

Direct Comparison of Online Tests Using Single-choice Items or Multiple-select Items in Pharmacology over One Year

Joachim Neumann¹, Stephanie Simmrodt¹, Beatrice Bader², Bertram Opitz², Ulrich Gergs^{1,*}

¹Institute for Pharmacology and Toxicology, Medical Faculty,
Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

²Center for multimedia-enhanced teaching and learning (@LLZ),
Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

*Corresponding author: ulrich.gergs@medizin.uni-halle.de

Received February 02, 2023; Revised March 05, 2023; Accepted March 13, 2023

Abstract One aim of testing medical students is to assess their level of knowledge. Another aim of examinations is that they can help to enhance the storage in memory for a longer time, at best, till they are in clinical practice. There is some debate what format of online examinations can be more useful to attain these aims. Here, in a formative online test, we compared single-answer items (SC) with multiple-select items (MS) for the very same learning objectives in a study of pharmacology in medical students. Medical students were randomly divided into two groups: group A was first given 15 SC (#1-15) followed by 15 different MS (#16-30). The opposite design was used for group B. One year later, four groups were formed from the previous two groups and were again given the same online test, but in a different order of questions. The main result was that all students fared better in the second test than in the initial test. But in the second test, students fared better when in the first test the same 15 items were asked in the MS mode than in the SC mode. In summary, we provide evidence that MS is better for knowledge retention over one year than SC. We speculate that MS is useful in formative tests to prepare students for any type of examination.

Keywords: single answer, multiple-select items, pharmacology, medical students

Cite This Article: Joachim Neumann, Stephanie Simmrodt, Beatrice Bader, Bertram Opitz, and Ulrich Gergs, "Direct Comparison of Online Tests Using Single-choice Items or Multiple-select Items in Pharmacology over One Year." *American Journal of Educational Research*, vol. 11, no. 3 (2023): 125-132. doi: 10.12691/education-11-3-4.

1. Introduction

Single best answers (single-answer or single-choice or SC, in which exactly one of, in the present context, five answers is known to hold true, generally called selective-response-formats) [1,2] are the classical tools used in medical examinations for medical students and postgraduate medical education in many countries. SC offers high levels of standardization, transparent marking, and the ability to test a wide range of knowledge in the available time. The SC format is easily electronically marked, stored and retrieved and is thus cost-effective.

However, there remain doubts whether SC offer the best option to encourage deep learning or whether SC simply leads to superficial learning or cramming [3,4]. Moreover, cueing is always a drawback in the SC format [5]. For instance, in our present study there is at least a 20% chance of guessing, because five answer options were given; in general, the chance of guessing will be $1/n \times 100$ in % if n -answer options are given and thus diminished if more and more answers are presented. A

hypothetically better way to assess knowledge lies in true multiple-answer questions in which one or more answers can be correct and the student is not aware of how many correct answers are to be anticipated (multiple-select items question format) [1,6,7]. The multiple-select items (MS) format belongs to the class of multiple response questions also known as selected response items, multiple mark items, or multiple-answer questions (for a review, see [2]). The MS format would not make guessing impossible but less probable: in the present study, for instance, four answer options were offered to each question and students were instructed that one, two, three or four correct answers were to be expected and no partial credit would be given. Hence, the guessing would be one in 15 possible answer options, reducing guessing to about 6.7%. Alternatively to MS, open questions would make guessing even more demanding but open questions require manually reviewing each test and are thus labor-intensive and not very easy to implement in many institutions, also in our hands [8,9,10,11]. In contrast to open questions, multiple-select item questions offer the advantage that they can be electronically marked and are thus cheaper to administer.

Here, we wanted to test both SC and multiple-select items question modes head-to-head in a formative online examination in pharmacology. Moreover, we asked the question of whether, one year later, these same students get better results if we exactly repeat these formative online tests (testing effect) in the very same mode (SC or MS). In addition, we asked whether multiple-select items or SC prepare examinees better for a subsequent multiple-select item exam or a SC exam format by enhancing retention or speculatively inducing transfer of knowledge. Our present goal of increasing in memory of students has been claimed to be a valid research topic and has been identified as a gap in the medical education literature (e.g., [12]). Clearly, testing is beneficial for the activation of the memory of students [13].

Thus, in the present work we wanted to directly compare the short- and long-term effects of two formats of formative online examination, namely SC and multiple-select items (MS), in pharmacology, immediately after first exposure of medical students to topics in pharmacology and one year later at the end of a three-semester pharmacology course. We also wanted to be able to decide which kind of questions are more beneficial to assess the knowledge of students at the time of examination and which kind of questions show a more pronounced testing effect. We hypothesized that the multiple-select items exams offer a better preparation for subsequent multiple-select items tests as well as SC tests. Our further goal was to improve the teaching mode for medical students, their competency in pharmacology and thus improve the treatment of future patients by these students once they have left medical school.

