest plot for the sensitivity analysis of answering peer-made questions on summative performance removing Walsh et al. (2018)

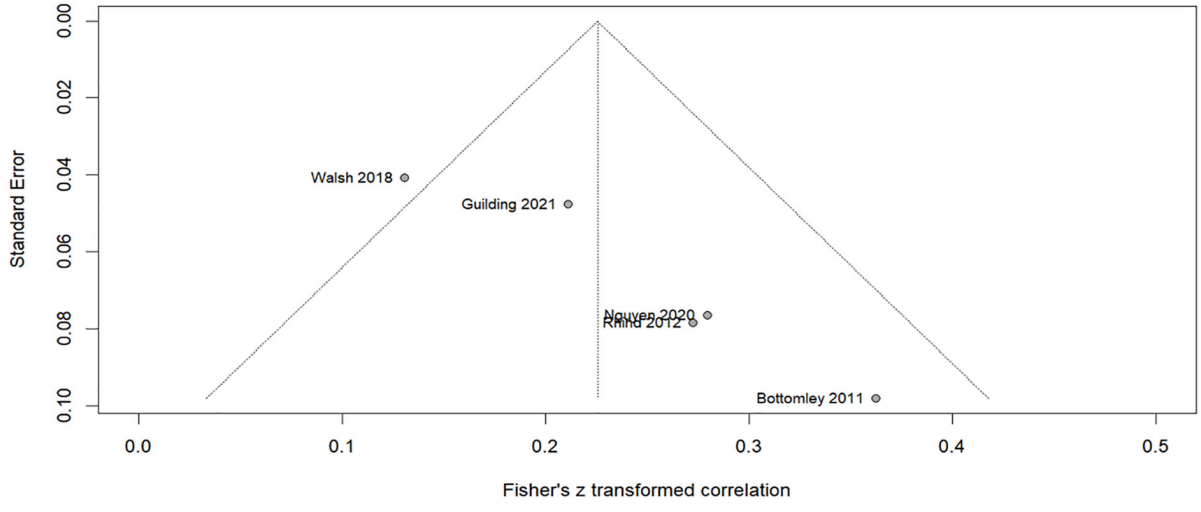

Figure 5: Funnel plot of the correlation between answering peer-made questions on summative performance

The Q-test revealed high heterogeneity of the outcomes (Q(3)=10.07 , p<0.180, Tau²=0.0094, I2=70.2%). A 95% prediction interval for the actual outcomes was -0.27 to 0.61. Calculation of the standardised residuals showed that no studies had a value more significant than +/- 2.00, and there were no outliers in this model. The Cook's distances were <1 for each study which demonstrates that none of the studies had a significant impact. Sensitivity analysis was performed using leave-one-out analysis. Removal of each study in turn did not significantly impact heterogeneity or correlation, see Figure 7. Visual inspection of the funnel plot did not show asymmetry as per Figure 8, indicating no publication bias. These findings collectively suggest that the meta-analysis results are robust.

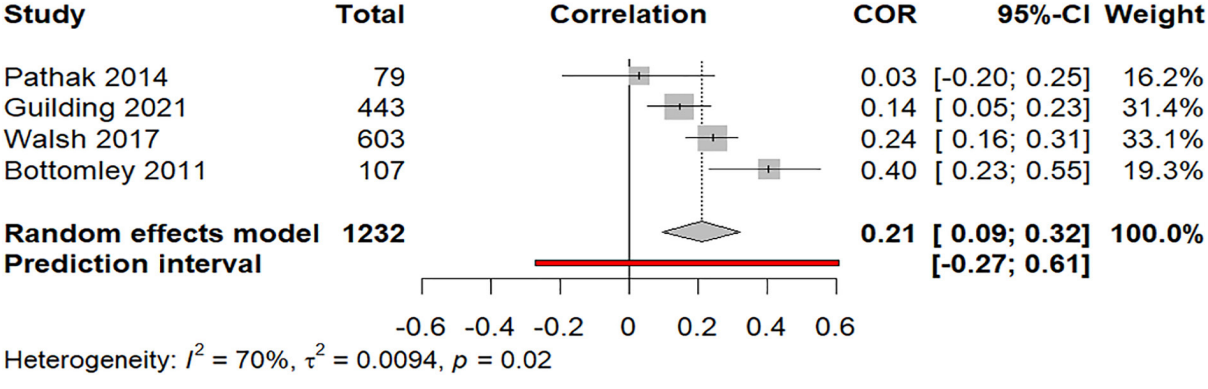

Figure 6: The forest plot showing meta-correlation of writing peer-made questions on summative performance

