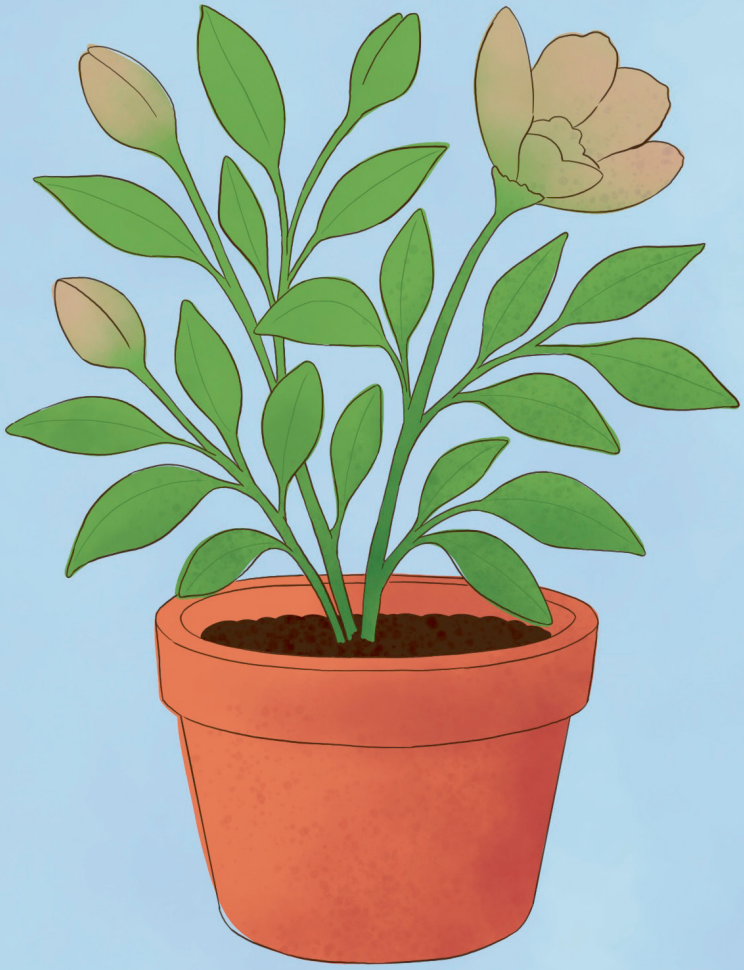
# Assessment for growth: fostering student learning through assessment innovations in medical education

Wijk, E.V. van



## Part I: Very short answer question

# Chapter 4

## The battle of question formats: A comparative study of retrieval practice using very short answer questions and multiple choice questions

Elise V. van Wijk
Mario de Jonge
Floris M. van Blankenstein
Roemer J. Janse
Alexandra M.J. Langers

*BMC Medical Education.* 2024;24(1):1547.
DOI: 10.1186/s12909-024-06538-0

## Abstract

**Background**: Retrieval practice is a highly effective learning strategy that enhances long-term retention by encouraging the active recall of information. However, the optimal question format for maximizing knowledge retention remains uncertain. In this study, we compared the effect of very short answer (VSAQ) versus multiple-choice question (MCQ) practice tests on students' knowledge retention. By analyzing these two formats, we aim to identify the most effective approach to retrieval practice, thereby helping to optimize its implementation and improve learning outcomes.

**Methods**: In this randomized within-subjects study, students (n = 45) practiced with both VSAQs and MCQs in an extracurricular lifestyle course, without receiving feedback. The final retention test consisted of identical questions in both formats. A 2 × 2 repeated measures ANOVA was used to determine the effect of question format in practice testing and final test on final test score. Additionally, digital questionnaires were used to explore students' test-taking experiences.

**Results**: The VSAQs were answered incorrectly more frequently on the practice tests and final test. There was no main effect of practice question format on final test performance, and no interaction effect between question format on the practice and final test. Regardless of question format, most students thought the practice tests were beneficial for learning.

**Conclusions**: We found no evidence indicating that either MCQ or VSAQ is more effective for knowledge retention during retrieval practice. The lower initial retrieval success in the VSAQs, indicated by the higher degree of incorrect answers on the practice tests, might have limited their effectiveness during retrieval practice. To optimize the use of VSAQs in retrieval practice, it seems important to improve initial retrieval success to maximize learning outcomes.

## Introduction

Retrieval practice is a highly effective learning strategy that enhances robust learning and long-term retention by requiring individuals to recall previously learned information from memory. This phenomenon, known as the testing effect or test-enhanced learning (TEL), is commonly implemented through practice tests or quizzes [1,2,3]. In health professions education, multiple-choice questions (MCQs) are frequently utilized as the practice test format. However, open-ended questions such as very short answer questions (VSAQs) are gaining popularity due to their advantageous psychometric properties and ability to provide deeper insights into students' misperceptions and common errors [4,5,6,7,8,9]. Despite the recognized learning benefits of retrieval practice, the question of which format – MCQs or open-ended questions – is most effective for enhancing learning and improving knowledge retention remains unresolved and is the subject of ongoing debate [10,11,12,13,14].



In health professions education, MCQs, mostly in the form of Single Best Answers (SBA), are commonly used to assess students due to their ease of marking, feasibility, and high reliability [15, 16]. However, their susceptibility to cueing, and more superficial recognition-based exam preparation [4, 7, 17] can decrease the effort of knowledge retrieval from memory, resulting in less effective retrieval practice and knowledge retention [18]. This aligns with the retrieval effort hypothesis, which states that retrieval requiring greater cognitive effort proves more beneficial for learning than easy retrieval with prompts or cues [19]. Consequently, students focusing on cues to respond to the question with the least effort (i.e., principle of least effort [20]), may fail to reproduce answers from memory on a later test without available answer options [21, 22].

Open-ended questions may be more effective for retrieval practice compared to MCQs, though the evidence is inconclusive [19]. Previous studies in cognitive psychology predominantly compared MCQs with open-ended Short Answer Questions (SAQs). Although SAQs allow for more extensive responses, such as several sentences, they are often employed in practice a way that mirrors VSAQs, with answers typically restricted to one or a few words. Several studies show that retrieval practice using SAQs is more effective than retrieval practice with MCQs, most likely because SAQs require more effort in retrieving information from memory and are typically experienced as more difficult [4, 6, 18, 23,24,25].

