# Assessment for growth: fostering student learning through assessment innovations in medical education

Wijk, E.V. van

**Part I: Very short answer question**

# Chapter 3

## Identifying academic success and underperformance: The discriminative power of very short answer questions and multiple-choice questions

Elise V. van Wijk
Floris M. van Blankenstein
B.N. Ruijter
J.H.T. Rohling
J. van der Kraan
F.W. Dekker
Alexandra M.J. Langers

*Submitted to Medical Science Educator (2025)*

## Abstract

**Background**: Multiple-choice questions (MCQs) are widely used in medical education, but are criticized for cueing and guessing. Very short answer questions (VSAQs), which require students to generate responses independently, may better assess knowledge. While VSAQs demonstrate higher item discrimination within individual exams, their effectiveness in distinguishing academic performance across multiple assessments remains unclear. This study examines whether VSAQs or MCQs more effectively distinguish students of varying performance levels across multiple summative examinations.

**Methods**: We analyzed retrospective data from six mixed-format examinations with VSAQs and MCQs of three cohorts of first- and second-year medical students. Academic performance was measured using grade point average (GPA) across assessments. Linear regression assessed the relationship of each question format with GPA, while ROC curves and C-statistics evaluated their ability to identify poor and excellent performing students (lowest and highest quintile of GPA).

**Results**: VSAQs showed higher item discrimination (Rir-values) than MCQs in all exams. VSAQs also had a stronger positive association with GPA compared to MCQs, and higher C-statistics, indicating superior discriminative ability.

**Conclusion**: VSAQs outperform MCQs in distinguishing academic performance levels across multiple assessments. Their integration into examinations enhances discriminative ability and may facilitate earlier identification of poor and excellent performing students, enabling targeted interventions and support of students.

## Introduction

Assessment of undergraduate medical students predominantly relies on multiple-choice questions (MCQs), especially in the single best answer (SBA) format, due to their high reliability and the efficiency of machine marking [1, 2]. However, MCQs have been criticized for their susceptibility to test-taking strategies, such as cueing – where students use cues in the question or answer options to deduce the correct answer without fully applying content knowledge – and guessing [3-6]. These factors introduce noise into MCQ scores, thereby diminishing their ability to accurately reflect students' true understanding [11]. In contrast, the very short answer question (VSAQ), an open-ended question requiring a concise answer, mitigates these issues by eliminating both cueing and guessing [7]. Consequently, VSAQs tend to exhibit higher item discrimination within individual examinations compared to MCQs [8-11], meaning that they can better differentiate between high- and low-performing students on a given test.

Although VSAQs consistently demonstrate superior item discrimination within single examinations, it remains unclear whether they also offer a better means of distinguishing among students with varying academic performance across multiple examinations. Interestingly, higher item discrimination in a single exam does not necessarily translate into a stronger ability to identify students as poor or excellent performers across several examinations. For example, Eijsvogels *et al.* [12] found that while extended matching questions (EMQs) demonstrate superior item discrimination compared to MCQs within examinations [13], they were less effective at identifying poor performing students, despite being better at identifying excellent performing students. Similarly, despite evidence supporting the higher item discrimination of VSAQs compared to MCQs within individual examinations [8-11], their effectiveness in distinguishing overall student academic performance across multiple assessments remains uncertain.

Beyond evaluating the item discrimination of individual assessment questions, it is important to investigate whether certain question formats are better suited for identifying poor and excellent performing students across multiple examinations. This broader perspective can offer insights into the consistency and robustness of these formats in distinguishing students across different performance levels, and formats with higher discriminative power may facilitate the early identification of underperforming students, thereby enabling timely interventions. In this study, we aim to 1) examine the relationship between question format (VSAQs *versus* MCQs) and academic performance; 2) evaluate the ability of VSAQs and MCQs to identify poor and excellent performing students. To address these aims, we first assess the item discrimination of both question formats within each examination, thereby verifying the assumption that VSAQs have superior discriminative ability within examinations [8-11]. We use the VSAQ- and MCQ-scores from two summative mixed-format examinations administered during the first and second year of an undergraduate medical curriculum. Our analysis includes two student populations: 1) all students who participated in the first-year examination, including those who may later leave the program, and 2) nominal students who participated in both the first- and second-year examinations. The first population offers a broader performance range, particularly among lower performing students, while the second population offers more datapoints (i.e., questions) per student, enhancing the reliability of the analysis.

