Hindawi
Education Research International
Volume 2021, Article ID 7767935, 10 pages
https://doi.org/10.1155/2021/7767935

Research Article

Comparison of Online Tests of Very Short Answer versus Single Best Answers for Medical Students in a Pharmacology Course over One Year

Joachim Neumann,[1] Stephanie Simmrodt,[1] Holger Teichert,[2] and Ulrich Gergs[1]

[1]*Institute for Pharmacology and Toxicology, Medical Faculty, Martin Luther University Halle-Wittenberg, 06097 Halle (Saale), Germany*
[2]*LLZ, Martin Luther University Halle-Wittenberg, 06097 Halle (Saale), Germany*

Correspondence should be addressed to Ulrich Gergs; ulrich.gergs@medizin.uni-halle.de

Received 27 February 2020; Revised 17 June 2020; Accepted 30 December 2020; Published 12 January 2021

Academic Editor: Christos Troussas

Single best answers (single choice, SC) are the classical tools used in medical examinations on all levels of education. In contrast, very short answer (VSA) probably requires sound prior knowledge and deeper learning than SC, and VSA should make cueing and guessing impossible. Hence, in a basic pharmacology course, we wanted to compare the SC and VSA formats directly at the end of the course and one year later. Medical students ($n = 211$) were given a formative online test. Two groups were randomly formed (A and B). Participants in group A were first given fifteen single choices (one out of five) pharmacology questions and thereafter fifteen very short answer questions (open question which were to be answered online and semiautomatically assessed). Very similar questions with regard to learning objectives but in opposite order were given to group B. After one year, about half of students from group A were again given the very same questions (AA) or the opposite questions (AB). Likewise, group B was again tested with the opposite questions (BA) or the same (BB). The SC questions in groups A, AA, AB, B, BA, and BB were in sum easier to answer than the corresponding VSA questions. Repeating the test after one year with the same students increased retention of right answers by about 1.5 points. In summary, direct comparison questions in the VSA format are more difficult for our students to answer than questions in the SC format, conceivably because cueing and guessing are eliminated. Knowledge retention is present by repeating the very same examination format online. Retention of knowledge is higher when starting with VSA (group B) both for a subsequent SC format or a VSA format. These data would argue for more use of the VSA format at least in pharmacology examinations.

1. Introduction

Single best answers (single choice, SC, generally called selective response formats cf. [1]) are the classical tools used in medical examination on all levels. SC has been well studied and evaluated to be fair, fast, quite inexpensive to give, and reliable; they are felt to be easily defensible, and in many studies, they were valid. However, SC formats certainly do not mirror the clinical situation with a certain patient where, for instance, there is not the choice of just five different drugs but a dozen drugs have to be chosen from [2, 3]. Cueing is another drawback of the SC format. Moreover, 20% of the right answers in SC can simply be guessed, if one answer of five answers is known to be correct. In addition, the reliability of the SC format heavily depends on the face validity of the distractors (the four wrong answers if five answers are given). Poor distractors make guessing and cueing even easier such that up to 50% of the right answers can be guessed. Another way to assess knowledge is an essay (generally called constructed response format cf. [1]). Various kinds of essays have been used. One possibility is to use short answers requiring one or three words as answer(s) (very short answers, VSA). Very short answers intuitively require sound prior knowledge and deeper learning, and the VSA format rules out cueing and guessing [1, 4]. Moreover, VSA does not require the quest for strong distractors. Hence,

one has argued that VSA is faster and easier to write for the medical educators than SC formats [5, 6]. Theoretically, SC offers retrieval cues, the right answer is given as an option (target information), and success of the test taker depends as much on familiarity with the content as on active memory retrieval [7]. They have shown using a short biology text in psychology students that VSA requires text processing and really understanding the new given text, whereas SC could be answered mainly based on prior biology high school knowledge [7]. There is the assumption that VSA in a formative test should drive students to deep learning, whereas SC would lead more to superficial or strategical learning [8]. Most educators would probably want to use tests that foster deep learning in prospective test takers. Another option to deepen learning that has been studied is adaptive testing with computers and has been used to provide personalized test questions. Others had tested clinical diagnosis in family medicine comparing SC and VSA (long list of possible answers was provided, uncued questions: abbreviated UnQ) but on paper sheets [3]. Basically, they reported a high correlation for the same topic in results of both formats: SC and UnQ. They noted that UnQ were always more difficult and retention after one year was better [3]. In our university in Halle, there is a longstanding interest in learning and forgetting, as Hermann Ebbinghaus, a founder of experimental psychology, taught and performed research here in Halle on memory loss and retention [9]. Others have made the point, that we find plausible, that testing and the testing effect in medical education should be employed because this "forced" retrieval stabilizes or improves the storage of the information in the brain, and this facilitates on the clinical ward, and the active retrieval will be needed at the job [10]. Production tests (such as VSA) lead to better retention over time, for instance, one month, than recognition tests (SC format) because VSA needs more effort in retrieval of information from memory than SC [9, 11, 12]. Tests in medical education can not only be regarded as neutral tools of measurement but also as tools to facilitate acquisition and retention of knowledge [10].

Here, we wanted to test both SC and VSA formats head to head in formative online examinations in basic pharmacology. Moreover, we asked the question whether one year later, these same students get better results if we repeat exactly these online test questions. In addition, we asked whether VSA or SC formats prepare better for a subsequent VSA or SC format. This increase in memory is a valid research topic and has been suggested to be looked at in more detail in medical education [10, 13, 14]. It is relevant also to test how good we prepare students for SC because most medical licensure exams have SC format (and not VSA).

Hence, our research questions (QR) were as follows: QR1, Is the SC format always easier for students than the VSA in basic pharmacology items? QR 2, Does the VSA format offers a better preparation for subsequent VSA as well as for the SC format in basic pharmacology? Preliminary results of this study have been published in abstract form [15].