For example, Gay *et al.* [24] compared the effects of six SAQ and MCQ tests, covering identical concepts, in two groups of 14 students in an educational research course. The study assessed students' performance on a final exam, which tested the same concepts using both question formats. The results showed that SAQ practice led to better knowledge retention when tested with SAQs on the final exam, while the retention was comparable between the two formats when tested with MCQs on the final exam. This finding is in line with the theory of transfer-appropriate processing (TAP), which refers to the idea that learning and knowledge retention are improved if the type of processing used during retrieval matches the type of processing used during encoding, in this case the question format [21, 26]. Another study compared VSAQs and MCQs in a formative online test at the end of a pharmacology course and after one year [27]. They found an overall increase in knowledge retention after one year, which was higher in the students who started with VSAQs in the formative test. Nonetheless, this study mainly assessed the effect of the order in which the questions were given in the tests (first VSAQs or first MCQs) and which

question format offered a better preparation in basic pharmacology. There are also studies that found no advantages of retrieval practice using SAQs against MCQs, which may be associated with the lower level of initial retrieval success in SAQs [14, 28,29,30].

In summary, consensus on the impact of different question formats on knowledge retention remains elusive, and previous studies mainly tested in a non-medical or simulated setting [18, 23,24,25, 29]. Therefore, we aimed to investigate the effectiveness of retrieval practice with VSAQs and MCQs on knowledge retention in a real-life educational stetting among health science students. We also evaluated students' experiences with the practice tests. Our hypothesis is that VSAQs will be more effective compared to MCQs as practice format. In line with the TAP theory, we expect that the benefit of VSAQs over MCQs will be most pronounced when students are also tested with VSAQs on the final retention test. Additionally, this effect may be influenced by the greater sensitivity of VSAQs as a memory test, as they require active retrieval rather than recognition. By investigating this, we can further optimize the implementation of retrieval practice in health professions curricula and improve learning outcomes.

# Methods

## Setting

This study was conducted during the 2024 Students Experienced in Lifestyle and Food (SELF) course at Leiden University Medical Center (LUMC). The SELF-course is an extracurricular program that covers lifestyle medicine topics not covered in the standard medical curriculum. It is offered annually in February-March at all medical schools in the Netherlands to (bio)medical or healthcare related students, and early-career doctors. The course comprises eight evening sessions (i.e., masterclasses) of two hours every week, each focusing on a specific lifestyle-related theme (*Appendix 1 – Course overview*). This years' masterclasses were each divided into two lectures, given by a different speaker. The masterclasses were given live in a lecture hall, with an option for students to attend virtually. No preparatory work was required for the masterclasses, and the presentations of the speakers were only shared with the students at the end of the course. Directly following each masterclass, students completed a short formative quiz (i.e., practice test), with the two highest scoring students receiving a small reward from the SELF-board (e.g., a cookbook). The course concluded with a final assignment, where students created a poster to encourage healthier eating habits among general practitioner patients. Upon successful completion of the course, students were awarded a certificate. Medical students who participated in the honours program of the university received additional in-depth practical assignments related to the themes, but not directly to the lectures, and were awarded with study credits upon successful completion.

## Participants

A total of 62 participants enrolled in the SELF-course, of 48 complete both the practice tests and the final test. We only included bachelor and master students enrolled in the course, and excluded participants who were already working in healthcare and no longer enrolled as students ($n$=2) to minimize potential confounding effects due to differences in baseline knowledge. Additionally, students who had missed more than two practice tests were excluded from the analysis ($n$=1), as they completed only a single practice test. This resulted in a final sample size of 45.

## Study design and procedure

We conducted an experimental study with a within-subjects design, in a real-world educational setting. We provided information to the students during the first masterclass of the SELF-course, which they also received by e-mail (*Figure 1*). The practice test of this first masterclass was created by the SELF-board, and not part of our study to give students enough time to think about participation. Our research team developed the subsequent four practice tests. Following the second masterclass, participants were randomized by using a computer-generated sequence to start the practice test with either four VSAQs about the first lecture followed by four MCQs about the second lecture (*version A*), or the counterbalanced version (*version B*). Prior to the first practice test, students were asked for demographical data (*Appendix 2* includes both demographical questions and questions of the first practice test). In the subsequent masterclasses, the practice tests formats were reversed each week (Additional file 3 provides a figure with a more detailed overview of the different versions). To isolate the effects of retrieval practice of the question format, students were not provided with the correct answers or the presentations after the tests. The final test was administered at the beginning of the last masterclass in week 8. The retention intervals for the different practice tests were six weeks for the first practice test, five weeks for the second, four weeks for the third, and three weeks for the last practice test (*Figure 1*), and consisted of questions used in the four practice tests ($n$=32; 16 VSAQs, 16 MCQs). Within the two groups (randomized to version A and B) students were further randomized to take one of the two counterbalanced versions of the final test (*version 1 or 2*). In version 1, the first two questions of each of the masterclass lectures were presented in VSAQ format, while the last two questions were in MCQ format (VSAQ | MCQ, *Figure 1 & Appendix 3 – Supplemental Figure 1*).). In version 2, this order was reversed, with the first two questions were in MCQ format and the last two in VSAQ format (MCQ | VSAQ, *Figure 1 & Appendix 3 – Supplemental Figure 1*). Since students received half of the final test questions in the same format as those in the practice test (congruent questions) and the other half in a different format (incongruent questions), we were able to assess the theory of TAP. After completing the final test, students competed a digital questionnaire regarding their experiences, and they received the correct answers of the final test.

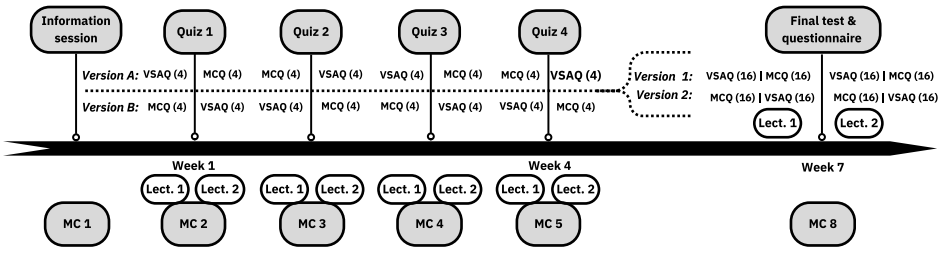



**Figure 1.** Study design with the different versions of the practice tests and final test. MC = masterclass; Quiz = practice test; Lect. = lecture; MCQ = multiple-choice questions; VSAQ = very short answer questions. The number of questions on the tests is reported between brackets.

## Development of practice tests and final test

The principal investigator (EvW) developed the practice test questions based on lecture materials, additional information and key messages provided by the speakers. Two educationalists experienced in assessment question writing along with the speakers, reviewed the questions to ensure their quality and alignment with the lecture content. The educationalists discussed their suggestions with the principal investigator, which led to adjustments of the original questions. All questions were designed to assess the first two stages of the concept of Bloom's Taxonomy: recall and understanding [31]. Each question was designed to suit both the VSAQ and MCQ format, without altering the question's structure or content. Students received the questions either in VSAQ or MCQ format. The MCQs had four answer options, with one correct answer (i.e., single best answer questions). Answer options were randomly ordered for each participant. The VSAQs required responses of no more than four words.