2. Methods

Medical students were offered eight days before the end of their fifth semester and the summative end of semester test in pharmacology (having been taught basic pharmacology in that semester) to partake in a formative online test (Figure 1A). We gave them the opportunity to get bonus points from the online test for their summative test (written SC format, under supervision, in lecture halls) eight days later. In order to pass the summative written test, students needed to provide the right answers in 60 % of the given questions. This was irrespective of any bonus points they might have earned. The questions in the summative test were not the same as in the online test. Only if they passed this 60 % threshold, the bonus points were used to improve their grades in pharmacology. These kind of bonus points are required in this context, at least in our institution, to achieve high participation rates of the eligible students in formative online tests as well as written formative tests, because in earlier studies without bonus points as few as 10 % of the eligible students took part and answered without motivation and little success [14,15]. This bonus point motivation was apparently successful because 200 students from 225 eligible students voluntarily entered the first online test (Figure 1A). Hence, bonus points and thus participation in the voluntary online test were not necessary to pass the summative test. Moreover, the original test takers cannot be identified by the data presented in this work and the students undergo

no conceivable risk by having their formative online test results published by the present study, because it is impossible to identify any individual student from the data presented in the present work as only mean values of test results are reported here. From this we concluded in accordance with the local Ethics Committee that participation in the online test was really voluntary and does not raise ethical issues. Moreover, the authors confirm that the work was carried out in accordance with the Declaration of Helsinki and the anonymity of participants was guaranteed. Students were not aware that they would be given the same questions one year later in the subsequent formative online test. Using this approach, we expected to assess via these questions directly their memorization of the learning objectives over the course of a year. Two groups (A and B) were randomly formed when the students checked into the learning environment (= ILIAS software, Figure 1A). Students were informed about the aims and purpose of the present study during lectures. They were instructed in the lectures that they would be offered in a random fashion, questions where exactly one of five answers (SC) were correct or that they were given four answer options to questions where one or up to four answers might be correct (MS). They were asked to inscribe electronically to one of two groups but were left unaware which group would offer which question format. By this approach, only very small differences in group size occurred which we regard as acceptable because large numbers of participants were reached (102 versus 98 in the groups A and B, respectively, Figure 1A). Using this random format, participants in group A ($n = 102$) were first given fifteen SC questions (i.e., one and only one out of five answers was right, see Figure 1B for sample questions) in pharmacology, followed by fifteen multiple-select questions (multiple-select items). In the multiple-select item format, four answers were presented (compare Figure 1B for sample format). This was done for the technical reasons that four answers were supported by the learning environment provided by our institution (ILIAS release 5.1.27) in the MS mode and five answers are required in the SC mode. The questions in the opposite order were given to group B ($n = 98$) (Figure 1A). Tests were taken by students at home without physical or electronic supervision by their educators, underscoring the voluntary nature of the whole present study. Ninety seconds at most in the mean were allocated to answer a test item. The software turned off after 22.5 minutes for the SC and multiple-select item sections such that the total online examination lasted 45 minutes. Students were technically not able to click at two answers or more in the SC format: this was forbidden by the software used.

The online examination took place at the very same time for all 200 students (5th of February 2019 online from 18:00 to 20:00 h). The re-test took place at 4th of February 2020 again from 18:00 to 20:00 h. Groups AA, AB, BA and BB (Figure 1A) comprised 43, 37, 33 and 35 students, respectively (for clarity also listed in Table 1). A clear advantage of online testing versus written testing might reside in the fact that total time taken to answer SC or multiple-select items questions for each student was automatically stored and thus could be easily analyzed in this study (Table 1).

A

$$SC_{\#1-15,16-30} = MS_{\#1-15,16-30}$$

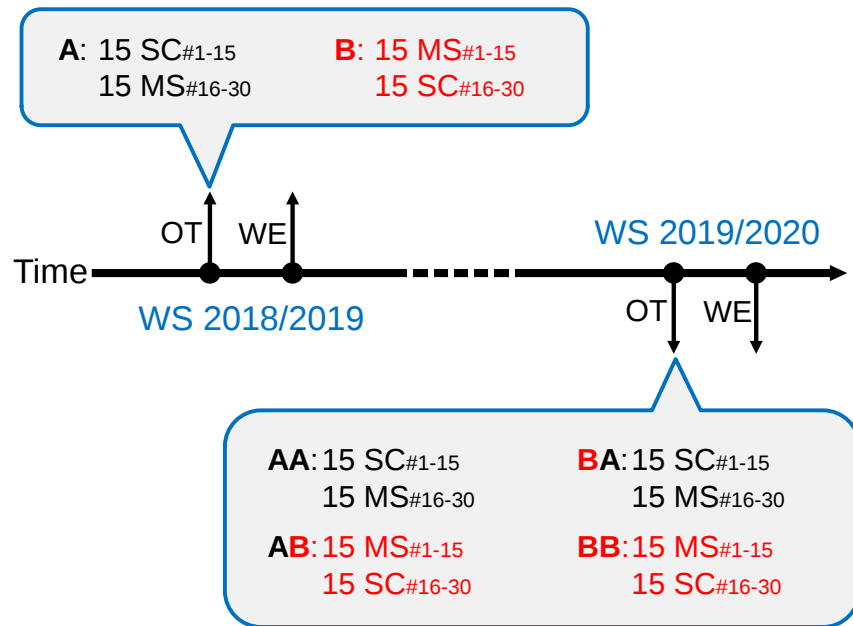
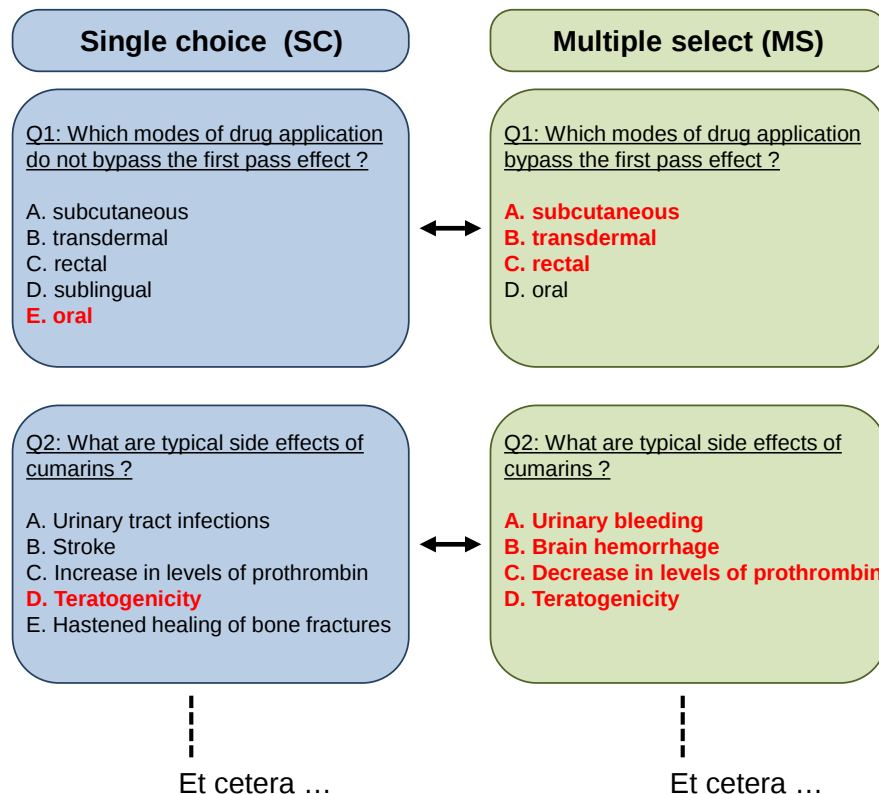
**B**