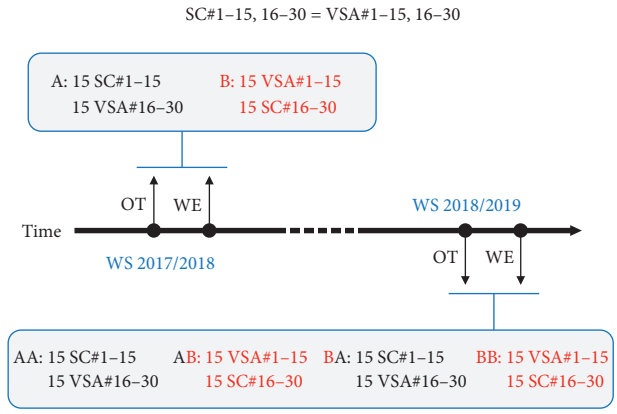

Figure 1: Schematic drawing illustrating the study design. The initial formative online test (OT) contained 30 online questions in the winter semester 2017/2018 (WS 2017/2018) in medical students at the Medical School, Halle-Wittenberg (single choice, SC or very short answer, VSA). Two groups A and B were randomly formed. Group A was online presented with 15 different SCs (labeled SC#1–15) and then subsequently 15 VSAs (labeled VSA#16–30). In group B, pairs in reverse order were formed: first 15 VSAs (labeled then VSA #1–15) and then subsequently 15 SCs (labeled then SC#16–30). In the winter semester 2018/2019 (WS 2018/2019), the same students got the same questions in different combinations as indicated. WE, written obligatory exam.

2. Materials and Methods

Medical students ($n = 211$) at the final week of their 5th semester (having been taught basic pharmacology for the first time in that semester by means of lectures) were given a formative online test in February 2018 (Figure 1). The number of students fits well to recent work on size requirements in assessment studies [16]. We gave students the opportunity to get bonus points from the online test for their summative test a week later. This was done because in a previous study [17], we noted that voluntary online tests were only taken by 20% of eligible students without a bonus program and by over 90% with a bonus program. Here, 87.6% (211) students from 241 eligible students entered the first online study, and 77.5% (186) from 240 eligible students took part at the second online test one year later. The same observations hold true in formative written tests, and it may therefore be generalized that a bonus system is always necessary for student participation rates of formative tests in tertiary education [18]. Students were not aware that the very same questions would be given one year later (February 2019) in the online test because in previous years, different questions were provided in the second online test. This was done to ensure comparability and to offer an easy way to test for keeping of their knowledge or even measure an improvement of their pharmacological knowledge. We expected an increase in knowledge in the second online test as in the subsequent sixth and seventh semester of Medical School, the students were exposed more deeply to pharmacology by taking part in seminars on basic and clinical pharmacology and lectures in clinical pharmacology.

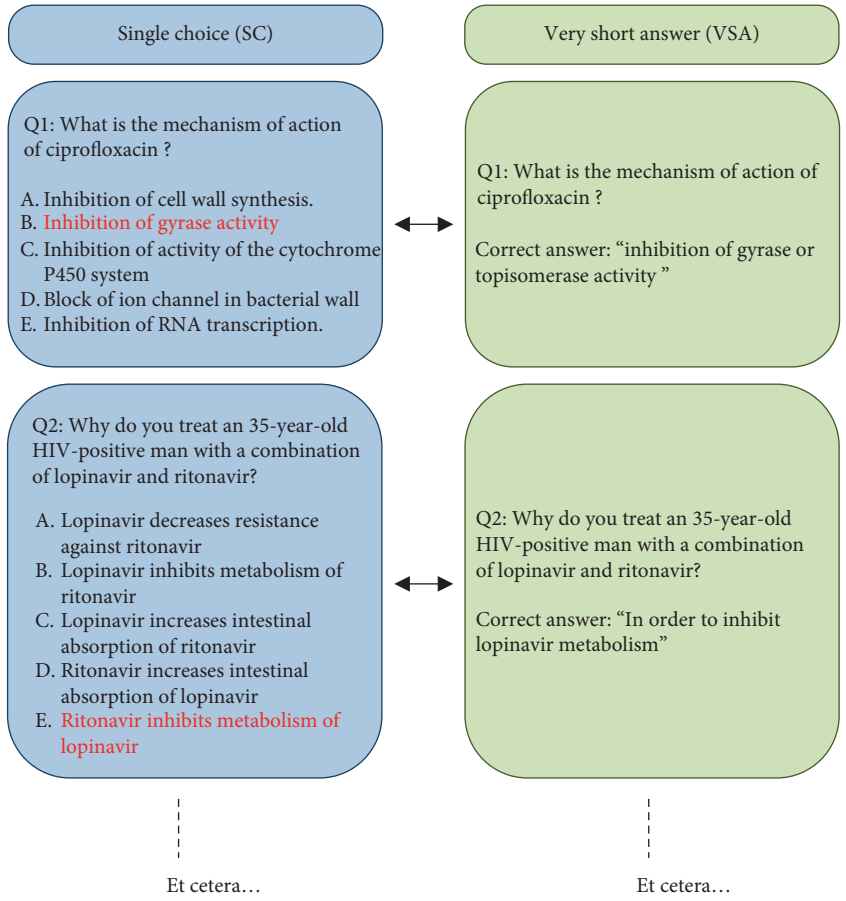

Figure 2: Two typical (in English translation from German) questions for single choice (SC) and very short answer (VSA).

Two groups were randomly formed by the software ILIAS 5.1 after checking in into the learning environment software (A and B, Figure 1 top). Participants in group A were first given fifteen single choice (i.e., one and only one out of five answers was right, SC; see Figure 2 for sample questions) questions in basic pharmacology and thereafter fifteen very short answer (VSA) questions; in other words, open questions were to be answered online and were semiautomatically assessed by JN; minor misspellings and valid alternative drugs were accepted, and typical right answers were stored in the software (ILIAS and StudIP), by the designers of the tests and sped up correction of the test one year later. For 211 or 186 students who participated in the present study, manual assessment of VSA was required each year (2018 or 2019) about three hours.

Sometimes, more than one right answer was possible for VSA (sample questions in Figure 2). The questions in opposite order were given to group B (Figure 1). Tests were taken by students at their home without supervision. 90 seconds at most in the mean were allocated to answer a test item. This was done because in earlier investigations, we noted that more time per item allocated dramatically increased the percentage of right answers; we have previously allowed six hours for the online test (different questions from the present tests), and nearly, all students answered rightly, whereas with the same test questions allowing only 90 seconds, only about 60% of the test items were answered properly [17]. If all questions were answered rightly, no further statistical analysis would have been possible, and therefore, the present protocol was used. One clear advantage of online testing resides in the fact that total time for SC or VSA for each student is stored and thus could additionally be analyzed in this study.

Content of the online tests was the learning objectives in lectures on basic pharmacology prior to the online test. Typically, two questions for each lecture were constructed. The content stretched over the whole subject matter of basal and systematic pharmacology, namely, pharmacodynamics, pharmacokinetics, antidiabetic drugs, drugs to reduce and stimulate thyroid function, antihypertensive drugs, sympatholytic drugs, drugs that inhibit the renin and angiotensin system, positive inotropic drugs, antiarrhythmic drugs, antibiotics, antiviral drugs, antimycotic drugs, cytostatic drugs, and drugs that inhibit thrombin function or coagulation.