## Scoring procedure

MCQs were automatically scored by using a standard letter key. For the VSAQs, a predefined list of acceptable answers was used, and these were automatically marked as correct in the digital assessment system. This predefined list was created by the item writer (EvW) together with the speakers. Subsequently, all incorrect answers were reviewed, and correct answers not initially included in the predefined list were added. Each correct answer was awarded one point, while incorrect or partially correct questions yielded no point, with a maximum score of 8 point on each practice test and 32 on the final test. Three investigators of the research team with a (bio)medical and educational background (EvW, MdJ, and AL) scored the VSAQs independently in a blinded manner. They indicated which of the answers for each question in the list they thought was correct, incorrect, or which of the answers they were not sure about. The individual scores were then reviewed and discussed a joint session to reach consensus on the final scores. Out of the total 2683 answers, only 77 responses (2.87%) required further discussion to reach consensus. If the question explicitly required a single response but multiple answers were given, the response was marked incorrect, even if one of the answers was correct. Conversely, for questions where the number of expected answers was not specified, the response was considered correct if at least one of the provided answers was correct. We also approved answers that were not explicitly mentioned during the lectures, and were thus not included in the predefined list of answers, but proved to be correct after all. One question, in both question formats, was removed during the reviewing process because the topic was not discussed during the lecture. The speakers were consulted about the answers that required more in-depth knowledge of the topic.

## Data collection

Directly at the end of the second, third, fourth, and fifth masterclass, participants undertook digital practice tests via the Remindo assessment system [32]. These tests could be taken either in the lecture hall or remotely, as approximately 20% of students attended the masterclasses online. For two students who were unable to access the digital system, paper-based tests were provided. Each practice test consisted of eight questions, from which the first four questions focused on the content from the first lecture, followed by four questions related to the second lecture. The order of the questions was fixed, ensuring a structured progression through the material covered in each lecture. On the digital practice tests, students were required to answer all questions, without the option to skip questions or to return to previous questions.

All students who participated in the practice tests completed every question on both the digital and paper version. This setup was consisted across all practice tests administered during the study. Participants were instructed not to use study materials during the test, but we did not use proctoring to verify this. We allocated 20 min to finish each practice test, after which all students were allowed to leave. The final test, administered at the start of the eighth masterclass, consisted of all questions derived from the four practice tests (*n*=32; 16 VSAQs and 16 MCQs). Upon initiation and completion of the final test, students answered questions in the digital assessment system regarding their experiences with the practice tests. For this study we analysed the four questions answered before the final test. The complete questionnaire can be found in *Appendix 4*.

### Data analysis

Descriptive statistics for the demographics and previous experience of participants were calculated. Continuous data, depending on their distribution, are reported as mean and standard deviation (SD) or median and interquartile range (IRQ). Categorical variables are expressed as a number (percentage). Total test scores were calculated by adding the points for each question. In line with previous research, we did not correct for guessing on the MCQs [14, 24, 27]. We reported the percentages of questions practiced in VSAQ or MCQ format that were incorrect/correct in the practice test and/or final test, and used a 2 (practice test format) × 2 (final test format) repeated measures ANOVA to analyse the main effects and interaction effect of question formats on practice and final test on final test score. The answers to the short questionnaire regarding student's experiences were reported as answer distributions. Missing data were not imputed. Statistical significance is denoted by p-values and 95% confidence intervals, with a p-value of less than 0.05 considered significant. All statistical analyses were performed using R version 4.3.1 (R Foundation for Statistical Computing, Vienna, Austria).

### Ethical approval

This study was approved by the LUMC Educational Research Review Board (OEC/ERRB/20231010/1). Participation in this study was voluntary and all students received verbal and written information prior to the study. Upon initiation of the first practice test, students provided informed consent.

## Results

### Demographics

Of the 62 participants in the SELF-course, 55 participants completed the practice tests. The final retention test was completed by 48 participants. We excluded three participants, because they were not a student (*n*=2), or had only completed less than two practice tests (*n*=1), resulting in a sample of 45 students (*Figure 2*). *Table 1* shows that the majority of participants studied medicine (71.1%), and were bachelor students (57.7%). Two students prepared for the final test. There were no differences in characteristics between the student groups in the counterbalanced versions (*version A and B*).

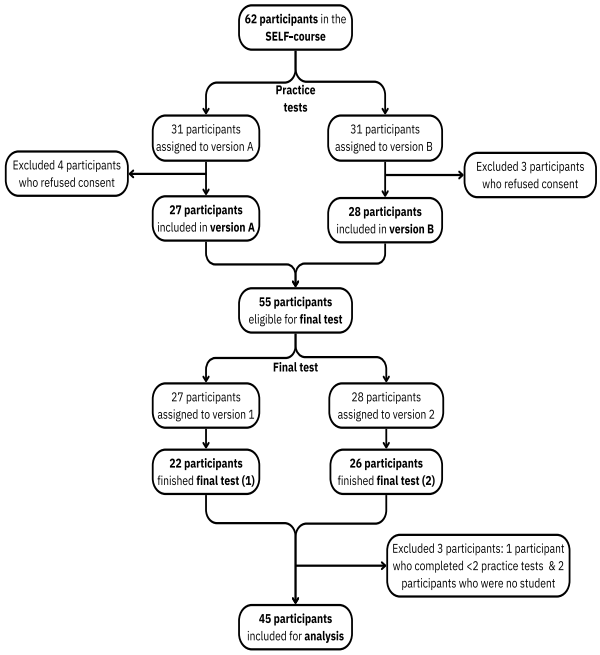



**Figure 2.** Flowchart of participants of the study

**Table 1.** Characteristics of the participants of the study

| | **Overall (*n*=45)** |
|---|---|
| Age, median (IQR) | 21 (20, 23) |
| Study, n (%) | |
| Biomedical Sciences | 7 (15.6) |
| Medicine | 32 (71.1) |
| Psychology | 2 (4.4) |
| Vitality and Aging | 3 (6.7) |
| Biomedical Sciences & Medicine | 1 (2.2) |
| Study year, n (%) | |
| Bachelor[a] year 1 | 4 (8.9) |
| Bachelor year 2 | 11 (24.4) |
| Bachelor year 3 | 11 (24.4) |
| Master[b] year 1 | 9 (20.0) |
| Master year 2 | 4 (8.9) |
| Master year 3 | 3 (6.7) |
| Waiting period[c] | 3 (6.7) |
| Honours program | |
| Yes | 15 (33.3) |
| Previous lifestyle courses | |
| Yes | 4 (8.9) |

IQR = interquartile range

[a]Bachelor = pre-clinical phase in medicine

[b]Master = clinical phase in medicine

[c]Waiting period before the clinical clerkships during the postgraduate phase.