Figure 1. (A) The initial formative online test (OT) consisted of 30 online questions in the winter semester and took place in the February of the year 2019 (WS 2018/2019) and was given to medical students at the Medical School, Halle-Wittenberg (single answer or single choice items format: SC or multiple select items format: MS). Two groups, A and B, were randomly formed. Group A was online presented with 15 different single answer items format questions (labeled #1-15: SC#1-15) and then subsequently 15 questions in MS (labeled #16-30: MS#16-30) questions. Group B got the same questions but in opposite format (#1-15 = MS and #16-30 = SC). In the following winter semester in February of the year 2020 (WS 2019/2020), the same students were distributed in four groups (AA, AB, BA, and BB) as indicated and got the same questions but the order of questions was switched as indicated by the Latin letters of the group. WE, written obligatory exam. **(B)** Typical (in English translation: German and English version can be found in the supplement) questions with single choice items format (SC) and multiple-select items format (MS). Here, the correct answers are red-labeled

Table 1. Summarized test results of all the investigated groups. Students in group A were first given fifteen SC items (A-SC#1-15) followed by fifteen MS items (A-MS#16-30). The reverse order was given to students in group B (B-MS #1-15 followed by B-SC #16-30). For the retest one year later, four groups have been formed (compare Figure 1A)

Group	Number of students (n)	Mean points ± SD	Mean time (min:s)	Cronbach's alpha	*Significance P values=
A (total)	102	18.7 ± 5.5	32:31	0.86	0.1084 vs. B (total)
B (total)	98	17.5 ± 5.3	34:17	0.84	
A-SC #1-15	102	10.0 ± 2.6	n.d.	0.64	0.1170 vs AA-SC
A-MS #16-30	102	8.6 ± 3.5*	n.d.	0.82	< 0.0001 vs. A-SC
B-SC #16-30	98	10.9 ± 3.2*	n.d.	0.79	< 0.0001 vs. A-MS
B-MS #1-15	98	6.7 ± 2.6*	n.d.	0.59	< 0.0001 vs. B-SC < 0.0001 vs. A-MS
AA (total)	43	21.0 ± 6.5*	31:07	0.91	< 0.031 vs. A (total)
AB (total)	37	22.5 ± 4.7*	25:34	0.84	< 0.0001 vs. A (total)
BA (total)	33	22.5 ± 5.4*	26:39	0.88	< 0.0001 vs. B (total)
BB (total)	34	21.7 ± 4.8*	28:58	0.88	< 0.0001 vs. B (total)
AA-SC #1-15	43	10.8 ± 3.2	n.d.	n.d.	
AA-MS #16-30	43	10.1 ± 3.6*	n.d.	n.d.	< 0.0001 vs. AB-SC < 0.0001 vs. BB-SC
AB-SC #16-30	37	13.1 ± 2.4*	n.d.	n.d.	< 0.0001 vs. AA-SC
AB-MS #1-15	37	9.4 ± 2.7*	n.d.	n.d.	< 0.0001 vs. BA-SC
BA-SC #1-15	33	11.8 ± 3.0	n.d.	n.d.	
BA-MS #16-30	33	10.7 ± 2.7*	n.d.	n.d.	< 0.0001 vs. AB-SC
BB-SC #16-30	34	12.6 ± 2.4*	n.d.	n.d.	< 0.0001 vs. BA-MS
BB-MS #1-15	34	9.1 ± 3.0*	n.d.	n.d.	< 0.0001 vs. BA-SC