2.1. Data Analysis. Mean values and SEM were calculated using Microsoft Excel 2016. For parametric and nonparametric tests, SPSS 25 (IBM, Ehningen, Germany) was used. A probability value (p value) less than 0.05 was regarded as significant. The ability of individual questions to discriminate the good from the marginal (bad) student was done by

Table 1: Test results of all groups of students.

Group	Number of students (n)	Mean points ± SD	Mean time (min:s)	Cronbach's alpha	Significance ($p < 0.05$)
A (total)	96	20.7 ± 4.1	36:25	0.650	
B (total)	115	21.3 ± 4.8^{a}	34:25	0.677	[a]vs. A (total)
A + B (total)	211	21.0 ± 4.4			
A (SC) $_{\#1-15}$	96	13.1 ± 1.9	17:02	0.603	
A (VSA) $_{\#16-30}$	89	8.1 ± 2.4^{b}	$19{:}12^{b}$	0.556	[b]vs. A (SC)
B (SC) $_{\#16-30}$	112	9.8 ± 2.9	17:42	0.610	
B (VSA) $_{\#1-15}$	115	$11.6 \pm 2.2^{c,d}$	$16{:}43^{c}$	0.499	[c]vs. B (SC) [d]vs. A (VSA)
AA (total)	44	20.9 ± 4.6	40:56		
AB (total)	42	21.2 ± 4.5	38:40		
BA (total)	44	22.1 ± 5.3	37:45		
BB (total)	52	$23.0 \pm 4.6^{e,f}$	37:46		[e]vs. AA (total) [f]vs. B (total)
AA + AB + BA + BB (total)	182	22.5 ± 3.8			
AA (SC) $_{\#1-15}$	44	12.4 ± 2.0^{b}	17:10		[b]vs. A (SC)
AA (VSA) $_{\#16-30}$	42	9.0 ± 2.4^{g}	$23{:}32^{g}$		[g]vs. AA (SC)
AB (SC) $_{\#16-30}$	41	10.5 ± 2.9	19:20		
AB (VSA) $_{\#1-15}$	42	10.9 ± 2.4	19:24		
BA (SC) $_{\#1-15}$	44	12.8 ± 2.2^{b}	16:48		[b]vs. A (SC)
BA (VSA) $_{\#16-30}$	42	$9.7 \pm 3.4^{d,h}$	$20{:}56^{i}$		[d] vs. A (VSA) [h]vs. BA (SC)
BB (SC) $_{\#16-30}$	49	11.7 ± 2.5^{c}	19:03		[c]vs. B (SC)
BB (VSA) $_{\#1-15}$	52	11.5 ± 2.1	18:09		

convention by use of a discrimination index calculated with the point-biserial correlation between students' scores on a particular item and their total test scores. The discrimination index can fall between −1 and +1, and a discrimination index larger than 0.3 indicates that the best students were able to answer that test item correctly, but the weaker ones were not [19]. Coefficient alpha reliability estimates (Cronbach's alpha) were calculated as a measure of the precision of the students' scores. The coefficient alpha can range from 0 to 1, with higher values indicating more precise student scores. A value close to 1 indicates that each student would be likely to achieve about the same score if a similar test was readministered, whereas a value of 0 indicates that scores on retesting would vary randomly [19].

3. Results

The mean times for taking tests are depicted in Table 1. All students finished the test in the given time frame. Testing times were not normally distributed using the Kolmogorov–Smirnov test in SPSS ($p > 0.05$). For instance, for the test in 2018, the frequency distribution of times is plotted in Figure 3. For convenience, we nevertheless report here the mean testing time as arithmetic mean (Table 1). The total mean testing time was longer in group B compared to group A (Table 1, the nonparametric test: Mann–Whitney test, $p < 0.05$). The numbers of correct answers in all groups (A or B, Figure S1) and subgroups (AA, AB, BA, BB, Figure S2) were likewise not normally distributed, but by convention and for convenience, we present here nevertheless arithmetic means, but used for comparing nonparametric tests (Table 1). The total numbers of correct answers were different between SC test items within group A and within group B (Figure 4, $p < 0.05$). Likewise, the numbers of correct answers in the VSA group were significantly different within groups A, B, AA, AB, BA, and BB (ANOVA for each $p < 0.05$). Moreover, VSA questions collectively were more difficult in group A as well as in group B compared to the SC format (Table 1). Moreover, combining all correct answers, results were better comparing all SCs and all VSAs in groups A and B as well as AA, AB, BA, and BB.

Interestingly, at least two effects were occurring: when SC was given initially (group A), the results were better ($p < 0.05$) than in students who were given SC in the second half of the test (group B). Moreover, SC tests gave higher marks compared to VSA tests ($p < 0.05$) (Table 1). Some learning objectives were often correctly answered in A and B in both formats; in at least one question, hardly any right answers (2%) were obtained in the very short answer format, whereas the companion single best question was correctly answered by nearly all students (96%) (Figure 4).

Cronbach's alpha measures that reliability is always higher in the SC format (A-AC or B-SC) than in the VSA format (A-VSA and B-VSA; Table 1). Total Cronbach's alpha is higher for group A or B compared to its subgroups (ASC; A-VSA; or B-SC and B-VSA) which is probably due to the higher number of items (30 versus 15, Table 1). Mean item discrimination (Figure S5) was better (higher) in SC than in VSA in initial tests (groups A and B) as in subsequent tests (AA, AB, BA, and BB).

In direct comparison, mean points (correct answers) in BB were lower than in AA, whereas no additional differences gained significance (Table 1). Interestingly, BB total was better than the initial test with the same design (Table 1 and Figures S1 and S2), namely, group B.