## Performance on practice tests and final test

The mean total score on the final test was 14% lower than the mean overall score of all four practice tests combined (mean percentage score (SD) practice tests: 66.2% (12.1) vs. final test: 52.7% (10.1); paired t-test: $t(44) = 6.99$; $p < 0.01$; 95%CI [0.10,0.17]). Successful retrieval practice, defined as correctly answered questions in both the practice test and the final retention test, occurred in 50.4% of questions practiced in the MCQ format vs. 39.8% of questions practiced in VSAQ format (*Table 2*). For both question formats approximately half of the initially correct answers were forgotten on the final test (23.5/50.4% for MCQs; 18.8/39.8% for VSAQs). 23.5% of the MCQs, and 18.8% of VSAQs were initially answered correctly on the practice test, but answered incorrectly in the final test. MCQs were more often answered correctly on the practice tests (73.9% vs. 58.6%), which is not surprising and was to be expected since MCQs also allow for guessing.

**Table 2**. Percentages of correct and incorrect answers on VSAQs and MCQs

| Practice test question format and performance | Final test performance[a] | | | | |
|---|---|---|---|---|---|
| | Correct | | Incorrect | | Total |
| | MCQ | VSAQ | MCQ | VSAQ | |
| **Multiple-choice questions** | | | | | |
| Correct | 28.3% | 22.1% | 8.7% | 14.8% | 73.9% |
| Incorrect | 4.0% | 2.7% | 8.9% | 10.5% | 26.1% |
| **Very short answer questions** | | | | | |
| Correct | 21.1% | 18.7% | 7.7% | 11.2% | 58.6% |
| Incorrect | 7.2% | 4.2% | 13.7% | 16.3% | 41.4% |

[a]Questions identical to practice test questions with a random 50/50 distribution of both VSAQs and MCQs

## Effect of question format on final retention test performance

The 2 × 2 (question format practice test × question format final test) repeated measures ANOVA showed no significant main effect of question format in practice tests on the final test score ($F(1,44) = 3.23$, $p = 0.08$, $\eta 2 = 0.02$). The mean proportion of correctly answered questions on the final test was similar for the content practiced with MCQs (M (SD) 57.1% (2.2)) and VSAQs (51.2% (2.3)). Question format on the final test had a significant main effect on the final test score ($F(1, 44) = 22.80$, $p < 0.01$, $\eta 2 = 0.10$) with lower proportional scores on VSAQs compared to MCQs, (M (SD) MCQ: 61.0% (2.2), VSAQ: 47.2% (2.1)). Figure 3 shows the proportion of correctly answered questions on the final test as a function of question format in the practice tests and on the final test (*Appendix 5* shows the table with the mean percentages). There was no interaction effect between question format on practice test and final test ($F(1, 44) = 0.16$, $p = 0.69$, $\eta 2 < 0.01$).

## Student's experiences

Of all students, 68.9% ($n$=31) students thought the practice tests helped them remember the content of the lectures better. There was no difference between the estimation of how much students thought they still remembered of the lectures practiced with VSAQs vs. MCQs (median of 0%, *Appendix 6* shows the histogram).

## Discussion

In this study we investigated the effectiveness of retrieval practice using MCQs and VSAQs on knowledge retention in an educational setting. We found no significant difference in knowledge retention between retrieval practice with MCQs and VSAQs, and no interaction effect between the question formats. This indicates that we found no evidence of an advantage associated with either question format in retrieval practice. However, VSAQs were more challenging, as reflected by lower scores on both practice and final tests. Most students reported that the practice tests were beneficial, regardless of the question format.

Our findings align with previous research comparing SAQs and MCQs, which also found no differences in the effectiveness of retrieval practice [14, 28,29,30, 33]. Similarly, we observed no significant interaction effect between practice and final test question format, indicating that matching question formats does not enhance the retrieval effect. While it challenges the TAP theory, it supports earlier findings suggesting that retrieval practice benefits do not depend on an exact match between conditions in practice and final test [11, 12, 14, 30, 33].

One explanation for the lack of a retrieval benefit in our study could relate to the level of initial retrieval success, which is often lower for VSAQs due to their greater difficulty [3, 14, 23, 34]. Rowland's meta-analysis [34] indicates that higher retrieval success often leads to a stronger retrieval effect. If initial retrieval is unsuccessful, it may lead to lower levels of knowledge retention [14]. However, a recent study suggests there may be no consistent relationship between initial retrieval success and the testing effect [35], underscoring the complexity of this relationship and the potential influence of other moderating factors, such as individual differences and contextual variations. Feedback might correct for unsuccessful retrieval, but we intentionally did not provide feedback after the practice tests to isolate the direct effects of the different question formats. Consistent with earlier findings [4, 6, 7, 27], VSAQs were answered incorrectly more often than the MCQs, resulting in lower retrieval success.

Previous studies show mixed results when corrective feedback is provided, with some indicating better learning outcomes for SAQs compared to MCQs, while others found no differences [11, 12, 30, 33]. Although feedback ensures exposure to correct answers, it also introduces indirect effects that are often not examined, complicating the assessment of direct retrieval effects of different question formats on learning [3, 36]. The ability to recognize the correct answer among MCQ options may also serve as implicit feedback, potentially enhancing the learning effect for MCQs. In studies showing greater knowledge retention with (V)SAQs compared to MCQs, practice and final test questions did not exactly match, and it was unclear whether feedback was provided or what the initial retrieval success levels were [24, 27]. The timing of the practice tests may further influence retrieval success. While administering them shortly after learning, as in our study, may enhance success, scheduling them late in the evening after a two-hour lecture might have hindered it, particularly for the more challenging VSAQs. In sum, and in line with previous research [3, 30, 33, 35, 37], finding the appropriate balance between retrieval success and effort is complex, as it may also be influenced by various other factors.

### Strengths & limitations

To our knowledge, this study is the first to compare the effect of VSAQ and MCQ practice tests among health science students using a controlled design within a real-life educational setting.

We used a rigorous methodology, including a within-subjects design, counterbalanced test versions, and identical questions in both tests. Although the use of identical questions across tests is not common practice in a real-life educational setting, it was necessary to ensure a pure measurement of the memorial effects of question formats. Students had comparable prior knowledge due to the absence of prior mandatory education on lifestyle topics, and most had not taken previous lifestyle courses. Moreover, there were no external incentives, such as preparatory assignments, that could influence learning effort. This design allowed us to directly measure retrieval practice effects without simulating a test environment [11, 18, 23]. The random distribution of both question formats in the final retention test equalized difficulty levels, and enabled analysis of how differences between initial and final test formats affected performance.