# Methods

## Setting

This retrospective cohort study was conducted in Leiden University Medical Center (LUMC), the Netherlands. The Dutch medical curriculum comprises a three-year bachelor's program followed by a three-year master's program. Although the bachelor courses are primarily assessed with MCQs, other formats such as extended matching, comprehensive integrated puzzle, open essay, and VSAQs are also included in the written assessments. Additionally, students participate in longitudinal training on various CanMEDS competencies [14] beyond the role of Medical Expert, such as communication skills, leadership, health promotion, and collaboration. To assess the discriminative ability of VSAQs and MCQs we analyzed summative examinations of two medical bachelor's courses: '*Regulation and Metabolism*' (RM), a first-year fundamental course, and '*Diseases of the Abdomen*' (DA), a second-year clinical course. These courses (6 and 7 weeks, respectively) address metabolic and gastrointestinal topics. In our prior study [8], we compared the psychometric properties of MCQs and VSAQs in formative assessments for both courses. Subsequently, the course coordinators added VSAQs to the previously MCQ-only summative exams, resulting in mixed-format examinations. This mixed format has now been used in both courses for three consecutive years (student cohorts 2020-2021; 2021-2022; 2022-2023), resulting in six summative mixed-format examinations available for our analyses. Near the end of the course and prior to the summative assessment, students were given opportunities to practice with the different question types. All assessments were administered digitally through RemindoToets (Paragin) system [15] and yielded study credits.

## Participants

The first population, *RM participants*, included all first-year bachelor medical students from the 2020–2021, 2021–2022, and 2022–2023 cohorts who completed the first sitting of the summative RM assessment. The second population, *RM & DA participants*, comprised all second-year bachelor medical students from the same cohorts who completed the first sitting of both the RM and DA summative assessments in consecutive years. Students who attempted less than 75% of all exams taken into consideration for the GPA calculation were excluded from both populations ($n$=47).

## Study design and data collection

We analyzed data from two summative mixed-format examinations from the RM and DA courses, administered between 2021 and 2024. For each assessment, we extracted individual questions, question formats, question scores (coded as 1 for correct and 0 for incorrect), residual item reliability-values (Rir) for each question, and student IDs from the Remindo assessment system. The Rir-value, representing the correlation between one test item and all other items of the test, measures item discrimination by indicating how well each question correlates with overall student performance on that test (after excluding that specific question) [16]. To assess the comparability of question formats within the examinations, we examined the distribution of each question format across the different themes and cognitive process dimensions based on Bloom's revised taxonomy [17]. Bloom's levels were assessed by two researchers of the research team (EvW, FvB). First, a sample of 50 questions, randomly selected from across the examinations, was reviewed together to establish a shared coding approach. Following this, each researcher coded a subset of 30 questions independently, with any discrepancies discussed to reach

consensus (Cohen's kappa = 0.89 before discussion). The remaining questions were then evaluated by the principal investigator (EvW), with consultation from the second researcher (FvB) in cases of uncertainty.

We measured academic performance with the grade point average (GPA), a widely recognized and standardized indicator of academic success [18, 19]. All exam grades from the first and second-year courses were retrieved from the university's administrative system. These numeric grades, combined with the corresponding study credits, were used to calculate students' GPA. The maximum score is 10, and grades of 5.5 or higher are considered sufficient to earn study credits. Failing grades (defined as <5.5) were included to ensure a realistic measurement of academic performance. We evaluated GPA based on two approaches: one using the most recent grade from the last sitting (including retake exams) and another using only the grade from the first sitting. For the first population (*RM participants*) we calculated the GPA based on first-year courses, while for the second population (*RM & DA participants*) the GPA included courses from both the first- and second-year. We included only exam grades. In courses where non-exam components such as presentations, reports, and participation, contributed to the final grade but did not result in a separate grade, these components were excluded. To ensure a fair representation of these courses, we allocated half of the available study credits to the exam grades, reflecting their partial contribution to the overall course assessment. Based on these GPAs, students were categorized by quintiles into '*poor performing*' (first quintile), '*average performing*' (second through fourth quintile), or '*excellent performing*' (fifth quintile) students.



## Data analysis

Descriptive statistics were calculated and presented for each examination, including total average scores, average VSAQ and MCQ scores, the distribution of VSAQs and MCQs across the themes, and the distribution of Bloom's cognitive dimensions assessed by VSAQs and MCQs. To compare the different question formats, we calculated separate z-scores for VSAQs and MCQs based on their respective absolute scores. For the *RM participants*, z-scores were derived from their VSAQ and MCQ scores on the RM assessment. For the *RM & DA participants*, z-scores were calculated separately for VSAQs and MCQs using absolute scores from both the RM and DA assessments, ensuring that each student had one VSAQ z-score and one MCQ z-score reflecting their performance across both exams. Item discrimination was determined using the mean of the Rir-values for each question format [16]. The Rir-value for a each question was calculated in relation to the entire exam.