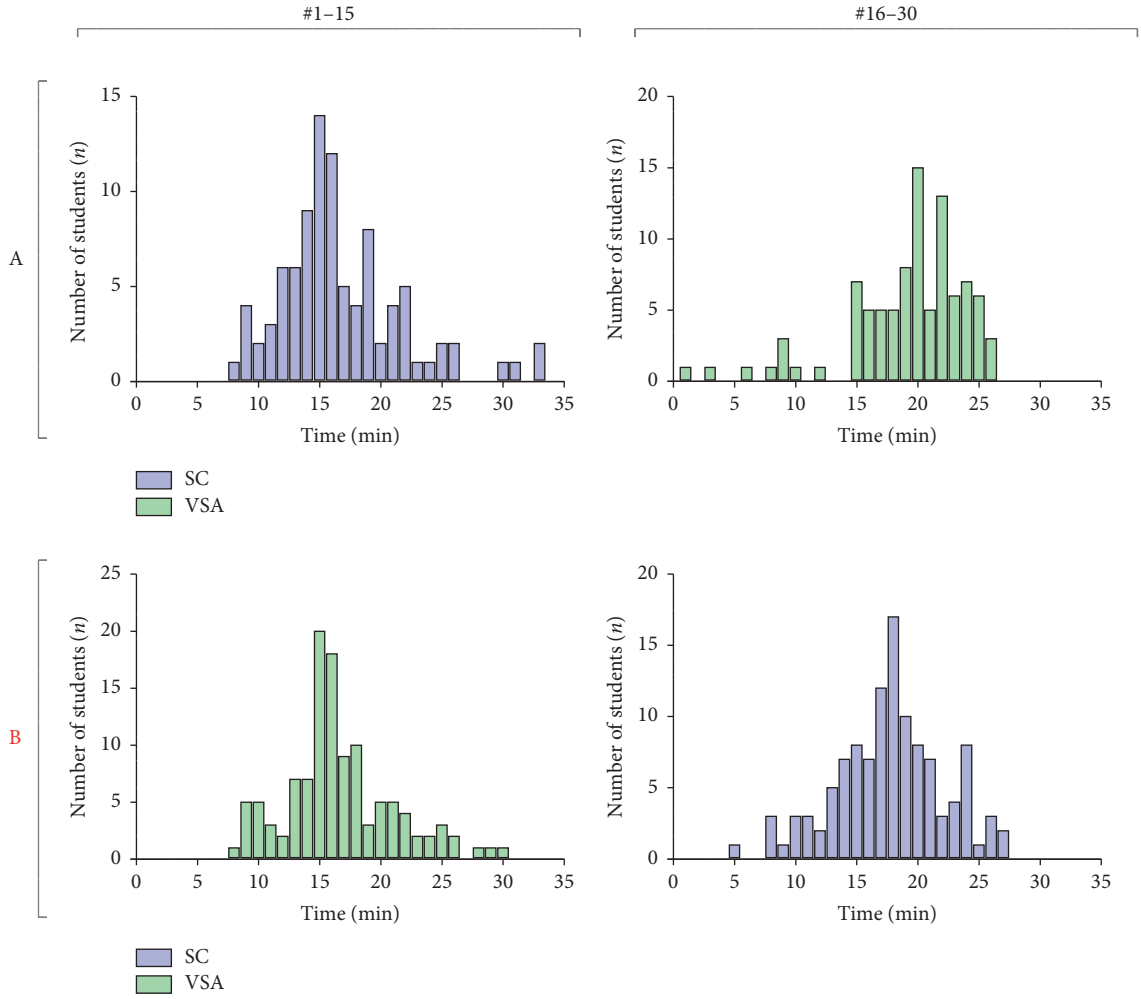

Figure 3: Distribution of times for working on a test in group A-SC (#1–15), in group A-VSA (#16–30), in group B-VSA (#1–15), and in group B-SC (#16–30).

Total 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) and, namely, amounted to 21 points and 22.5 points (correctly answered questions), an increase in 1.5 points which can be interpreted as an effect size of about 0.1, indicative of a small effect size. Interestingly, total time spent on the questions was higher in total group A (36 : 25) than in total group B (34 : 25), though more points were obtained in total group B. Looking more closely on the format of question, it turns out that in group A-SC, less time was required than in B-SC, but much more time was required in group A-VSA than in group B-VSA (Table 1). A possible interpretation might be that VSA is so demanding that at the end of the test, it takes more time to muster the questions than at the beginning of a test. Likewise, in the second online test, the VSA part always takes longer than the SC part (e.g., AA-SC and AA-VSA in Table 1). Group BB is a special case: there is no improvement in BB-VSA versus B-VSA; perhaps, the students were already at their maximum of performance.

Figure 4 depicts the direct comparison between the same learning objectives (#1–#15), initially answered as SC (A-SC) or as VSA (B-VSA), and Supplementary Figure S3 depicts the direct comparison between the different learning objectives (#1–#15 versus #16–#30) of the very same group. SC was in sum easier to answer than VSA, but the paired differences are not large with the exception of question #2. Indeed, in question #2 (a question for treatment of drug intoxications) in the SC format, it was simply asked for which combination of drugs their functional antagonist would take longest to act; in the VSA, it was explicitly asked that at what time vitamin K would at least need to counteract an overdosage of phenprocoumon (accepted answer, 24 hours or more). In questions #1, #2, #3, #4, #5, #6, #8, #9, #10, #11, #12, #13, and #15, there were significant differences between pairs: from these pairs, A-SC had more points than B-VSA in pairs #1, #2, #3, #4, #6, #8, #9, #10, #11, #13, and #15. In pairs, B-VSA gave more points only in questions #5 and #12 than A-SC, possibly because the answer options distracted the students in the appropriate SC format. In detail, #5 in the VSA format was asking about the main life-threatening side effect of thiamazole (agranulocytosis) and #12, what is the effect of atropine on conduction velocity in the AV node (increase in velocity). There were no significant

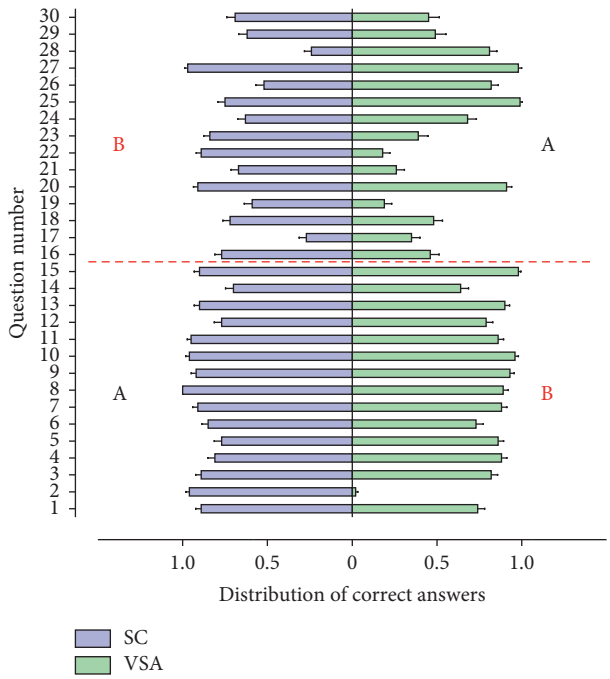

Figure 4: Distribution of correct answers in groups A and B in the winter semester 2017/2018. When single best answer questions (SC) were provided first (group A-SC#1–15), the results were better ($p < 0.05$) than in students that were given SC questions last (group B-SC#16–30). The same phenomenon was noted for very short answer questions (B-VSA#1–15 versus A-VSA#16–30).

differences between pairs #7 (how would you treat diabetes in pregnancy? insulin) and #14 (which bronchodilator acts via phosphodiesterase inhibition? theophylline).