The content of the online tests included the learning objectives in lectures prior to the online test. Typically, two questions for each given lecture were constructed. The content stretched over the whole subject matter of basal and systematic pharmacology, covering topics such as pharmacodynamics, pharmacokinetics, autonomic pharmacology, antiarrhythmic drugs, drugs that lower blood pressure, and antibiotics (compare Figure 1B, and please see original questions in the original German version and the English translations in supplementary data files). All questions have been asked before in previous cohorts of medical students, but students were not made aware of that fact. The number of students who previously were tested with these questions amounted to 70 students (questions #1, 2, 3, 4, 5, 6, 7, 8, 9,10), 190 students (questions # 11,13, 14, 15, 16,17, 20, 21) and 239 students (questions # 12, 18, 19, 22, 23, 24, 25, 26, 27, 28, 29, 30) respectively. We selected questions with difficulty in a range of from 20.5 to 61.6 %. Discrimination coefficients ranged from 0.21 to 0.46.

The results of the online exam were made visible for the students that participated using the educational electronic environment Stud.IP. The students were listed on a spreadsheet, but we did not publish their names but as is the case in all the examinations within the Medical Faculty in Halle, we only listed the student number ("Matrikelnummer"). As such, no one from outside the Medical Faculty could read the result sheets, and even then, they could only read the student number but not the name of the student and likewise, the students did not see the names of their fellow students in the system.

2.1. Data Analysis

Mean values and standard deviations or standard errors of the mean were calculated using Microsoft Excel (2016,

from German distributor in Munich, Germany) and Student's t-tests, analysis of variance and post hoc tests (with Bonferroni correction) were performed with Graphpad Prism 5 (San Diego, California, USA). Variances between the groups were checked and did not differ substantially which supports the application of t-tests and analysis of variance. Cronbach's alpha was calculated with IBM SPSS Statistics for Windows, Version 25.0. (IBM Corporation, Armonk, NY, USA [purchased locally from IBM, Ehningen, Germany]). By convention, P-values smaller than 0.05 are regarded as significant. Graphs were prepared with Prism 5 and Microsoft PowerPoint (2016, from the German distributor in Munich, Germany).

3. Results

3.1. Initial Test (See Time Line in Figure 1A)

The mean times for taking the test (based on the software for the online test) are depicted in Table 1 and were not higher in group B than group A. The testing time distribution is depicted in Figure 2A-B and is much skewed to longer times. The total time spent on the questions was similar in group A and in group B (Table 1). In group A-SC, less time was required than in group A-MS (= multiple-select items), but more time was required in group B-SC than in group B-MS (Table 1). The total numbers of correct answers were similar in group A and in group B (Table 1). The total numbers of correct answers were similar in the SC format in group A and in group B (Table 1). Likewise, the total numbers of correct answers were similar in the SC format in group A (#1-15) and in group B (# 16-30) (Table 1). The distribution of the total points obtained (maximum 30 points) was depicted in

Figure 2C and Figure 2D. In group A (and in group B), the SC questions were answered better than the MS questions. Eight out of 15 questions were answered correctly more often in A-SC than in B-SC (Supplementary Figure S1C). Direct comparison, however, was possible for questions #1-15 in group A-SC and questions #16-30 in B-SC (plotted as bars in Supplementary Figure S1C). In a comparison (Supplementary Figure S1A) of the initial part of the test in groups A and B in nine questions (#1, #2, #3, #5, #6, #7, #8, #9, #12, #13, #15), the SC format led to more correct answers compared to the multiple-select items format, which is a significant difference (Table 1).

Comparing only the second half (Supplementary Figure S1B) of the questions (#16-30), questions #16, #19, #24, #25, #27, #28, #29, #30 were more often correctly

answered in the SC format) than in the multiple-select items format ($P < 0.001$). This indicates that for a given question, multiple-select item questions are usually but not always more difficult to answer than SC questions. In more detail, in direct comparison, all pairs, with the exception of #22, exhibited significantly higher scores in SC than in multiple-select items (Supplementary Figure S1A). The number of negative values for the discrimination index was three in group A and two in group B (Figure 2E). Moreover, negative discrimination indexes were seen in four questions in the multiple-select items format and once in the SC format (Figure 2E).

Cronbach's alpha-values, which measure reliability, were listed in Table 1 and are quite high in A and B and lower in the subgroups A-SC, A-MS, B-SC, B-MS.