Linear regression analysis was performed with GPA as linear outcome variable and the z-scores of the VSAQ and MCQ as covariates, separately for each cohort and across all cohorts. First, univariate regression analyses were conducted to assess the variance explained by each question format individually (reported as adjusted R2). Next, both formats were included in a multivariate model to account for shared variance and determine their relative predictive value. Student GPAs from the courses of the first year (RM participants) and the first two year (*RM & DA participants*) were used as outcome. Similarly, the same student GPAs were used to create new categorical variables: '*poor performing*' versus '*other*' (average and excellent) and '*excellent performing*' versus '*other*' (average and poor). Here, '*poor performing*' students were defined as those in the lowest quintile, '*average performing*' in the second through fourth quintiles, and '*excellent performing*' in the fifth quintile. Receiver Operating Characteristic (ROC) curves were used to evaluate the ability of MCQs and VSAQs to distinguish between poor and non-poor performing students,

and also between excellent and non-excellent performing students. The Concordance-statistic (C-statistic or area under the ROC curve) was calculated as a measure of discriminatory power, reflecting how well these formats differentiate between performance levels. Significant differences between the C-statics of VSAQs and MCQs were assessed using paired bootstrapping with 95% confidence intervals. All statistical analyses were performed using R version 4.1.0 (R Foundation for Statistical Computing, Vienna, Austria).

### Ethical approval

This study received ethical approval from the LUMC Educational Research Review Board (OEC/ERRB/20241008/1). Informed consent was not required, as no new participants were recruited for this study. The data analyzed were retrospective and already available from the assessment system as part of the students' regular curriculum. We collected no additional demographic data, and all data were aggregated to ensure that individual students could not be identified.

## Results

### Descriptives and item discrimination

The students included in our study for each mixed-format examination, along with the average total, VSAQ, and MCQ scores are presented in *Table 1*. In all mixed-format examinations, the average Rir-values of the VSAQs were consistently higher compared to the MCQs (*Table 1*). The distribution of question formats across the themes was generally even, with no themes exclusively assessed by a single format (*Appendix 1 – Supplemental Table 1*). Across Bloom's taxonomy levels, the distribution within exams was largely balanced, except in RM, where VSAQs appeared more frequently classified as '*Remember*' questions and less as '*Understand*' questions (*Appendix 2 – Supplemental Table 2*).

**Table 1.** The scores and Rir-values of VSAQs and MCQs within the mixed-format examinations.

| | Total examination | | Very short answer questions | | | Multiple-choice questions | | |
|---|---|---|---|---|---|---|---|---|
| | Max points | Total score[b] mean (SD) | N | Score, mean (SD) | Rir-values, mean (SD) | N | Score, mean (SD) | Rir-values, mean (SD) |
| RM 2021[a] (*n*=323) | 74.50 | 53.40 (9.19) | 42 | 27.95 (6.36) | 0.30 (0.13) | 32 | 24.50 (3.53) | 0.21 (0.11) |
| RM 2022 (*n*=296) | 71 | 45.49 (8.79) | 29 | 15.53 (4.63) | 0.28 (0.11) | 36 | 27.29 (4.61) | 0.24 (0.10) |
| RM 2023 (*n*=287) | 73 | 49.96 (9.74) | 31 | 18.97 (5.40) | 0.32 (0.12) | 38 | 29.29 (4.79) | 0.24 (0.11) |
| DA 2022 (*n*=282) | 91 | 59.58 (10.35) | 22 | 13.39 (3.41) | 0.30 (0.11) | 48 | 33.34 (4.53) | 0.18 (0.10) |
| DA 2023 (*n*=255) | 90 | 64.38 (9.48) | 30 | 20.75 (4.14) | 0.26 (0.09) | 40 | 28.50 (3.99) | 0.19 (0.11) |
| DA 2024 (*n*=260) | 71 | 48.55 (8.18) | 31 | 20.62 (4.18) | 0.27 (0.13) | 20 | 13.66 (2.64) | 0.19 (0.11) |

RM = Regulation and Metabolism; DA = Diseases of the Abdomen; N = number of questions.

[a]RM 2021 & DA 2022 = cohort 2020-2021; RM 2022 & DA 2023 = cohort 2021-2022; RM 2023 & DA 2024 = cohort 2022-2023.

[b]Average total score also includes the scores on question formats other than VSAQs and MCQs (comprehensive-integrated-puzzle (CIP) (maximum points: 4), open-ended question without word limitation (maximum points: 8), 6-step pharmacology question (maximum points: 8), hotspot question (maximum points: 1)).

## Relationship between question format and academic performance

We conducted a linear regression analysis to examine the effects of VSAQ and MCQ z-scores on GPA. First, we assessed the variance explained by each question format separately using univariate models ($R^2$ values). Among all *RM participants*, the $R^2$ values ranged from 0.62 to 0.66 for VSAQ z-score and from 0.44 to 0.57 for MCQ z-score. In all *RM & DA participants*, the $R^2$ values ranged from 0.68 to 0.71 for VSAQ z-score and from 0.59 to 0.67 for MCQ z-score.