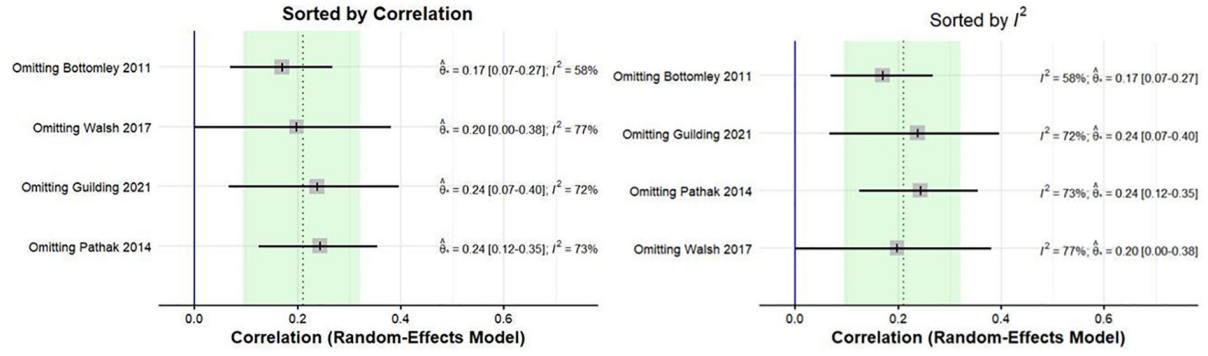

Figure 7: The Forest plots generated showing the impact of leave-one-out analysis on writing peer-made questions and summative performance sorted by correlation and I^2.

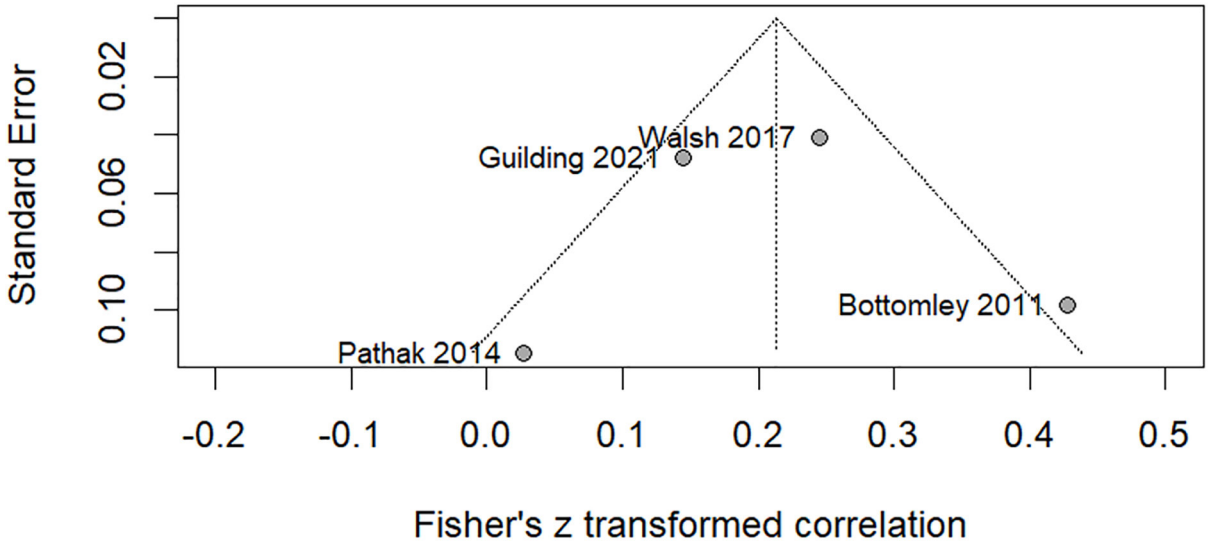

Figure 8: Funnel plot of the correlation between writing peer-made questions on summative performance