The absence of feedback, while allowing a direct measurement of retrieval effects, may have influenced our results, particularly for VSAQs, were students showed lower initial retrieval success. Research indicates that feedback can enhance the testing effect by reinforcing correct responses and correcting errors, especially when initial retrieval success is unsuccessful [11, 12, 34, 36, 38]. Including feedback might have amplified the differences between question formats, improving VSAQ performance more significantly than MCQ performance. Additionally, the limited number of test questions and sample size, though larger than in most prior studies [23, 24, 29], may have affected the robustness of our findings. A larger sample size could reveal differences between MCQ and VSAQ formats, warranting further investigation.



While our within-subjects design and random assignment of students to counterbalanced versions minimized confounding variables, such influences cannot be entirely eliminated in a real-world educational setting. Nonetheless, our study design more accurately reflects the uncontrolled conditions typical of real-world educational settings, enhancing the generalizability of the findings. Variations in lecture attendance (online vs. in-person) and test format (digital vs. paper) might have influenced student motivation and test performance. Furthermore, the voluntary, low-stakes nature of this course may differ from traditional graded courses, potentially affecting both students' engagement with the material and their test performance. However, this low-stakes setting reduced additional studying between tests, providing a clearer view of retrieval practice effects. It remains possible, though, that some motivated students independently sought feedback. Because we compared the effect of two different question formats on retrieval practice without including a control group of students who did not complete practice questions, we cannot conclude that learning was directly supported by the practice questions in our setting. However, the general effectiveness of practice testing has been consistently demonstrated in previous research.

## Implications for practice and future research

Our findings suggest no clear advantage of either question format in enhancing learning through practice tests, regardless of the format used in the final test. Achieving an optimal balance between retrieval effort and retrieval success may enhance the effectiveness of retrieval practice. Practically, this indicates that teachers can choose either VSAQs or MCQs for practice testing, as performance was not influenced by practice question format or alignment with the final test format. However, VSAQs may offer learners better insights into knowledge gaps and misperceptions compared to MCQs, making them a preferred diagnostic

tool to guide future learning [8, 9], particularly when assessing higher-order skills such as knowledge application. In educational settings, VSAQs could be prioritized when the goal is to address specific misconceptions, as they provide deeper insights into student understanding. For lower-order skills like remembering and understanding, as examined in this study, VSAQs are more likely to reflect retrieval failures or memory interference, which can be valuable for formative assessments but may require additional feedback to maximize their effectiveness.

A key advantage of VSAQs is their ability to reduce 'foresight bias' common with MCQs, providing a more reliable estimation of students' understanding, and better predicting retrieval success [39, 40]. This can help teachers identify underperforming students earlier and intervene more effectively. Additionally, VSAQs eliminate the need to create plausible distractors, a challenging and time-consuming aspect of MCQ construction [41,42,43,44], thereby saving teachers time and improving test quality. To address teachers' limited experience with VSAQs, providing clear instructions or workshops on constructing effective VSAQs would be beneficial.

To enhance the effectiveness of retrieval practice, particularly with VSAQs, we recommend integrating immediate or delayed feedback [11, 12, 45], and implementing repeated spaced practice retrieval [37]. As alternative for teacher-learner feedback, self-assessment by students in immediate self-feedback VSAQs might be a good option to use in formative assessments [9]. In this format, students give self-feedback on the understanding of the correct answers, which can help students recognize their knowledge gaps and guide further learning. Hybrid questions, such as combining VSAQs with MCQs or using stepwise MCQs, offer another promising approach [38, 39, 46]. These formats engage students in initial effortful retrieval (i.e., answering an open question without cues) followed by multiple-choice options for direct feedback, which balances retrieval effort and success. Further research is needed to evaluate the effectiveness of these formats in health professions education. Moreover, future research could explore factors influencing the effectiveness of retrieval practice with different question formats, including the level of initial retrieval success, individual differences in memory strength, learning strategies, and student motivation. Studies in diverse educational contexts, with longer retention intervals, or questions that assess higher levels of cognitive learning could further unravel the impact of retrieval practice on learning outcomes across different question formats.

## Conclusion

We found no evidence to suggest that either question format, MCQ or VSAQ, is more effective for knowledge retention through retrieval practice. Despite higher retrieval effort in VSAQs, their lower initial retrieval success may have limited their effectiveness in enhancing retention. Nevertheless, practice testing with VSAQs offers valuable insights into knowledge gaps and provides a more reliable estimation of students' understanding. To optimize learning outcomes and enhance knowledge retention, it seems important to increase initial retrieval success, which could be achieved through feedback or repeated practice sessions.