The multiple regression analyses showed a significant positive association between both VSAQ and MCQ z-scores and GPA (*Table 2*). However, for all *RM participants* (*n*=906), the VSAQ z-score had a stronger association with GPA when using first sitting grades (β = 0.66, t(903) = 21.55, p < .001, 95% CI [0.60, 0.72]) than the MCQ z-score (β = 0.37, t(903) = 12.20, p < .001, 95% CI [0.31, 0.42]), with the model explaining 75% of the variance in GPA (F(2, 903) = 942.44, p < .001, adjusted $R^2$ = .75). Similarly, in all *RM & DA participants* (*n*=797), the VSAQ z-score had a stronger positive association (β = 0.54, t(794) = 22.23, p < .001, 95% CI [0.49, 0.59]) compared to the MCQ z-score (β = 0.37, t(794) = 15.03, p < .001, 95% CI [0.32, 0.41]), with the model accounting for 68% of the variance (F(2, 794) = 1202.38, p < .001, adjusted $R^2$ = .68). These results were similar when using last sitting grades and for the analyses of the separate cohorts (*Table 2*).

**Table 2.** Linear regression model parameters with grade point average as outcome based on the first and last sitting grades.

| | First sitting grades | | | | | | Last sitting grades | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | VSAQ z-score | | | MCQ z-score | | | VSAQ z-score | | | MCQ z-score | | |
| **Cohort** | **ß** | **t** | **95% CI** | **ß** | **t** | **95% CI** | **ß** | **t** | **95% CI** | **ß** | **t** | **95% CI** |
| *RM participants* | | | | | | | | | | | | |
| 2020-2021[a] (*n*=323) | 0.68 | 14.60 | 0.59–0.77 | 0.26 | 5.65 | 0.17–0.35 | 0.61 | 15.08 | 0.53–0.69 | 0.21 | 5.16 | 0.13–0.29 |
| 2021-2022 (*n*=296) | 0.62 | 12.10 | 0.52–0.72 | 0.48 | 9.35 | 0.38–0.58 | 0.49 | 11.32 | 0.40–0.57 | 0.39 | 9.09 | 0.31–0.47 |
| 2022-2023 (*n*=287) | 0.67 | 12.18 | 0.57–0.78 | 0.39 | 7.12 | 0.29–0.50 | 0.56 | 11.42 | 0.47–0.66 | 0.33 | 6.75 | 0.24–0.43 |
| Total (*n*=906) | 0.66 | 21.55 | 0.60–0.72 | 0.37 | 12.20 | 0.31–0.43 | 0.56 | 21.31 | 0.51–0.61 | 0.30 | 11.66 | 0.25–0.36 |
| *RM & DA participants* | | | | | | | | | | | | |
| 2020-2021 (*n*=282) | 0.52 | 13.44 | 0.44–0.59 | 0.35 | 9.20 | 0.28–0.43 | 0.43 | 12.25 | 0.36–0.50 | 0.31 | 8.65 | 0.24–0.38 |
| 2021-2022 (*n*=255) | 0.49 | 11.86 | 0.41–0.57 | 0.40 | 9.60 | 0.31–0.48 | 0.39 | 10.49 | 0.31–0.46 | 0.35 | 9.59 | 0.28–0.43 |
| 2022-2023 (*n*=260) | 0.61 | 13.69 | 0.52–0.70 | 0.36 | 7.94 | 0.27–0.44 | 0.53 | 13.57 | 0.45–0.61 | 0.29 | 7.52 | 0.22–0.37 |
| Total (*n*=797) | 0.54 | 22.23 | 0.49–0.59 | 0.37 | 15.03 | 0.32–0.41 | 0.45 | 20.74 | 0.41–0.49 | 0.31 | 14.44 | 0.27–0.36 |

RM=Regulation and Metabolism; DA = Diseases of the Abdomen; VSAQ=very short answer question; MCQ=multiple-choice question; CI=confidence interval. SE = standard error. All p-values <0.001.

[a]Cohort 2020-2021 = RM 2021 & DA 2022; Cohort 2021-2022 = RM 2022 & DA 2023; Cohort 2022-2023 = RM 2023 & DA 2024.

### Identification of poor and excellent performing students

We calculated the C-statistic to assess the ability of VSAQ and MCQ scores to identify poor performing (lowest GPA quintile) and excellent performing (highest GPA quintile) students. In all *RM participants* (*n*=906), VSAQ z-scores showed higher C-statistics (i.e., greater discriminative ability) for both poor and excellent performing students compared to MCQ z-scores (*Figure 1A*). However, for poor performing students using first sitting grades, the difference between the C-statistics of VSAQ and MCQ was not significant. In all *RM & DA participants* (*n*=797), the C-statistic of the VSAQ z-score remained significantly higher than that of the MCQ z-score for poor performing students. For excellent performing students, there was no significant difference in discriminative ability (*Figure 1B*). The findings across separate cohort analyses followed the same pattern; either the C-statistic for the VSAQ z-score was significantly higher than that of the MCQ z-score, or there was no significant difference between them (*Appendix 3 & 4 – Supplemental Figure 1 & 2*).