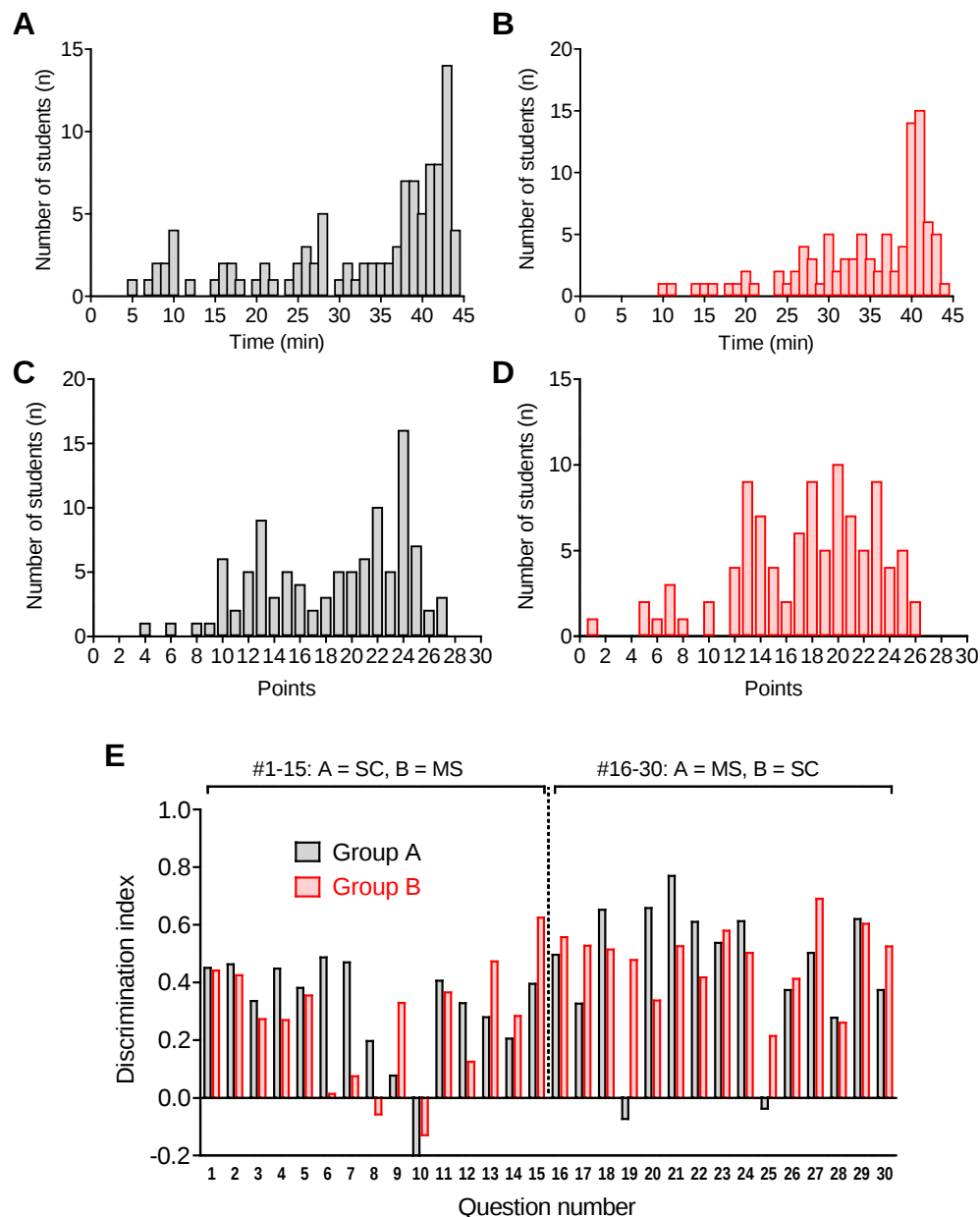


Figure 2. (A, B) Distribution of time required for tests in group A (A) and group B (B). The ordinates give the number of students who required the time (abscissae, time in minutes) for completion of all online questions. (C, D) Distribution of points reached in the online test in WS 2018/2019 for group A (C) and group B (D). The ordinates give the number of students who obtained the point scores (abscissae, points) in groups A and B. In (E), the item discrimination indices (ordinate in E) for the test questions (abscissa) in group A and group B are given as pairwise comparison

3.2. Second test (see time line in Figure 1A)

In direct comparison, mean points (correct answers) were always higher in the second test compared to the first test (Table 1), that is, in A compared to AA and AB or in B compared to BA and BB (Table 1). This can be clearly seen for each individual question comparing always the first two rows for each question in Supplementary Figure S2. Repeating SC or MS in the same order led to a gain of points (e.g., AA-SC vs. A-SC, see Table 1, Supplementary Figure S2 for individual questions). Likewise, total correct points in the initial online test (combining A and B) were surpassed in the second online test (combining AA and AB and BA and BB). In groups AB, BA, and BB (but not AA), MS was more difficult than the SC part in these groups (Table 1); group AA is a special case. There was no improvement in AA-SC versus A-SC and no difference between AA-SC and AA-MS. In all other subgroups in the second test, the performances in both SC and MS were higher than in the appropriate first test (Supplementary Figure S2). For instance, one might compare AB-SC and AB-MS with A-SC and A-MS (Table 1). Numerically, in BB-SC, the highest number of correct answers was obtained.

4. Discussion

4.1. Retrieval

The current study suggested that MS was better for knowledge retention than SC using pharmacological questions in medical students. The retrieval of knowledge, which we demand from a student by giving him or her a test, can affect later retention: retention of a fact was shown to be better if it is tested at all, compared to groups which have not been sitting an exam, and the benefit was independent of corrective feedback [16]. The testing effect has been reviewed in some detail in terms of its general importance in human psychology [17]. The same group of psychologists applied this re-test strategy more recently to medical students in their university and also found a benefit in the participants: their students fared better in subsequent written tests [12]. However, they did not use pharmacology questions in their work. Others have tested whether previous online MC tests enhanced knowledge retention for subsequent workshops (the didactic intervention) for board certified pediatricians, but not medical students [18]. Their control groups of pediatricians did not receive a previous MC test. After the workshops, both groups of pediatricians were given (online) the very same MC questions [18]. It was found that retention was better (measured as performance in the MC tests after the workshop) if a pre-test was done [18]. In contrast to our study where we used 15 MC tests, only five MC tests were given in the pre-test and only five different MC tests were given in the knowledge test period [18].