## References


1. Dunlosky J, KA R, Marsh E, Nathan M, Willingham D. Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology - PubMed. Psychological science in the public interest : a journal of the American Psychological Society. 2013 Jan;14(1).
2. Butler AC, Karpicke JD, Roediger HL. The effect of type and timing of feedback on learning from multiple-choice tests. Journal of Experimental Psychology Applied. 2007;13(4):273-81.
3. Karpicke JD. Retrieval-based learning: A decade of progress.: Academic Press; 2017.
4. 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.
5. 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.
6. 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.
7. 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.
8. Bala L, Westacott RJ, Brown C, Sam AH. Twelve tips for introducing very short answer questions (VSAQs) into your medical curriculum. Medical Teacher. 2022:1-8.
9. Lertsakulbunlue S, Kantiwong A. Development and validation of immediate self-feedback very short answer questions for medical students: practical implementation of generalizability theory to estimate reliability in formative examination designs. BMC Medical Education 2024 24:1. 2024;24(1).
10. Greving S, Richter T. Examining the Testing Effect in University Teaching: Retrievability and Question Format Matter. Frontiers in Psychology. 2018;9.
11. Kang SHK, McDermott KB, Roediger HL. Test format and corrective feedback modify the effect of testing on long-term retention. European Journal of Cognitive Psychology. 2007;19(4-5):528-58.
12. Little JL, Bjork EL, Bjork RA, Angello G. Multiple-Choice Tests Exonerated, at Least of Some Charges: Fostering Test-Induced Learning and Avoiding Test-Induced Forgetting. Psychological Science. 2012;23(11):1337-44.
13. Mcdaniel MA, Roediger HL, Mcdermott KB. Generalizing test-enhanced learning from the laboratory to the classroom. Psychonomic Bulletin & Review. 2007;14(2):200-6.
14. Smith MA, Karpicke JD. Retrieval practice with short-answer, multiple-choice, and hybrid tests. Memory. 2014;22(7):784-802.
15. Al-Rukban MO. Guidelines for the construction of multiple choice questions tests. Journal of Family & Community Medicine. 2006;13(3):125-33.
16. Schuwirth L, van der Vleuten C. Written Assessment. ABC of Learning and Teaching in Medicine: Wiley-Blackwell; 2017. p. 65-9.
17. Schuwirth LWT, Vleuten CPM, Donkers HHLM. A closer look at cueing effects in multiple-choice questions. Medical Education. 1996;30(1):44-9.
18. Larsen DP, Butler AC, Roediger III HL. Test-enhanced learning in medical education. Medical Education. 2008;42(10):959-66.
19. Pyc MA, Rawson KA. Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? Journal of Memory and Language. 2009;60(4):437-47.
20. Samuels SJ. Effects of pictures on learning to read, comprehension and attitudes. Review of Educational Research. 1970;40(3):397-407.
21. Morris CD, Bransford JD, Franks JJ. Levels of Processing versus Tranfer Appropriate Processing. Journal of Verbal Learning and Verbal Behavior. 1977;16(5):519-33.
22. Roediger HL, Gallo DA, Geraci L. Processing approaches to cognition: The impetus from the levels-of-processing framework. Memory. 2002;10(5-6):319-32.
23. Butler AC, Roediger HL. Testing improves long-term retention in a simulated classroom setting. European Journal of Cognitive Psychology. 2007;19(4-5):514-27.
24. Gay LR. The comparative effects of multiple-choice versus short-answer tests on retention. Journal of Educational Measurement. 1980;17:45-50.
25. Greving S, Richter T. Practicing retrieval in university teaching: short-answer questions are beneficial, as multiple-choice questions are not. Journal of Cognitive Psychology. 2022;34(5):657-74.
26. Veltre MT, Cho KW, Neely JH. Transfer-appropriate processing in the testing effect. Memory. 2015-11-17;23(8).
27. Neumann J, Simmrodt S, Teichert H, Gergs U. Comparison of Online Tests of Very Short Answer versus Single Best Answers for Medical Students in a Pharmacology Course over One Year. Education Research International. 2021;2021:1-10.
28. Moreira BFT, Pinto TSS, Starling DSV, Jaeger A. Retrieval Practice in Classroom Settings: A Review of Applied Research. Frontiers in Education. 2019;4.
29. Lau KY, Ang JYH, Rajalingam P. Very Short Answer Questions in Team-Based Learning: Limited Effect on Peer Elaboration and Memory. Medical Science

Educator. 2023;33(1):139-45.

30. Bloom BS. Taxonomy of Educational Objectives: The Classification of Educational Goals: Longmans, Green; 1956 1956. 240 p.
31. RemindoToets [Available from: https://www.paragin.nl/remindotoets/.
32. McDaniel MA, Wildman KM, Anderson JL. Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. Journal of Applied Research in Memory and Cognition. 2012;1(1).
33. McDermott KB, Agarwal PK, D'Antonio L, Roediger HL, McDaniel MA. Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. Journal of Experimental Psychology: Applied. 2014;20(1).
34. Rowland CA. The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. Psychological Bulletin. 2014;140(6).
35. Green ML, Moeller JJ, Spak JM. Test-enhanced learning in health professions education: A systematic review: BEME Guide No. 48. Medical Teacher. 2018-4-3.
36. Vaughn KE, Rawson KA, Pyc MA. Repeated retrieval practice and item difficulty: Does criterion learning eliminate item difficulty effects? Psychonomic Bulletin & Review. 2013;20(6).
37. van den Broek GSE, van Gog T, Jansen E, Pleijsant M, Kester L. Multimedia Effects During Retrieval Practice: Images That Reveal the Answer Reduce Vocabulary Learning. Journal of Educational Psychology. 2021;113(8):1587-608.
38. Koriat A, Bjork RA. Illusions of Competence in Monitoring One's Knowledge During Study. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2005;31(2):187-94.
39. van den Broek GSE, Gerritsen SL, Oomen ITJ, Velthoven E, van Boxtel FHJ, Kester L, et al. Optimizing Multiple-Choice Questions for Retrieval Practice: Delayed Display of Answer Alternatives Enhances Vocabulary Learning. Journal of Educational Psychology. 2023;115(8):1087-109.
40. Little JL, Frickey EA, Fung AK. The role of retrieval in answering multiple-choice questions. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2019;45(8).
41. Little JL, Bjork EL. Optimizing multiple-choice tests as tools for learning. Memory & Cognition. 2015;43(1):14-26.
42. Gierl MJ, Bulut O, Guo Q, Zhang X. Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. Review of Educational Research. 2017;87(6).
43. Ryan AT, Judd T, Wilson C, Larsen DP, Elliott S, Kulasegaram K, et al. Timing's not everything: Immediate and delayed feedback are equally beneficial for performance in formative multiple-choice testing. Medical Education. 2024/07/01;58(7).
44. Park J. Learning in a New Computerized Testing System. Journal of Educational Psychology. 2005;97(3).
45. Park J, Choi B-C. Higher retention after a new take-home computerised test. British Journal of Educational Technology. 2008;39(3).

# Appendix

## Appendix 1 – Course overview

***Theme 1: Sleep***

**Lecture 1**

- What is good sleep?
- What is the influence of sleep on (development of) health/disease?
- What is the influence of too little or too much sleep on the development of disease?
- What feasible advice can you give to patients regarding sleep?

**Lecture 2**

- What is the influence of caffeine on sleep and our biological clock?
- What is the influence of night shifts on our health?
- What is the influence of the other pillars (movement, nutrition, relaxation, mental health) on sleep?
- To what extent does our biological clock influence health and disease? And what is the role of the type of light (natural and artificial) on this?
- What is the best approach when you work night shifts? Is this universal or does it vary per person?

***Theme 2: Mental Health***

**Lecture 1**

- How do psychiatric medication and lifestyle interact?
- What is the influence of exercise on psychiatric disorders?
- What is the relationship between nutrition and mood?
- How can you use this information as a doctor in the clinic?

**Lecture 2**

- What is SOLK/ALK? (SOLK stands for "Somatisch Onvoldoende verklaarde Lichamelijke Klachten" in Dutch, which translates to "Somatically Unexplained Physical Complaints". ALK stands for "Aanpassingsstoornis met Lichamelijke Klachten", which translates to "Adjustment Disorder with Physical Complaints".)
- What is the relationship between emotion and SOLK/ALK complaints?
- How do you deal with patients with SOLK/ALK and what are the treatment options and advice?
- How can you as a doctor make use of this?
- What are the experiences of patients with these complaints?