Direct comparison of the two formats (asked for in the second half of the first test, Figure 1) for the remaining learning objectives (#16–30) is shown in Figure 4. Here, in seven (#16, #18, #19, #21, #22, #23, and #30) of 15 questions in the SC format (B-SC), students obtained more points than in the VSA format (A-VSA). In five pairs (questions #17, #20, #24, #27, and #29), no differences between pairs was noted. This lack of differences can in some cases be explained by the fact that most students could answer the questions properly, and they were apparently too easy (#20 (what is the mechanism of action of acyclovir? inhibition of nucleic acid synthesis) and #27 (what is the mechanism of action of ciprofloxacin? inhibition of gyrase or topoisomerase), Figure 4). However, in questions #17, #24, and #29, the questions were of intermediate difficulty, and no reason for a lack of differences is apparent. This could be due to the learning objectives (#16–#30) or exhaustion of students in the second half of this test. Somewhat unexpected, questions #25 (which coagulation factors are decreased after phenprocoumon treatment? II, VII, IX, X), #26 (which protein is activated by heparin? antithrombin III), and #28 (name one penicillin derivative worth trying against β-lactams-forming staphylococci? oxacillin or flucloxacillin or dicloxacillin) were much more often rightly answered in the VSA format than the SC format, probably because the distractors were too strong. In Figure 4, we plotted for group A the questions in both halves (A-SC and A-VSA) that occur in the same subsequent order: for instance, the first question #1 is seen next to #16, the first question of the VSA format. The objectives in, e.g., #1 and #16 are different; however, it is apparent that many questions in the subsequent format obtained lower scores than in the first format (Figure 4). This seems to contrast group B where the first format is often not better than the second format asked at the same relative position (Figure 4).

After one year, the second test (Figure 1) was given and is summarized in Supplementary Figures S4 and S5. There is no clear-cut picture for SC; in Figure S4A, it is obvious that there is even a small decline in knowledge in question #3 and in increase in question #13. With the VSA format, there is after one year a clear decrease in knowledge which was in paired comparison often significant (questions declined #16, #19, #21, #25#, 27, #29, and #30, Figure S4C). Only in question #17, there was an increase in score; one year later, the others remained unchanged. So the VSA format given after the SC format in the same test (group A, Figure 1) does not prepare well for a VSA format given as the second half again one year later (AA-VSA). This might be due to exhaustion or the different learning objectives in #1–#15 in comparison to #16–#30.

A higher knowledge was detected in the VSA format if first SC format and then the VSA format were given (group A), and after one year, one starts with SC (group BA) for the learning objectives which were initially given as VSA format (group B). Here, the very same questions are in sum better answered, present for many pairs of questions (Figure S4D). There is a substantial significant gain in knowledge in questions #2 and #13, whereas the others remained unchanged. Apparently, initial VSA (group B) prepares well for another SC format (group BA).

However, VSA seems over all, also to prepare well for VSA (Figure S4B). The BA group is better in questions #5, #7, #19, #11, #12, #13, #14, and #15 but worse in questions # 4, #6, and #8 compared to A-VSA for the same questions, whereas there was no difference in #1, #2, and #3.

The order of questions seems to be important; if the SC format is given as second format (Figure 1), this is a good preparation for second answers to SC (Figure S4G); questions #17 and #28 are answered better in the test one year later (BB-SC) than in the previous year (B-SC), whereas no differences were noted in the other questions. This lack of increase is at least for #20 and #27 due conceivably to the question being too easy in the first place. Nevertheless, the overall results are better for questions #1–#15 compared to #16–#30 independent of the format (Figures 4 and S5). Therefore, the questions #16–#30 may be more difficult than the questions #1–#15.

The VSA format does prepare over a one-year interval in principle well for VSA, if VSA is given first (Figure S4E). Here, as a whole, a gain in knowledge after one year was an apparent. Specifically, a gain in question #2 but a decrease in #4 and no change in the remaining question pairs was noticed. Again, part of a lack of increase (but not the decline in knowledge) could result from giving too easy questions.

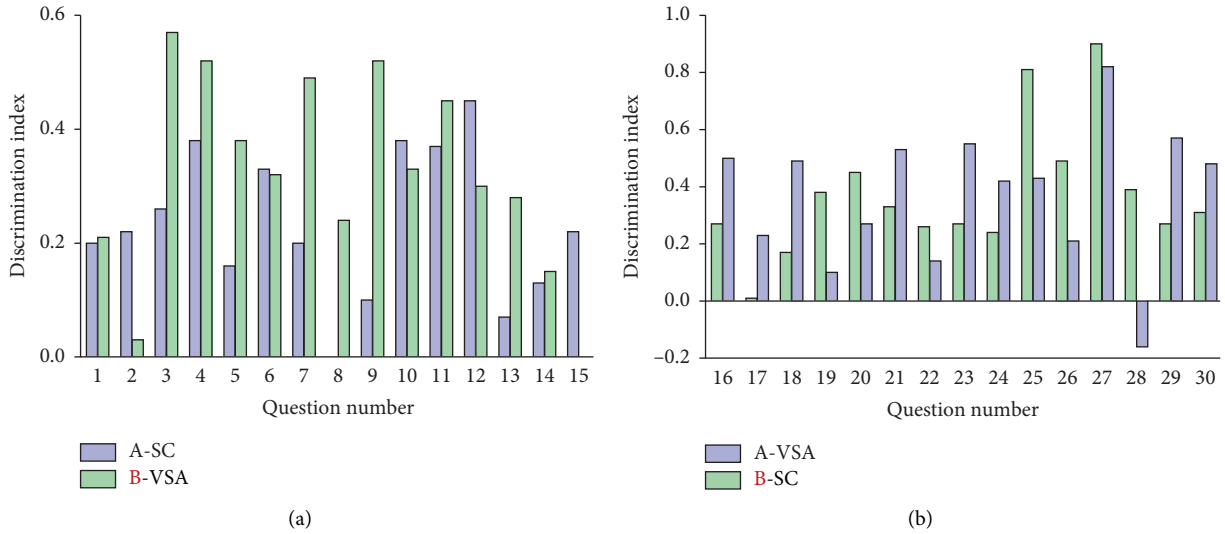

Figure 5: Comparison of discrimination indices of SC versus VSA questions for questions #1–15 (a) and for questions #16–30 (b).