4.2. MS vs. SC

Direct comparison in the present work shows that also in pharmacology, SC questions are easier to answer than MS questions in agreement with the published literature

[14,19], but now in the subject matter of pharmacology. However, we also noticed MS prepares examinees better than SC for a subsequent test. Hence, it is tempting to speculate that MS is useful in formative tests to prepare students for any type of examination and also for clinical practice in pharmacology. To the best of our knowledge, this study is the first head-to-head comparison of the MS format to the SC format in pharmacology and the first to use MS to assess one-year memory of pharmacological questions. Repeatedly, it has been proposed to use MFT tests more often in education because this format is highly efficient and more reliable than SC [20,21]. Our data confirm this proposition.

4.4. Scoring Methods

Numerous scoring methods have been suggested to give partial credit in the MS format and are reported to increase reliability, which is per se in the MS format possibly higher than for the SC format [21-27]. Others, however, raised the concern that partial credits in MS, sacrifices validity to some extent [27] and hence was not used here. In the present study, in group A, Cronbach's alpha, a measure of reliability, was larger in the MS format than in the SC format, while in group B the opposite effect was observed (Table 1). This could be due to the exam question design or be a positional effect as in the second half of the test for both groups A and B Cronbach's alpha was larger. Cronbach's alpha increases if all 30 questions of group A and B are combined. This is reasonable as it is known that more questions in most tests will increase the reliability of the test.

4.5. Limitations

One can argue that the voluntary format of the formation of the A and B group might generate a systematic bias. This cannot be excluded. We would argue that at least in prior examinations (albeit using other pharmacology questions) the mean points achieved by both groups were not significantly different (data not shown). It might be informative in future studies, to devise an algorithm where half of the students choose their group at will and half are randomly chosen by the software: thereafter, one can follow four groups and find out whether they behaved differently in the subsequent intervention. Another limitation of our study lies in the fact that we could not rule out cheating in our students, as they were not supervised. However, the tests were given online for all students at the very same time. We cannot rule out that students used help from other students (texting) or looked answers up in textbooks or online. Nevertheless, if they did this, it would have been intellectually demanding and would translate to a classic open-book test: open-book tests have been well studied and are a useful assessment tool (for a review, see [28]). Even if students contacted each other or used electronic sources, the time allocated for each question (90 seconds) was too small and/or their motivation was somehow limited as their test results show wide variation between students and no student solved all items properly (Table 1, Figure 2C and Figure 2D). Moreover, one can ask why we did not use the same number of possible answers in the SC

(five answer options) or MS (four answer options). This was due to the pre-set option of the Stud.IP software we use at our institution and might be a drawback of the current study which should be avoided if more flexible software will be available to us in the future. Furthermore, it can be argued that the stems for the questions are not identical when using the SC-format compared to the MS-format. However, we could not figure out a different approach but to adapt the wording (see supplementary data files for original questions in direct comparison in both formats), this drawback might also be improved in future studies. Thus, a caveat is in order that in the present work the effect of re-testing is easier to interpret than the effect of different question formats.

4.6. Conclusions

In summary, we provide preliminary evidence that MS is better for knowledge retention than SC, which should change our way of testing medical students at least in formative if not summative tests.

Acknowledgements

We thank PD Dr. A. Aslan, Institute of Psychology, Martin-Luther University Halle-Wittenberg, Halle, Germany, for help with the study design and interpretation, Prof. Dr. A. Wienke, Institute for Medical Epidemiology, Biometrics and Informatics for help with statistical analysis, and Dr. iur. M. Struwe, Ethical Committee of the University Hospital Halle for advice in ethical issues.