***Theme 3: Nutrition***

**Lecture 1**

- What role does (highly) processed food play in disease?
- What concrete nutritional advice should you give in clinical practice?
- How do you deal with varying information on the internet about nutritional advice?
- What is your opinion about a healthy diet?

**Lecture 2**

- What is the connection between plant-based food and lifestyle and how can this be optimized for a healthy lifestyle?
- How can one transition to a plant-based diet?
- What are the ethical and medical-legal considerations regarding (plant-based) nutrition?

***Theme 4: Microbiome and food transition***

**Lecture 1**

- What is the microbiome and what is its function?
- What is the relationship between the microbiome, health, and disease?
- What are the effects of diet and lifestyle on the microbiome and are there lifestyle interventions that can improve the microbiome?
- What are recent scientific developments in the field of the microbiome?

**Lecture 2**

- No learning goals, because of a last-minute schedule change

***Theme 5: Drugs and addiction***

**Lecture 1**

- The student understands what addiction means both in practice as the working mechanism in the brain.
- The student understands how an addiction affects daily life.
- The student has insight into how sex and gender identity influence addiction.
- The student has insight into how to be careful with prescribing addictive substances between different sexes and gender identities.
- The student is able to discuss the topic of addiction with people of different sexes and gender identities.

**Lecture 2**

- The student understands the determinants of why people start vaping.
- The student recognizes the risk group that starts vaping.
- The student knows the dangers and physical harm to the body caused by vaping.
- The student is able to discuss the topic of vaping in the consultation room, also with the youth.
- The student is aware of the possible treatments to stop vaping.
- The student understands the urgency to stop the use of nicotine and vapes in the population.

***Theme 6: Young and old***

**Lecture 1**

- What are the first 1000 days?
- Which nutrition, lifestyle, and parenting factors influence development during the first 1000 days?
- What advice should be given to future parents in the consultation room, among others in the field of nutrition?

**Lecture 2**

- What is the effect of lifestyle on aging?
- How can lifestyle interventions ensure healthy aging?
- What are blue zones?
- What is the influence of the lifestyle of people living in blue zones on aging?

***Theme 7: Movement***

**Lecture 1**

- What is the effect of movement on health?
- What role does movement play in the development of diseases and disease prevention?
- What kind of lifestyle interventions are there in the field of movement, and how can these be implemented?

**Lecture 2**

- What is the importance of movement in aging?
- How does movement lead to better health outcomes and quality of life in the elderly?
- How can peer coaching be used for movement interventions and healthier aging?
- What do you advise patients/elderly in the consultation room about movement?
- What are the movement norms? Do people in the Netherlands move enough?
- What do you advise patients in the consultation room about movement?

***Theme 8: In practice***

**Lecture 1**

- How does behavioural change work?
- Why is behavior ('bad' habits) so difficult to change?
- Why do behavioural changes come to one patient and not to another?
- What makes lifestyle interventions effective or ineffective?
- What is the best approach for behavioural change for the individual? And from the government?

**Lecture 2**

- Bringing together all discussed topics and lifestyle pillars and concretizing with advice for practice.
- Do's and don'ts in the consultation room in the field of lifestyle.
- How can you as a doctor best influence the lifestyle behavior of patients?
- How do you deal with lifestyle problems in the consultation room?
- Cases from practice: good and bad examples.

## Appendix 2 – Practice test and demographical questions

***Practice test 1: Mental Health***

Welcome to the first quiz of the SELF-course!

The quiz consists of eight questions in total, of which four are multiple-choice questions and four are very short answer questions (VSAQs). For the VSAQs, you are expected to give a short answer (maximum of four words). This first test will be preceded by a number of questions about personal data. We would like to use this for the analyses of the research. If you agree with the informed consent (indicate below), your (anonymized) data will be used for our study. If you do not agree, your data will only be used for the course and scoring.

It is not allowed to use study materials while taking the test. Good luck!

***Demographical questions***

1. What is your age?
2. What is your gender?
    - o Female
    - o Male
    - o Non-binary

3. What do you study?
    - o Biomedical Sciences
    - o Medicine
    - o Clinical Technology
    - o Psychology
    - o Nursing
    - o Vitality and Aging
    - o I already finished my study, and work as a:
    - o Other, namely:

4. Which study year are you in?
    - o Bachelor year 1
    - o Bachelor year 2
    - o Bachelor year 3
    - o Master year 1
    - o Master year 2
    - o Master year 3
    - o Waiting period before clinical clerkships
    - o Not applicable

5. Do you participate in the honours program?
    - o Yes

- o No

6. Did you participate in previous SELF-courses or other courses related to prevention and life style?
   - o Yes, how many SELF- or orther lifestyle courses did you follow?
   - o No

***Version A***

**Very short answer questions**

*You have to answer these questions with a very short answer (one to four words)*

1. Which lifestyle factor has both a preventive protective effect on anxiety disorders and an efficacy in the treatment of anxiety disorders?
2. Which lifestyle intervention is mentioned as the first treatment step in the guideline for depressive disorders?
3. Which category of psychopharmaceuticals has the highest chance of obesity as a side effect?
4. What type of diet is associated with an increased risk of mental health issues?



**Multiple-choice questions**

1. In about 40% of the complaints, no clear physical cause is found in the general practice. In what percentage of cases do such complaints become chronic?
   - A. 10%
   - B. 30%
   - C. 50%
   - D. 70%

2. Pain is an important protective mechanism of the brain. What is it trying to protect you from?
   - A. Emotional stress
   - B. Physical damage
   - C. Physical fatigue
   - D. Expected danger

3. What is the first intervention for patients suspected of having SOLK/ALK?
   - A. Additional examination
   - B. Psycho-education
   - C. Lifestyle interventions
   - D. Referral to a specialist

4. There is a theory that suggests that the brain actively makes estimates about upcoming sensory input instead of passively registering it. This could play a role in the development of chronic pain complaints. What is this theory called?
   - A. Extended cognition
   - B. Neurophenomenology

C. Predictive coding
D. Unconscious inference

***Version B***

**Multiple-choice questions**

1. Which lifestyle factor has both a preventive protective effect on anxiety disorders and an efficacy in the treatment of anxiety disorders?
   A. Activity
   B. Sleep
   C. Quit smoking
   D. Diet

2. Which lifestyle intervention is mentioned as the first treatment step in the guideline for depressive disorders?
   A. Activity
   B. Sleep
   C. Quit smoking
   D. Diet

3. Which category of psychopharmaceuticals has the highest chance of obesity as a side effect?
   A. Antidepressants
   B. Antipsychotics
   C. Benzodiazepines
   D. Stimulants