Discussion

The impact of answering peer-made questions on summative performance

Our results confirm a weak positive correlation (Spearman's coefficient=0.22, 95% CI:0.15 to 0.29, p<0.0001) between answering peer-made questions and summative examination performance. Based on the prediction interval, it is unlikely that negative effects would be observed in future studies. Previous studies agree on a positive correlation but vary on the precise estimate and cannot fully rule out a negative result (16-20). This aligns with education theory, as answering questions fulfils four of Bloom's taxonomy domains: analysis, application, understanding, and remembering, thus enhancing learning.

The true correlation is possibly estimated by the meta-analysis. Sensitivity analysis demonstrates that the highest source of heterogeneity can be attributed to Walsh, et al. (2018) (20), and removal of this study increases the coefficient to 0.25 (95% CI:0.19 to 0.31, p<0.0001) whilst reducing heterogeneity. Walsh, et al. (2018) (20) was assessed as having a high risk of bias overall, whereas the remainder have low to moderate risks of bias. A major concern in Walsh, et al. (2018) (20) came from potential bias in selection of participants. Participants were excluded if they did not contribute to questions, answers or comments after the initial induction lecture. This adds a potential confounding variable which may explain the source of heterogeneity. A possible explanation could be that students who did not contribute were less likely to perform well in the summative examination, thereby underestimating the correlation.

Another limitation is the failure to control for confounding factors like previous performance, demographics, and socio-economic background. Only Guilding, et al. (2021) (17) considered previous performance, but this was excluded for consistency in this analysis. Failure to control for previous performance likely contributes to underestimation of the correlation. Previous performance in summative assessments is an established predictor of future performance andis, therefore, an important confounder to adjust for when seeking to isolate the benefit of

question bank usage (25). This is consistent with Guilding et al. (2021) (17), who found a strong correlation between stage 3 and stage 4 scores.

The results indicate that peer-made question banks have a positive impact on students' formative examination performance, even with a possible underestimation as discussed above. As a student-implemented intervention, peer-made question banks are low cost and require minimal faculty intervention (17). When considering this in the context of their demonstrated effectiveness, they may pose an attractive and feasible option for medical education providers.

A major limitation of this meta-analysis is the reliance on Spearman's correlation, which confirms only a monotonic relationship and is less effective for non-linear relationships. While Spearman's can indeed confirm a positive monotonic correlation, as a correlation of ranking it is poor at quantitative determination of non-monotonic or non-linear relationships. It could be theorised that there is a non-linear relationship between the number of questions answered and examination score, with the additional benefit determined by the number of previously answered questions, following a sigmoidal relationship. A large scale prospective randomised trial would allow for thorough investigation and quantification of the benefit of this learning method. Alternatively, robust prospective observational studies account for other factors known to influence exam performance, such as previous performance, socio-economic background and standardised test scores (24-26). Proving and quantifying the curve would enable identification of the second point of minimal improvement, thereby providing educational providers and students a guide on the number of questions required to be attempted for maximum performance, in the most efficient manner. This, of course, would likely vary between curricula and examination faced.

The impact of writing peer-made questions on summative performance

Our results show a weak positive correlation (Spearman's co-efficient=0.21 (95% CI: 0.09 to 0.32, $p<0.0004$) between writing peer-made questions and summative performance. While the prediction interval is wide, the direction of effect remains positive. The correlation is similar to answering questions, which contradicts Bloom's taxonomy, where writing questions should indicate mastery and show a stronger correlation (15).