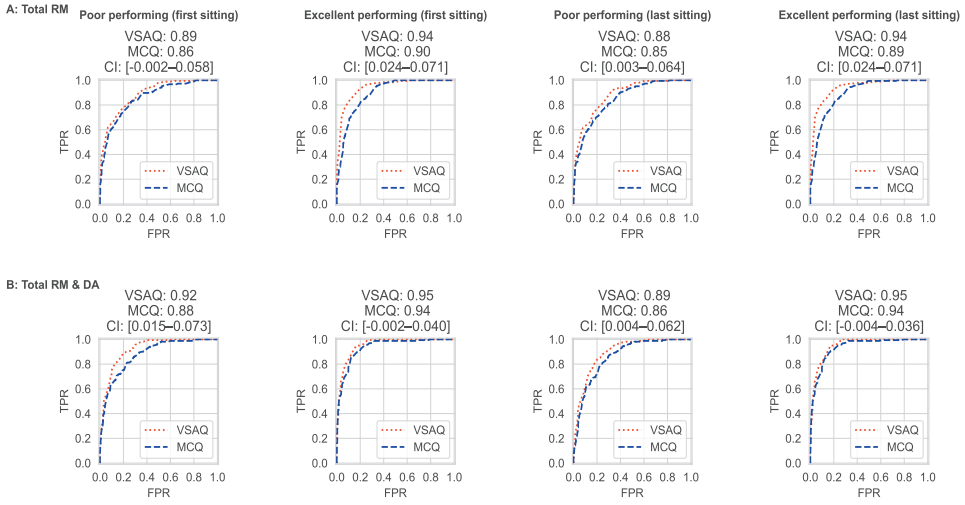



**Figure 1.** Receiver Operating Characteristic (ROC) curves for both poor and excellent performing students using the first sitting or last sitting grades from A) the total RM participants, and B) the total RM & DA participants. Red = VSAQs; Blue = MCQs; CI = 95% confidence interval; TPR = true positive rate; FPR = false positive rate. C-statistics are shown above each graph, which were calculated from the area under the ROC curve.

## Discussion

In this study, we examined whether VSAQs or MCQs are more effective in distinguishing undergraduate medical students with varying academic performance levels based on GPA. Our findings reveal that while both question formats are suitable for distinguishing students across performance levels, VSAQs consistently demonstrate superior effectiveness compared to MCQs in both the linear regression analysis

and ROC curves. Notably, the difference in discriminative ability between VSAQs and MCQs was slightly more pronounced among first-year students (*RM participants*) compared to the combined group of first- and second-year students (*RM & DA participants*). A plausible explanation for this observation is the exclusion of mostly poor performing students who did not nominally progress to the second year, resulting in a more homogeneous group of students with less variability in academic performance. Consistent with previous research [8-11], VSAQs also demonstrated higher item discrimination than MCQs within the individual examinations.



Our findings differ from the findings of Eijsvogels *et al.*, who found that MCQs were more effective than EMQs in identifying poor performing students [12]. This discrepancy may arise from differences in question format and scoring methodology. While EMQs share certain features with VSAQs — such as reduced reliance on guessing compared to MCQs — they still provide a list of plausible options that could lead to guessing. In contrast, VSAQs require students to generate concise responses without external cues, inherently minimizing the likelihood of guessing and offering a more direct measure of their knowledge. Furthermore, Eijsvogels *et al.* [12] penalized incorrect answers on MCQs, potentially increasing the effectiveness of this question format. By normalizing performance scores within mixed-format examinations, we ensured a fair comparison, eliminating potential biases from variations in exam quality or instructional approach. Additionally, our analysis included multiple examinations and a larger sample size, enhancing the robustness of our findings.

The GPA used in this study to assess academic performance is primarily based on assessments that predominantly consist of MCQs, which could bias the results toward MCQs due to their alignment with the format used to calculate GPA. However, despite this potential bias, our findings indicate a clear advantage for VSAQs over MCQs in distinguishing poor and excellent performing students. The superior performance of VSAQs can be attributed to their format, which requires students to generate responses independently, eliminating opportunities for guessing and relying instead on their actual understanding [3-6]. This characteristic makes VSAQs a more accurate reflection of student's knowledge level. While guessing may inflate MCQ scores for poor performing students in a single examination, this effect will diminish when performance is averaged across multiple exams, thereby reducing noise introduced by guessing.

## Strengths and limitations

This study is the first to evaluate the discriminative ability of VSAQs across multiple assessments using GPA as a measure of academic performance, rather than focusing exclusively on item discrimination with individual exams. By utilizing real-world summative assessment data from three distinct student cohorts, we minimized the biases associated with voluntary participation and ensured a robust, and representative sample size. This also allowed us to aggregate results and average out variance across examinations and populations, thereby enhancing the reliability of our findings.

The inclusion of both first-year students with greater variation in academic performance, and nominal second-year students allowed for a more comprehensive analysis of performance across different populations, while mitigating selection bias. Moreover, mixed-format examinations enabled within-subjects comparisons of VSAQs and MCQs, under comparable instructional and assessment conditions. Standardizing scores further reduced variability and ensured reliable comparisons.