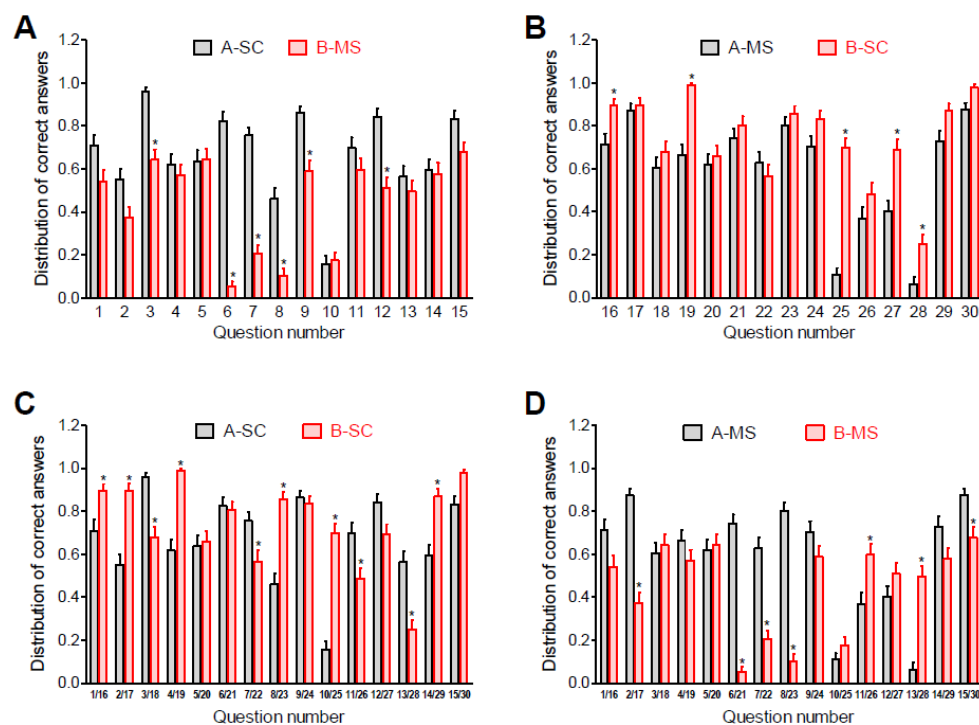
Statement of Competing Interests

The authors have no competing interests.