This result must be interpreted in the context of high heterogeneity among the included studies ($Q(3)=10.07$, $p=0.180$, $\tau^2=0.0094$, $I^2=70.2\%$). Sensitivity analysis did not identify a single study as a major source of heterogeneity; rather, all studies contributed substantially. This high heterogeneity may be due to the small number of included studies and the limited sample sizes within them. As a result, the overall effect estimate should be interpreted with caution, as high heterogeneity may reduce the reliability and generalizability of the findings.

Walsh, et al. (2018) (20) and Bottomley, et al. (2011) (16) reported a higher correlation coefficient for writing questions than answering, whilst Guilding, et al. (2021) (17) found the opposite. This may be explained by differences in methodology; Bottomley, et al. (2011) (16) motivated students by granting module marks from quality and quantity of questions submitted; Walsh, et al. (2018) (20) demonstrated the ease of writing questions and encouraged students to submit a question during the initial lecture; Guilding, et al. (2021) (17) introduced questions via email which may have reduced engagement; Pathak, et al. (2014) only examined the correlation between writing questions and summative examination performance. There is a significant deviation in Pathak's methodology (21) when comparing it to the other studies, which likely contributed to the overall heterogeneity – students were asked to write questions for an exam at the end of each week over a six-week period, yielding a correlation of 0.03. One week is a limited period for the intervention to take effect and may explain the reason why this correlation contradicts existing medical educational theory. Another possible explanation for the high heterogeneity is cultural and organizational differences. The four studies included are from four educational institutions in three countries. Differences in course teaching style and the impact of culture on the style of learning could account for a source of heterogeneity.

This meta-analysis suggests a positive correlation between writing peer-made question banks and summative performance though future research is needed to determine whether writing questions has a greater impact than answering. Whilst statistically significant, it is acknowledged that these correlations are relatively weak in magnitude, and, therefore, caution should be taken when considering the extent to which they may be impactful in practice. The limitations of this study include the inability to conclusively exclude future potential negative intervention effects and inability to quantify the correlation in writing peer-made questions as greater than answering. These limitations represent the current research gap in this area which should

be the topic of future studies. To address this, it is essential to concurrently study both the impact of writing and answering questions in a peer-made question bank. A large observational study could be set up accounting for potential confounders, including potential lack of student motivation in writing questions. Such a study could motivate students by offering prizes for contribution or module marks for quality and quantity of questions contributed.

Strengths

The strengths of this meta-analysis include a comprehensive literature search and a robust mathematical model. Adherence to the PRISMA 2020 guidelines demonstrates that this review was conducted with transparency and rigour, enhancing the credibility of the results. The inclusion of studies from various medical and allied health disciplines adds generalisability to the findings.

One of the main strengths is that the application of correlation meta-analyses as a research method in medical education is rare. The use of one here highlights its potential as a tool in advancing medical education research, providing more precise and reliable estimates of intervention effects.

Limitations

Many of the study limitations have been mentioned above, including those attributable to the topic's current research gap which is an area for further research. Whilst strictly adhering to the PRISMA 2020 guidelines, it was originally planned to register the review in PROSPERO. However, they only accept registrations that have a health-related outcome, and as this is a medical education study, it was rejected. Following this, the research team were unable to find an appropriate place to register the work. It is suggested that pre-registration should be strongly considered in future studies to enhance methodological transparency and reduce the risk of bias. Furthermore, allied medicine courses were included due to the low number of medicine-specific articles ($k<10$). This increased the generalisability but reduced the specificity of the findings as it pertains to undergraduate medical students. In addition, the low number of included studies limits the use of some meta-analysis tools. For example, sub-group analysis based on study characteristics including subject, country, and date of publications would have yielded more data for analysis. In addition, whilst the funnel plot did not show significant asymmetry the lack of further publication bias analysis, including Egger's test reduces the power in publication bias detection which may have led to an overestimation of the effect size.

Research implications, policy suggestions, and areas for future research

The strength of correlation between writing and answering peer-made MCQs varies across studies. This is likely due to differing student engagement strategies, intervention durations, and institutional and cultural learning styles. Furthermore, classroom-based exercises appear to foster greater engagement than online submission, whilst shorter interventions periods (e.g. six weeks) may be insufficient to allow for measurable learning benefits to develop.