However, this study also has limitations. While GPA is an objective and widely used measure of academic achievement, it simplifies complex learning outcomes and may overlook critical thinking and skill development [19]. Additionally, the variability in content and distribution between VSAQs and MCQs within real-world examinations could introduce bias. However, we mitigated this by aggregating data from multiple examinations and cohorts, standardizing scores, and conducting analyses across two student populations.

## Implications for practice and future research

Our findings indicate that VSAQs have a higher discriminative ability than MCQs, effectively distinguishing students both within individual examinations and across multiple examinations based on GPA. This suggests that VSAQs provide a more robust and valid question format for evaluating academic performance, supporting their implementation to enhance the discriminatory power and reliability of assessments. Incorporation of VSAQs into assessments can also enhance the ability to identify students in need of early interventions and support, while also recognizing those who excel. Moreover, VSAQs could play a valuable role in the selection process for medical studies by assessing study success potential while simultaneously providing an authentic preparation for the curriculum. Teachers could leverage VSAQs to implement tailored strategies for improving learning outcomes and to identify exceptional students for advanced opportunities. Future studies could expand academic performance measures beyond GPA, incorporating assessments of skill development and work-place learning. These additional metrics would provide a more holistic evaluation of the effectiveness of VSAQs and MCQs across different educational contexts. Additionally, exploring whether VSAQs are associated with long-term academic and professional success, including performance in real-world clinical settings, would provide deeper insights into their utility.

## Conclusion

This study highlights the superior effectiveness of VSAQs over MCQs in distinguishing undergraduate medical students with varying academic performance levels. Integrating VSAQs into assessments enhances the discriminative power and robustness of assessments and may improve early identification of both poor and excellent performing students, allowing for targeted interventions, tailored support, and advanced opportunities. Future research could explore the broader applicability of VSAQs across diverse educational settings and assess their potential to predict long-term academic achievements.

## References


1. Al-Rukban MO. Guidelines for the construction of multiple choice questions tests. Journal of Family & Community Medicine. 2006;13(3):125-33.
2. Schuwirth L, van der Vleuten C. Written Assessment. ABC of Learning and Teaching in Medicine: Wiley-Blackwell; 2017. p. 65-9.
3. Sam AH, Hameed S, Harris J, Meeran K. Validity of very short answer versus single best answer questions for undergraduate assessment. BMC Medical Education. 2016;16(1):266.
4. Schuwirth LWT, Vleuten CPM, Donkers HHLM. A closer look at cueing effects in multiple-choice questions. Medical Education. 1996;30(1):44-9.
5. Desjardins I, Touchie C, Pugh D, Wood TJ, Humphrey-Murto S. The impact of cueing on written examinations of clinical decision making: a case study. Medical Education. 2014;48(3):255-61.
6. Sam AH, Wilson R, Westacott R, Gurnell M, Melville C, Brown CA. Thinking differently – Students' cognitive processes when answering two different formats of written question. Medical Teacher. 2021;43(11):1278-85.
7. 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.
8. van Wijk EV, Janse RJ, Ruijter BN, Rohling JHT, van der Kraan J, Crobach S, et al. Use of very short answer questions compared to multiple choice questions in undergraduate medical students: An external validation study. PloS One. 2023;18(7):e0288558.
9. Sam AH, Field SM, Collares CF, van der Vleuten CPM, Wass VJ, Melville C, et al. Very-short-answer questions: reliability, discrimination and acceptability. Medical Education. 2018;52(4):447-55.
10. Sam AH, Peleva E, Fung CY, Cohen N, Benbow EW, Meeran K. Very Short Answer Questions: A Novel Approach To Summative Assessments In Pathology. Advances in Medical Education and Practice. 2019;Volume 10:943-8.
11. Mee J, Pandian R, Wolczynski J, Morales A, Paniagua M, Harik P, et al. An experimental comparison of multiple-choice and short-answer questions on a high-stakes test for medical students. Advances in Health Sciences Education. 2023;29(3).
12. Eijsvogels TMH, van den Brand TL, Hopman MTE. Multiple choice questions are superior to extended matching questions to identify medicine and biomedical sciences students who perform poorly. Perspectives on Medical Education. 2013;2(5):252-63.
13. Fenderson BA, Damjanov I, Robeson MR, Veloski JJ, Rubin E. The virtues of extended matching and uncued tests as alternatives to multiple choice questions. Human Pathology. 1997;28(5):526-32.
14. Frank J, Danoff D. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. Medical Teacher. 2007;29(7).
15. RemindoToets [Available from: https://www.paragin.nl/remindotoets/.
16. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education: Classical test theory and item response theory. Medical Education. 2010;44(1):109-17.
17. Krathwohl DR. A Revision of Bloom's Taxonomy: An Overview. Theory Into Practice. 2002;41(4).
18. Brookhart SM, Guskey TR, Bowers AJ, McMillan JH, Smith JK, Smith LF, et al. A Century of Grading Research. Review of Educational Research. 2016;86(4).
19. York TT, Gibson C, Rankin S. Defining and Measuring Academic Success. Practical Assessment, Research, and Evaluation. 2015;20(1).