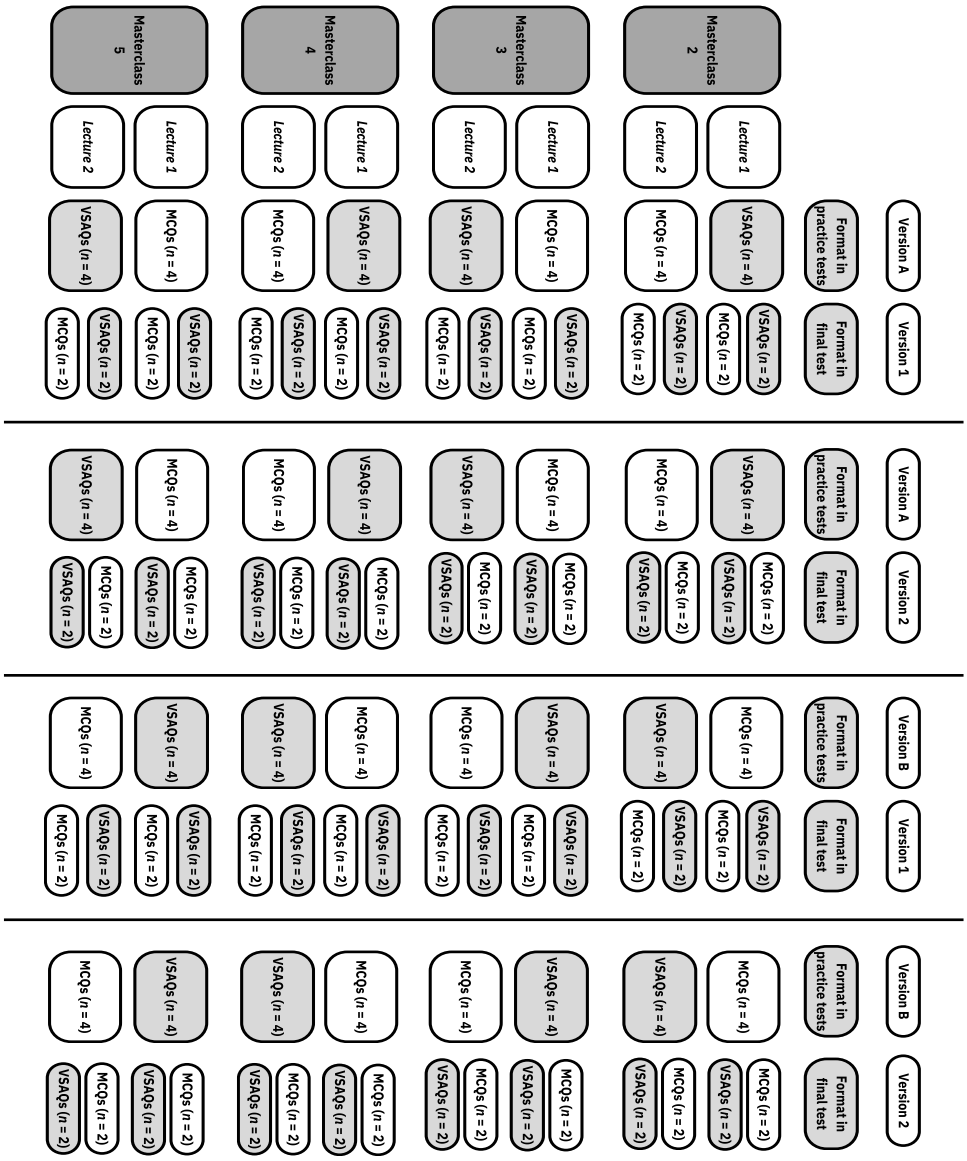
4. What type of diet is associated with an increased risk of mental health issues?
   A. Fermented food
   B. Highly processed food
   C. Herbs and spices
   D. Fat fish

**Very short answer questions**

*You have to answer these questions with a very short answer (one to four words)*

1. In about 40% of the complaints, no clear physical cause is found in the general practice. In what percentage of cases do such complaints become chronic?
2. Pain is an important protective mechanism of the brain. What is it trying to protect you from?
3. What is the first intervention for patients suspected of having SOLK/ALK?
4. There is a theory that suggests that the brain actively makes estimates about upcoming sensory input instead of passively registering it. This could play a role in the development of chronic pain complaints. What is this theory called?

## Appendix 3 – Supplemental Figure 1




**Supplemental Figure 1**. Detailed study design with versions A and B (practice tests) and versions 1 and 2 (final test)

## Appendix 4 – Questionnaire

***Before the final test:***

1. Have you prepared for the practice test? (yes or no). If yes, how? *<open answer>*
2. Have you prepared for the final test? (yes or no). If yes, how? *<open answer>*
3. Do you feel that you have better remembered the content of the lectures by taking the practice tests?
   - o Yes
   - o No
4. What percentage of the practice test about lecture 1 [title]; lecture 2 [title]; lecture 3 [title]; lecture 4 [title]; lecture 5 [title]; lecture 6 [title]; lecture 7 [title]; lecture 8 [title] do you think you have remembered?

***After the final test:***

5. How much of the material from this test do you think you will still know in 6 months? (*answer in percentage*)
6. Can we approach you to retake the same final test in 6 months?
   - o Yes
   - o No

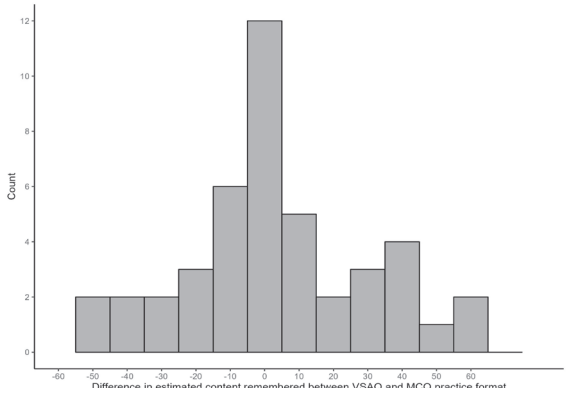
## Appendix 5 – Supplemental Table 1

**Supplemental Table 1**. Mean percentages correctly answered questions on the final test for the different question formats on the practice test and final test.

| Question format practice test | Question format final test | |
|---|---|---|
| | Multiple-choice question | Very short answer question |
| Multiple-choice question | 64.5% (3.1) | 49.7% (2.7) |
| Very short answer question | 57.5% (3.1) | 44.8% (3.3) |

Standard deviations in parentheses.

## Appendix 6 – Supplemental Figure 2




**Supplemental Figure 2**. The distribution of the difference in estimated content remembered practiced with VSAQs and MCQs. The percentages of lectures practiced with MCQs are subtracted from the percentages of lectures practiced with VSAQs.