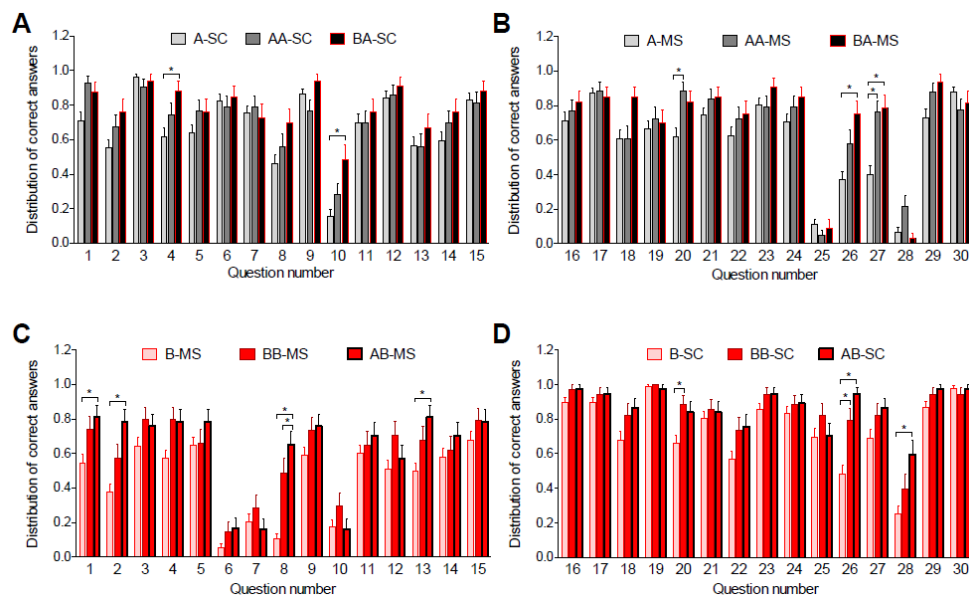
References

- [1] Desjardins, I., Touchie, C., Pugh, D., Wood, T. J., and Humphrey-Murto, S., "The impact of cueing on written examinations of clinical decision making: a case study," *Medical education*, 48 (3). 255-261. 2014.
- [2] Verbic, S., "Information value of multiple response questions," *Psihologija*, 45 (4). 467-485. 2012.
- [3] Elstein, A. S., "Beyond multiple-choice questions and essays: the need for a new way to assess clinical competence," *Academic medicine : journal of the Association of American Medical Colleges*, 68 (4). 244-249. 1993.
- [4] Veloski, J. J., Rabinowitz, H. K., Robeson, M. R., and Young, P. R., "Patients don't present with five choices: an alternative to multiple-choice tests in assessing physicians' competence," *Academic Medicine* [online], 74 (5). 539-546. 1999.
- [5] Schuwirth, L. W., van der Vleuten, C. P., and Donkers, H. H., "A closer look at cueing effects in multiple-choice questions," *Medical education*, 30 (1). 44-49. 1996.
- [6] Krebs, R., "The Swiss Way to Score Multiple True-False Items: Theoretical and Empirical Evidence," *Advances in Medical Education*, edited by A. J. J. A. Scherpbier, et al., Vol. 12, Springer Netherlands, Dordrecht, 1997, 158-161.
- [7] Krebs, R., "Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche Ausbildung," Bern, 2004.
- [8] Neumann, J., Gergs, U., Simmrodt, S., Aslan, A., and Teichert, H., "Direct comparison of very short answer versus single best answer questions for medical students in a pharmacology course," *Association for Medical Education in Europe, Vienna, Abstractbook*; Vol. 1068, 2019.
- [9] Sam, A. H., Hameed, S., Harris, J., and Meeran, K., "Validity of very short answer versus single best answer questions for undergraduate assessment," *BMC medical education*, 16 (1). 266. Oct. 2016.
- [10] Sam, A. H., Field, S. M., Collares, C. F., van der Vleuten, C. P. M., Wass, V. J., Melville, C., Harris, J., and Meeran, K., "Very-short-answer questions: reliability, discrimination and acceptability," *Medical education*, 52 (4). 447-455. Feb. 2018.
- [11] Sam, A. H., Fung, C. Y., Wilson, R. K., Peleva, E., Kluth, D. C., Lupton, M., Owen, D. R., Melville, C. R., and Meeran, K., "Using prescribing very short answer questions to identify sources of medication errors: a prospective study in two UK medical schools," *BMJ open*, 9 (7). e028863. Jul. 2019.
- [12] Wood, T., "Assessment not only drives learning, it may also help learning," *Medical education*, 43 (1). 5-6. 2009.
- [13] McConnell, M. M., St-Onge, C., and Young, M. E., "The benefits of testing for learning on later performance," *Advances in health sciences education : theory and practice*, 20 (2). 305-320. 2015.
- [14] Melzer, A., Gergs, U., Lukas, J., and Neumann, J., "Rating Scale Measures in Multiple-Choice Exams: Pilot Studies in Pharmacology," *Education Research International*, 2018 (1). 1-12. 2018.
- [15] Neumann, J., Simmrodt, S., Teichert, H., and Gergs, U., "Poster Abstracts, 21st Annual Meeting of the International Association of Medical Science Educators, Burlington, VT, USA, June 10-13, 2017: Problems when using online mock examination in pharmacology," *Medical Science Educator*, 27 (S1). 19-94. 2017.
- [16] Roediger, H. L., and Karpicke, J. D., "The Power of Testing Memory: Basic Research and Implications for Educational Practice," *Perspectives on psychological science : a journal of the Association for Psychological Science* 1 (3). 181-210. 2006.
- [17] Roediger, H. L., and Karpicke, J. D., "Test-enhanced learning: taking memory tests improves long-term retention," *Psychological science*, 17 (3). 249-255. 2006.
- [18] Feldman, M., Fernando, O., Wan, M., Martimianakis, M. A., and Kulasegaram, K., "Testing Test-Enhanced Continuing Medical Education: A Randomized Controlled Trial," *Academic medicine : journal of the Association of American Medical Colleges*, 93 (11S). S30-S36. 2018.
- [19] Newble, D. I., Baxter, A., and Elmslie, R. G., "A comparison of multiple-choice tests and free-response tests in examinations of clinical competence," *Medical education*, 13 (4). 263-268. 1979.
- [20] Frisbie, D. A., "The Multiple True-False Item Format: A Status Review," *Educational Measurement: Issues and Practice*, 11 (4). 21-26. 1992.
- [21] Tarasowa, D., and Auer, S., "Balanced Scoring Method for Multiple-mark Questions," *Proceedings of the 5th International Conference on Computer Supported Education*, SciTePress - Science and Technology Publications, 06.05.2013 - 08.05.2013, 411-416.
- [22] Albanese, M. A., and Sabers, D. L., "Multiple True-False Items: A Study of Interitem Correlations, Scoring Alternatives, and Reliability Estimation," *Journal of Educational Measurement*, 25 (2). 111-123. 1988.
- [23] Haladyna, T. M., "The Effectiveness of Several Multiple-Choice Formats," *Applied Measurement in Education*, 5 (1). 73-88. 1992.
- [24] Tsai, F.-J., and Suen, H. K., "A Brief Report on a Comparison of Six Scoring Methods for Multiple True-False Items," *Educational and Psychological Measurement*, 53 (2). 399-404. 1993.
- [25] Wu, B. C., *Scoring Multiple True False Items: A Comparison of Summed Scores and Response Pattern Scores at Item and Test Levels.*, Taiwan, 2003.
- [26] Bauer, D., Holzer, M., Kopp, V., and Fischer, M. R., "Pick-N multiple choice-exams: a comparison of scoring algorithms," *Advances in health sciences education : theory and practice*, 16 (2). 211-221. 2011.
- [27] Siddiqui, N. I., Bhavsar, V. H., Bhavsar, A. V., and Bose, S., "Contemplation on marking scheme for Type X multiple choice questions, and an illustration of a practically applicable scheme," *Indian journal of pharmacology*, 48 (2). 114-121. 2016.
- [28] Durning, S. J., Dong, T., Ratcliffe, T., Schuwirth, L., Artino, A. R., Boulet, J. R., and Eva, K., "Comparing Open-Book and Closed-Book Examinations: A Systematic Review," *Academic medicine : journal of the Association of American Medical Colleges*, 91 (4). 583-599. 2016.

Supplementary



Supplementary Figure S1. Distribution of correct responses to SC or MS questions #1-30. Mean values plus standard error of the mean are plotted in the ordinates. Variance analysis (ANOVA) revealed significant differences between all groups compared here ($P < 0.001$). By pairwise comparison (Bonferroni post-tests), differences between individual pairs of questions were analyzed and labeled in the graphs ($*P < 0.05$). (A) group A-SC compared to B-MS. This graph compares the correct answers for objectives #1-15 using both question formats. (B) group A-MS compared to B-SC. This graph compares the correct answers for objectives #16-30 using both question formats. (C) group A-SC compared to B-SC. (D) group A-MS compared to B-MS. The graphs in (C) and (D) compare the correct answers for the same question format but using different questions (e.g., #1 versus #16)



Supplementary Figure S2. Comparison of the distribution of correct answers between groups from WS 2018/2019 (first online test) and from WS 2019/2020 (second online test). Mean values plus standard error of the mean are plotted in the ordinates. Variance analysis (ANOVA) revealed significant differences between all groups compared here ($P < 0.001$). By pairwise comparison (Bonferroni post-tests), differences between individual pairs of questions were analyzed and labeled in the graphs ($*P < 0.05$). (A) A-SC versus AA-SC and BA-SC. (B) A-MS versus AA-MS and BA-MS. (C) B-MS versus BB-MS and AB-MS. (D) B-SC versus BB-SC and AB-SC.