# Appendix

## Appendix 1 – Supplemental Table 1

**Supplemental Table 1.** Distribution of VSAQs and MCQs across the themes within the examinations.

| | RM 2021 | | RM 2022 | | RM 2023 | | DA 2022 | | DA 2023 | | DA 2024 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | VSAQ | MCQ | VSAQ | MCQ | VSAQ | MCQ | VSAQ | MCQ | VSAQ | MCQ | VSAQ | MCQ |
| Theme 1 | 2 | 1 | 1 | 1 | 1 | 3 | 2 | 4 | 4 | 3 | 4 | 1 |
| Theme 2 | 5 | 7 | 6 | 9 | 5 | 5 | 4 | 5 | 4 | 7 | 3 | 3 |
| Theme 3 | 23 | 6 | 13 | 11 | 14 | 13 | 2 | 4 | 2 | 5 | 4 | 1 |
| Theme 4 | 4 | 11 | 5 | 8 | 5 | 10 | 3 | 4 | 6 | 3 | 7 | 1 |
| Theme 5 | 5 | 6 | 3 | 6 | 4 | 6 | 3 | 8 | 2 | 7 | 4 | 2 |
| Theme 6 | 3 | 1 | 1 | 1 | 2 | 1 | 2 | 11 | 6 | 4 | 5 | 4 |
| Theme 7 | *NA* | *NA* | *NA* | *NA* | *NA* | *NA* | 5 | 7 | 3 | 5 | 1 | 4 |
| Theme 8 | *NA* | *NA* | *NA* | *NA* | *NA* | *NA* | 1 | 5 | 3 | 6 | 3 | 4 |

RM=Regulation and Metabolism; DA=Diseases of the Abdomen; VSAQ=very short answer question; MCQ=multiple-choice question.

RM: Theme 1: Regulation of the temperature, Theme 2: Regulation of the reproduction, Theme 3: Regulation of the thyroid, Theme 4: Stomach, bowl and liver, Theme 5: Metabolism, Theme 6: Nutrition.

DA: Theme 1: Abdominal swelling, Theme 2: Stomach complaints, Theme 3: Jaundice, Theme 4: Acute abdominal pain, Theme 5: Chronic abdominal pain and defecation disorders, Theme 6: Blood loss, Theme 7: Anatomy, Theme 8: Other.