Giving VSA format initially, and then second, the SC format first, leads to results not hugely different from first giving SC and then VSA and one year later, SC second (AB-SC). Here, higher scores were noted only in questions #17 and #28 (Figure S4H). If initially SC is given and after one year VSA (AB-VSA, Figure S4F), there was only a higher score one year later in questions #7 and #14, whereas others were not different or even lower in questions #1 and #4, one year later (Figure S4F).

In Figure 3, it is nicely notable how much longer time is required for VSA when starting with SC (Figure 3, group A). This seems due to the position not the format of the question as the reverse is seen in Figure 3 group B. The distribution function for group A or B and all subgroups (A-SC, A-VSA, B-SC, and B-VSA) is shown in Figure S1. We speculate that A-SC is easy to answer, but A-VSA is more demanding as high scores are hardly reached (Figure S1). In contrast, in group B, it seems to be reversed: VSA questions are more properly answered than SC (Figure S1). Discrimination indices are sometimes different between VSA format and the SC format (Figure 5). In Figure 5(a) but not in Figure 5(b), the discrimination indices for VSA are significantly higher than in SC (χ^2 test 0.05). In all paired questions, the number of the discrimination index is higher in VSA than SC with one telling exception; question #2, here as mentioned above, it was crucial to know that only the active metabolite of clopidogrel inhibits clot formation. This fact was unknown to nearly all test takers (see above), and therefore, it does not discriminate but gives random numbers.

The discrimination indices for the exams in 2019 gave similar results (data not shown). Usually, the discrimination indices are higher for AA-VSA than for AA-SC. A similar pattern appears if VSA is given first and then SC for BB. The differences in discrimination indices between questions in groups AB and BB were significant (χ^2 test $p < 0.05$) but not for groups AB and BA. In contrast, there is no clear pattern discernible. For some questions, VSA shows higher discrimination indices and for some questions lower. One can speculate that in initial questions, SC discriminate better, but as attenuation and exhaustion sets in, VSA discriminates better. Except questions #13 and #14 which are very difficult in the VSA format and were answered properly by less than 10% of students, this might easily explain their good discrimination indices.

4. Discussion

We find it noteworthy that the gain in knowledge by giving the questions twice is astonishingly small; in an earlier formative study [17], we gave medical students the very same written SC test of 30 questions (18 right answers were the passing grade) at the beginning of the course in basic pharmacology, and in a subsequent summative examination (15 weeks later), this translates into the finding that the mean value in the pretest amounted to 11.5 points and the mean points in the final exam were 25.8, an increase by 14.3 points or by 124%. We speculated that the gain in the written test was higher due to the more stressful environment: written paper examination, supervision by teachers, all students in one lecture hall, all students sitting with an empty chair in between, no cooperation of students allowed, no interfering, and visual interference allowed [17]. In contrast, here the online tests were taking place at home, without supervision, and therefore, the examination was in all probability less stressful. Another explanation would be that the first online test was given after 14 weeks in a basic pharmacology course; hence, students started with better results and were not so emotionally challenged by the prospect of failure than the students preparing for a summative examination. The time allocated cannot explain any difference because it was the same in all tests (in mean 90 seconds per question). On the other hand, others noted comparably small gains in student's

knowledge even in smaller interval; clinical students in intensive care rotation were given the same questions initially and four weeks later; the 32 participants experienced an increase in exam points from baseline (65.7) by 4.6 points [20] abstract. Others used the subsequent design in psychology students; students were asked to read a text on a computer screen; then, students were first given nine open ended (VSA) questions on that same text, and then, the same students were subjected to nine MCQ (one SC out of four) tests; here, SC was easier than VSA and Cronbach's alpha was higher in VSA than in SC [7].

What practical conclusions can we draw from this work? Written exams are probably much superior as a formative tool than online tools as concerns gain of knowledge. However, usually resources are scarce, and thus, administrators prefer less costly solutions. Hence, there is a small benefit of online test which should be seized in clinical education. Others have tried new tools to improve mobile learning such as adaptive tests on smartphones. It might be relevant for a subsequent study to allot students more time for the initial online test because an increase in time drastically increased the number of properly answered tests (for SC) [17]. One year later on, one could give the same questions but only allow 90 seconds to answer. On the other hand, the testing effect might be relevant under these conditions; it is possible that the knowledge of an impending second tests improved retention. Others likewise noted that questions that required a response, such as VSA, led to better scores one year later in a mock medical licensure exam than SC format questions [21].

The time for the same learning objectives, given in two different formats, shows that more time was required when taking first SC and then VSA (group A) question formats. In comparison to VSA and then SC (group B), here, two interpretations are possible: the learning objectives were more difficult in group A than in group B, or VSA prepares the mind to answer SC faster. This finding is reproducible as this difference was also present when taking the same test one year later (compare group A and group AA or group B and group BB; Figure S4). Students who switch to a different format behave differently: switching from group A to AB made the task conceivably more difficult as a longer time was used by students to answer questions (A vs. AB). However, switching from an initial VSA to a subsequent VSA led to less time demands (B vs. BA) and might be interpreted that VSA prepares for both VSA and MCQ, which is not unreasonable. A drawback of VSA is the cost for manually reviewing the answers. There has been much progress in automatic assessment of VSA (reviewed in [22]) which may increase cost effectiveness of this approach.