Based on the above points, the following policy recommendations are posed:

- Encourage question writing through initiatives such as module marks, academic credits, or prizes for high-quality contributions.
- Promote structured and active learning, such as classroom-based exercise, to enhance student participation and knowledge retention.
- Conduct further research simultaneously exploring the impact of both writing and answering questions whilst attempting to control for confounding factors such as prior academic performance and student motivation.

Conclusions

This study confirms a positive correlation between writing and answering peer-made questions and summative performance, despite its limitations. Students who write and answer more peer-made questions tend to perform better in summative exams. Furthermore, it is possible that the degree of correlation is underestimated due to confounding factors. While negative effects from answering questions can be ruled out, the impact of writing questions remains unclear and cannot be conclusively shown to improve performance. Future large-scale observational studies are needed to examine both writing and answering questions, controlling for confounding factors.

These findings may be considered in guiding future educational policies, as the meta-analysis demonstrates the positive correlation between answering peer-made question banks and summative performance, the extent of which may have been underestimated. The implementation of peer-made question banks at medical university courses may improve cohort performance at a low financial and Time investment for faculty. Furthermore, at a macro-level, multiple universities could pool peer-made question banks for a common syllabus,

e.g. US medical schools for the USLME, potentially heightening the student's summative performance through access to more syllabus-specific questions.

Suggestion

- Concurrently study the impact of both writing and answering peer-made MCQs – for example a large observation study that accounts of potential confounders such as potential lack of student motivation in writing questions. Contribution prizes or module marks could be offered as motivation.

- Consider a large-scale prospective randomised trial to allow for thorough investigation and quantification of the benefits of MCQ writing and answering.

- Alternatively, consider a robust prospective observational study.

- Control for factors known to influence examination performance, including previous performance, socio-economic background, and standardised test scores.

- Aim to prove and quantify the curve to identify the second point of minimal improvement. This would help guide educational providers and students on the number of questions to be attempted to maximise performance in the most efficient manner.

- Consider pre-registering the review protocol through a platform such as IDESR or INPLASY to enhance methodological transparency and reduce the risk of bias.

Acknowledgements and Declarations

We would like to thank Guilding, et al. (17) for providing the raw data from their study. The authors declare that they have no competing interests nor received financial support for this project.

Authors' Contribution

All authors contributed to the discussion, read and approved the manuscript, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conflict of interest

The authors declare that they have no conflicts of interest.

Declaration on the use of AI

The authors of this manuscript declare that no artificial intelligence (AI) was used during the writing process.