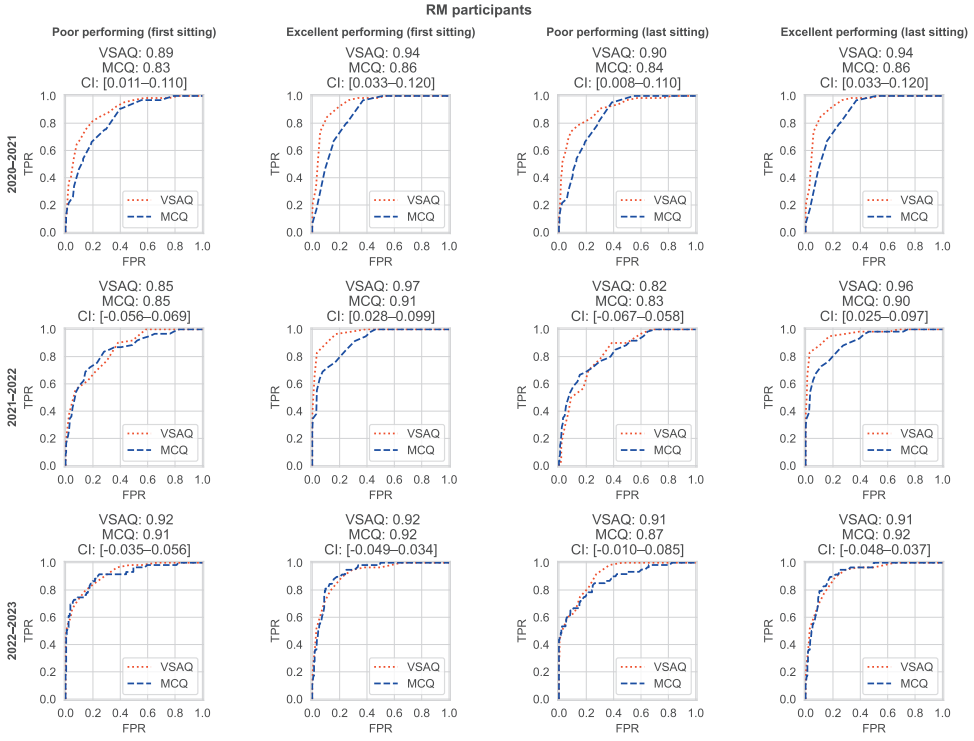
## Appendix 2 – Supplemental Table 2

**Supplemental Table 2.** Distribution of Bloom's taxonomy levels of VSAQs and MCQs within examinations.

| | RM 2021 | RM 2022 | RM 2023 | DA 2022 | DA 2023 | DA 2024 |
|---|---|---|---|---|---|---|
| Remember | | | | | | |
| MCQ | 43.8% | 47.4% | 52.5% | 41.7% | 42.5% | 55.0% |
| VSAQ | 80.9% | 72.4% | 77.4% | 54.6% | 40.0% | 35.5% |
| Understand | | | | | | |
| MCQ | 53.3% | 47.4% | 40.0% | 22.9% | 27.5% | 25.0% |
| VSAQ | 14.3% | 13.8% | 12.9% | 9.1% | 13.3% | 16.1% |
| Apply | | | | | | |
| MCQ | 3.1% | 5.3% | 7.5% | 35.4% | 30.0% | 20.0% |
| VSAQ | 4.8% | 13.8% | 9.7% | 36.4% | 46.7% | 48.4% |

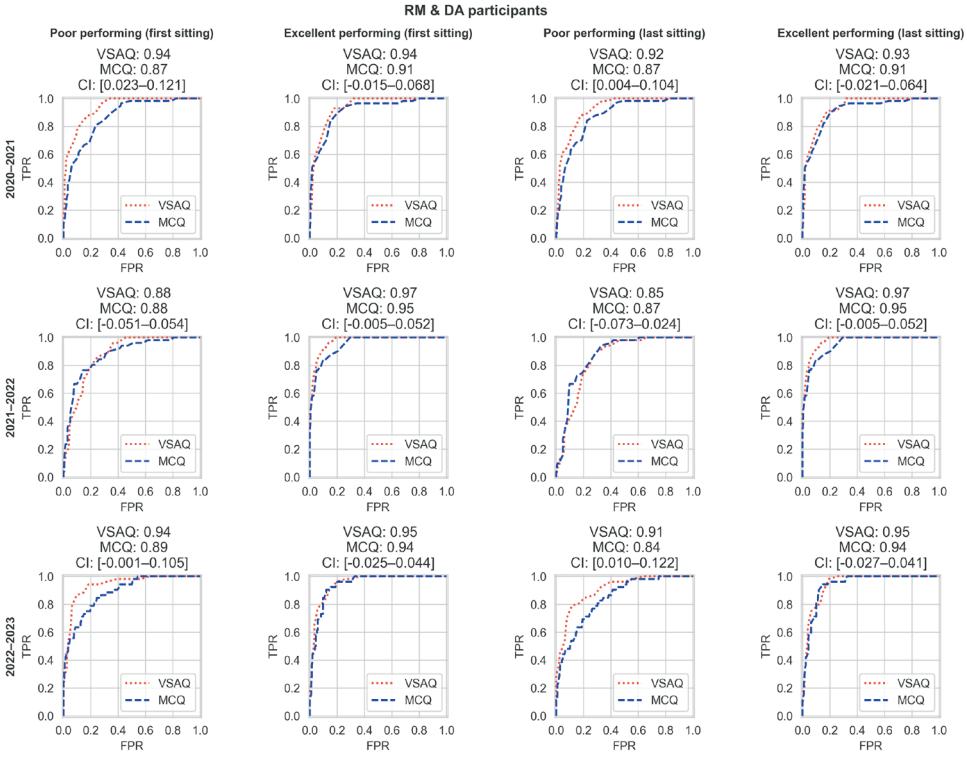
RM=Regulation and Metabolism; DA=Diseases of the Abdomen; VSAQ=very short answer question; MCQ=multiple-choice question.

## Appendix 3 – Supplemental Figure 1




**Supplemental Figure 1.** Receiver Operating Characteristic (ROC) curves for both poor and excellent performing students using the first sitting or last sitting grades from the RM participants cohort 2020-2021 (RM 2021), cohort 2021-2022 (RM 2022), and cohort 2022-2023 (RM 2023). Red = VSAQs; Blue = MCQs; CI = 95% confidence interval; TPR = true positive rate; FPR = false positive rate. C-statistics are shown above each graph, which were calculated from the area under the ROC curve.

## Appendix 4 – Supplemental Figure 2




**Supplemental Figure 2.** Receiver Operating Characteristic (ROC) curves for both poor and excellent performing students using the first sitting or last sitting grades from the RM & DA participants cohort 2020-2021 (RM 2021 & DA 2022), cohort 2021-2022 (RM 2022 & DA 2023), and cohort 2022-2023 (RM 2023 & DA 2024). Red = VSAQs; Blue = MCQs; CI = 95% confidence interval; TPR = true positive rate; FPR = false positive rate. C-statistics are shown above each graph, which were calculated from the area under the ROC curve.