Our work confirms and extends the work from [6, 23, 24]. Medical students always answered—in our nomenclature—SC better than VSA questions [6, 23, 24]. However, they assessed all clinical learning objectives (Table 1) in [23], whereas we focused here on basic pharmacology. In a subsequent study, Sam et al. [6] used an approach similar to ours; in the first group of students, they gave electronically 60 SCs and then on paper, 60 VSAs for the same learning objectives, and in the second group, in the reverse order. We noted in contrast to Sam et al. [23] that in some questions, VSA fared similar to SC. This divergent finding could be in part explained by the different protocol; we used two groups and started in one group with SC and the other with VSA, while Sam et al. [23] used only one group and gave first VSA then SC. In this way, students might have anticipated the proper answer in SC from the previous VSA (this interpretation is already discussed in [6]). This has previously been noted by [4]. A better consistent response to SC than to VSA has been already described using paper examinations for both formats [25]. They noted that the more incompetent students gained most from the SC type format and were easily identified by the VSA type format. Using this approach, they noted that differences between the results on matched questions in SC and VSA declined with an increase in the clinical competence, less difference in registrars than in medical students [25]. The general observation that SC is easier and is of benefit for the incompetent is not novel but dates back to Hurd (Hurd, 1932). However, Sam et al. [6, 23, 24] did not perform a follow-up online assessment in the subsequent academic year like we did in the present work. We asked 30 different questions, whereas [23] gave only 15 questions (and gave 60 questions in [6]) first as VSA and then the same learning objectives as SC. In another study, they compared drug prescribing skills in students [24]. They wanted to prepare students for the PSA (Prescribing Safety Assessment) exam in Britain. Hence, they mimicked that PSA exam by asking the students online to give the concrete drugs, dose, route of application, and relevant drug interaction for 50 different clinical scenarios using a VSA style answer (2.5 min given per question). In addition, they asked in a SC type question to decide which of five drugs to apply in this scenario (1 min per question given) [24]. They noted that students always fared better in SC than in VSA [24]. They concluded that VSA is more reliable and discriminates better than SC. In addition, they make the valid point that wrong answers in VSA can be used by educators to better understand where misconception prevail which should be addressed during the remaining time in medical school or at least the next intake of medical students by changing the curriculum [19] [24]. This view was shared by students because in a survey of students who undertook the study, Sam et al. [6] reported that the students regarded VSA to mimic more closely the clinical reality than SC. Many years before, students already had also deemed VSA to be more useful to prepare for the ward than SC, at the time paper examinations were studied [25]. In a clinical setting (OSCE examinations), similar results were obtained: students performed better in written SC than in written VSA; when in one group, the same learning objectives were first tested as SC and then as VSA and in the second group vice versa [1]. Students themselves rated the SC format as easier as the VSA format [1].

Comparing the groups B-VSA and BB-VSA (Figure 1, S4, and Table 1), there is no gain in correct answers; this might mean that the students were already performing at their best. However, they were still far away from answering all questions correctly, casting doubt on the hypothesis that VSA prepares best for VSA. When giving VSA in the second

half of the test, there is a gain in correct answers from 7.73 (A-VSA) to 10.1 (AA-VSA, Table 1), supporting a learning effect in VSA (if one starts from a low level). Surprisingly, SC repetition decreases the number of correct answer from 13.2 (A-SC) to 11.8 (AA-SC, Table 1). This might indicate that SC is not the perfect way to increase strong memorization of learning objectives but is prone to loss of knowledge. If this is repeated by other independent groups, this would deeply question our current use of SC in medical education.

One limitation of our study lies in the fact that we could not rule out cheating in our students. However, the tests were given online for all students at the very same time. Mean time allocated for each test item was only 90 seconds. We cannot rule out that they used help from other students (texting), looked results up in textbooks, or online. Nevertheless, if they did this, it was very demanding for them and would translate to a classical open book test, which has been well studied and is a useful assessment tool (review [27]). In addition, the time allocated was too small or their motivation was limited as their test results show wide variation between students, and no student solved all items properly. Hence, the test was discriminatory, as offered. As concerns of motivation of students in mobile testing, an interesting new invention is the use of games to pose tests to students.

It is not without precedence that open questions can be better answered if the subject matter fits than the SC format; examples are given in [5]. Others noted that better scores were obtained in a final MCQ examination if previously, the learning objectives had been tested in a SC format than a simple context-free recall MCQ format [21], supporting our data. However, those researchers [21] only studied questions dealing with social sciences (population health, legal questions, organizational questions, and ethical issues); whereas, our novel contribution to the field is the concentration solely on pharmacology questions and the longer time scale (one month [21] versus one year (this study)). Others detected a higher correlation of SC interim questions to SC final examination than between uncued questions (similar to VSA) and final SC final examination; this interpreted that SC prepares better for SC than VSA [19]. Uncued questions had a wider range than MCQ and were attributed to marginal knowledge of marginal students [19] which we also note for our data. They also noted that median values of discrimination indices were higher for uncued questions versus MCQ similar to our findings. These data may further indicate that VSA formats are more useful than multiple choice tests for identifying marginal students who are in need of special assistance and tutoring, as suggested by others [19]. One concern when using VSA in formative or summative tests in a particular medical school might be that students are poorly prepared for national board examinations that use the SC format. This concern is not borne out in the literature. At least comparing in one medical school in the USA in subsequent academic years the final nationwide results (USMLE), these grades were not worse for students prepared with uncued questions versus MCQ [19].

Online formative examinations given at home to detect an increase in knowledge have been published by others in medical education; for instance, pediatricians were subjected to (online) SC format tests to enhance knowledge retention for subsequent seminars for pediatricians [28]. After the workshops, participants were given (online) the very same MC questions [28]. It turned out, retention was better (measured as performance in the MC test after the workshop) if a pretest was performed in comparison to the control group [28]. In contrast to our study, only five SC questions were used in the pretest and five different SC questions were given in the knowledge test period [28]. A difference to the present study is that we used a more prolonged interval between two tests (one year) than earlier work (in the psychological lab or in medical students in the clinic) where typically, a delay between two tests of a few minutes up to one week was used (reviewed in [29, 30]).

In summary, we can answer the research questions (QR) in the following way: QR1, direct comparison shows that in basic pharmacology, SC is easier than VSA. QR2, VSA prepares better than SC for a subsequent test. We speculate that VSA is useful in formative tests to prepare students for any type of examination and also for clinical practice in pharmacology.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

J.N. designed the research; S.S., U.G., and H.T, performed research; S.S., J.N., H.T, and U.G. analyzed data; U.G., and J.N. wrote the manuscript.

Acknowledgments

The authors thank PD Dr. Alp Aslan, Institute of Psychology, Martin-Luther-University Halle-Wittenberg, Halle, Germany, for help in design and interpretation of the study.