References

1. Ali SH, Ruit KG. The impact of item flaws, testing at low cognitive level, and low distractor functioning on multiple-choice question quality. Perspect Med Educ. 2015;4:244251.
2. Thompson AR, O'Loughlin VD. The Blooming Anatomy Tool (BAT): A discipline-specific rubric for utilizing Bloom's taxonomy in the design and evaluation of assessments in the anatomical sciences. Anat Sci Educ. 2015;8:493501.
3. Zaidi NLB, Grob KL, Monrad SM, Kurtz JB, Tai A, Ahmed AZ, et al. Pushing Critical Thinking Skills With Multiple-Choice Questions: Does Bloom's Taxonomy Work? Acad Med. 2018;93(6):856-9.
4. Pugh D, De Champlain A, Touchie C. Plus ça change, plus c'est pareil: Making a continued case for the use of MCQs in medical education. Med Teach. 2019;41(5):569-77.
5. Heist BS, Gonzalo JD, Durning S, Torre D, Elnicki DM. Exploring Clinical Reasoning Strategies and Test-Taking Behaviors During Clinical Vignette Style Multiple-Choice Examinations: A Mixed Methods Study. J Grad Med Educ. 2014;6(4):709-14.
6. Sam AH, Westacott R, Gurnell M, Wilson R, Meeran K, Brown C. Comparing single-best-answer and very-short-answer questions for the assessment of applied medical knowledge in 20 UK medical schools: Cross-sectional study. BMJ Open. 2019;9(9):e032550.
7. Kilgour JM, Tayyaba S. An investigation into the optimal number of distractors in single-best answer exams. Adv in Health Sci Educ. 2016;21:571–85.
8. Tai J. Cheating the system- Tackling the problem at UCL: RUMS Review. [Internet]. 2019 [Accessed 3 May 2024]. Available from: https://rumsreview.com/2019/03/10/cheating-the-system-tackling-the-problem-at-ucl/.
9. Wynter L, Burgess A, Kalman E, Heron JE, Bleasel J. Medical students: what educational resources are they using?. BMC Med Educ. 2019;19:36.
10. Makus D, Kashyap A, Labib M, Humphrey-Murto S. A Curriculum Ignored? The Usage of Unofficial Commercial and Peer Learning Resources in Undergraduate Medical Education at a Canadian Medical School. Med Sci Educ. 2023;33(6):1379-88.
11. Passmedicine-Medical student finals/UKMLA resource [Internet]. 2024 [Accessed 15 July 2024]. Available from: https://www.passmedicine.com.
12. Passmedicine-Medical student finals/UKMLA resource: Revision [Internet]. 2024 [Accessed 15 Jul. 2024]. Available from: https://www.pastest.com.
13. Clemmons KR, Vuk J, Jarrett DM. Educational Videos Versus Question Banks: Maximizing Medical Student Performance on the United States Medical Licensing Examination Step 1 Exam. Cureus. 2023;15(4):e38110.
14. Kühbeck F, Berberat PO, Engelhardt S, Sarikas A. Correlation of online assessment parameters with summative exam performance in undergraduate medical education of pharmacology: a prospective cohort study. BMC Med Educ. 2019;19:412.
15. Bloom BS. Taxonomy of educational objectives: Cognitive and affective domains. New York: David McKay; 1956.
16. Bottomley S. Denny P. A participatory learning approach to biochemistry using student authored and

evaluated multiple-choice questions. Biochem Mol Biol Educ. 2011;39:352-61.
17. Guilding C, Pye RE, Butler S, Atkinson M, Field E. Answering questions in a co-created formative exam question bank improves summative exam performance, while students perceive benefits from answering, authoring, and peer discussion: A mixed methods analysis of PeerWise. Pharmacol Res Perspect. 2021;9:e00833.
18. Nguyen KA, Lucas C, Leadbeatter D. Student generation and peer review of examination questions in the dental curriculum: Enhancing student engagement and learning. Eur J Dent Educ. 2020;24:548–58.
19. Rhind SM. Pettigrew GW. 'Peer generation of multiple-choice questions: Student engagement and experiences'. Journal of Veterinary Medical Education. 2012;39(4):375–9.
20. Walsh JL, Harris BHL, Denny P, Smith P. Formative student-authored question bank: perceptions, question quality and association with summative performance. Postgrad Med J. 2018;94(1108):97-103.
21. Pathak R, Aye AM. The relationship between web-based MCQ authoring by students and their performance in medical Physiology. Indian Journal of Applied Research. 2015;5(10):1-2.
22. Wu A, Choi E, Diderich M, Shamim A, Rahhal Z, Mitchell M, et al. Internationalization of Medical Education - Motivations and Formats of Current Practices. Med Sci Educ. 2022;32(3):733-45.
23. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. "Basics of Meta-Analysis: I2 Is Not an Absolute Measure of Heterogeneity." Research Synthesis Methods. 2017;8(1):5–18.
24. Schwarzer G, Chemaitelly H, Abu-Raddad LJ, Rücker G. "Seriously Misleading Results Using Inverse of Freeman-Tukey Double Arcsine Transformation in Meta-Analysis of Single Proportions." Research Synthesis Methods. 2019;10(3):476–83.
25. Langensee L, Rumetshofer T, Mårtensson J. Interplay of socioeconomic status, cognition, and school performance in the ABCD sample. NPJ Sci Learn. 2024;9(1):17.
26. Geraci L, Kurpad N, Tirso R, Gray KN, Wang Y. Metacognitive errors in the classroom: The role of variability of past performance on exam prediction accuracy. Metacognition Learning. 2023;18:219–36.