Supplementary Materials

Figure S1. Distribution of points in the first online test in WS 2017/2018 for groups A, A-SC, A-VSA, B, B-SC, and B-VSA. Figure S2. Distribution of points in the second online test in WS 2018/2019 for groups AA-SC, AA-VSA, BA-SC, BA-VSA, AB-SC, AB-VSA, BB-SC, and BB-VSA. Figure S3. Pairwise comparison of distribution of correct answers in group A-SC versus A-VSA (A) and group B-VSA versus B-SC (B) in the winter semester 2017/2018. Figure S4. Pairwise comparison of the distribution of correct answers between groups from WS 2017/2018 (first online test) and from WS 2018/2019 (second online test). (A) A-SC versus AA-SC; (B) A-SC versus BA-SC; (C) A-VSA versus AA-VSA; (D) A-VSA versus BA-VSA; (E) B-VSA versus BB-VSA; (F) B-VSA versus AB-VSA; (G) B-SC versus BB-SC;

and (H) B-SC versus AB-SC. Figure S5. Comparison of the distribution of correct answers between all groups from WS 2018/2019 (second online test). (A) Comparison for questions 1–15; (B) comparison for questions 16–30. (*Supplementary Materials*)

References

[1] I. Desjardins, C. Touchie, D. Pugh, T. J. Wood, and S. Humphrey-Murto, "The impact of cueing on written examinations of clinical decision making: a case study," *Medical Education*, vol. 48, no. 3, pp. 255–261, 2014.

[2] A. S. Elstein, "Beyond multiple-choice questions and essays," *Academic Medicine*, vol. 68, no. 4, pp. 244–249, 1993.

[3] J. J. Veloski, H. K. Rabinowitz, M. R. Robeson, and P. R. Young, "Patients don't present with five choices," *Academic Medicine*, vol. 74, no. 5, pp. 539–546, 1999.

[4] L. W. T. Schuwirth, C. P. M. Vleuten, and H. H. L. M. Donkers, "A closer look at cueing effects in multiple-choice questions," *Medical Education*, vol. 30, no. 1, pp. 44–49, 1996.

[5] I. Damjanov, B. A. Fenderson, J. J. Veloski, and E. Rubin, "Testing of medical students with open-ended, uncued questions," *Human Pathology*, vol. 26, no. 4, pp. 362–365, 1995.

[6] A. H. Sam, S. M. Field, C. F. Collares et al., "Very-short-answer questions: reliability, discrimination and acceptability," *Medical Education*, vol. 52, no. 4, pp. 447–455, 2018.

[7] Y. Ozuru, S. Briner, C. A. Kurby, and D. S. McNamara, "Comparing comprehension measured by multiple-choice and open-ended questions," *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, vol. 67, no. 3, pp. 215–227, 2013.

[8] D. I. Newble and N. J. Entwistle, "Learning styles and approaches: implications for medical education," *Medical Education*, vol. 20, no. 3, pp. 162–175, 1986.

[9] H. Ebbinghaus, *Über das Gedächtnis: Untersuchungen zur Experimentellen Psychologie*, Duncker & Humblot, Berlin, Germany, 1885.

[10] D. P. Larsen, A. C. Butler, and H. L. Roediger III, "Test-enhanced learning in medical education," *Medical Education*, vol. 42, no. 10, pp. 959–966, 2008.

[11] A. C. Butler and H. L. Roediger III, "Testing improves long-term retention in a simulated classroom setting," *European Journal of Cognitive Psychology*, vol. 19, no. 4-5, pp. 514–527, 2007.

[12] M. A. McDaniel, H. L. Roediger, and K. B. McDermott, "Generalizing test-enhanced learning from the laboratory to the classroom," *Psychonomic Bulletin & Review*, vol. 14, no. 2, pp. 200–206, 2007.

[13] D. P. Kulasegaram, A. C. Butler, and H. L. Roediger III, "Repeated testing improves long-term retention relative to repeated study: a randomised controlled trial," *Medical Education*, vol. 43, no. 12, pp. 1174–1181, 2009.

[14] T. Wood, "Assessment not only drives learning, it may also help learning," *Medical Education*, vol. 43, no. 1, pp. 5-6, 2009.

[15] J. Neumann, U. Gergs, S. Simmrodt, A. Aslan, and H. Teichert, *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, Austria, 2019.

[16] A.-S. Aubin, M. Young, K. Eva, and C. St-Onge, "Examinee cohort size and item analysis guidelines for health professions education programs," *Academic Medicine*, vol. 95, no. 1, p. 151, 2020.

[17] J. Neumann, S. Simmrodt, H. Teichert, and U. Gergs, "Problems when using online mock examination in pharmacology," *Medical Science Educator*, vol. 27, no. 1, p. S54, 2017.

[18] A. Melzer, U. Gergs, J. Neumann, and J. Lukas, "Rating scale measures in multiple-choice exams: pilot studies in pharmacology," *Education Research International*, vol. 2018, Article ID 8615746, 12 pages, 2018.

[19] B. A. Fenderson, I. Damjanov, M. R. Robeson, J. J. Veloski, and E. Rubin, "The virtues of extended matching and uncued tests as alternatives to multiple choice questions," *Human pathology*, vol. 28, no. 5, pp. 526–532, 1997.

[20] D. Piquette, R. Brydges, A. Goffi, C. Lee, B. Mema, and C. Walsh, *Assessing Competency of Subspecialty Residents in Critical Care Clinical Reasoning: Validity Evidence in Support of the Script Concordance Test*, Association for Medical Education in Europe, Basel, Switzerland, 2018.

[21] M. M. McConnell, C. St-Onge, and M. E. Young, "The benefits of testing for learning on later performance," *Advances in Health Sciences Education*, vol. 20, no. 2, pp. 305–320, 2015.

[22] S. Burrows, I. Gurevych, and B. Stein, "The eras and trends of automatic short answer grading," *International Journal of Artificial Intelligence in Education*, vol. 25, no. 1, pp. 60–117, 2015.

[23] A. H. Sam, S. Hameed, J. Harris, and K. Meeran, "Validity of very short answer versus single best answer questions for undergraduate assessment," *BMC Medical Education*, vol. 16, no. 1, p. 266, 2016.

[24] A. H. Sam, C. Y. Fung, R. K. Wilson et al., "Using prescribing very short answer questions to identify sources of medication errors: a prospective study in two UK medical schools," *BMJ Open*, vol. 9, no. 7, Article ID e028863, 2019.

[25] D. I. Newble, A. Baxter, and R. G. Elmslie, "A comparison of multiple-choice tests and free-response tests in examinations of clinical competence," *Medical Education*, vol. 13, no. 4, pp. 263–268, 1979.

[26] W. Hurd, "Comparisons of short answer and multiple choice tests covering identical subject content," *The Journal of Educational Research*, vol. 26, no. 1, pp. 28–30, 1932.

[27] S. J. Durning, T. Dong, T. Ratcliffe et al., "Comparing open-book and closed-book examinations," *Academic Medicine*, vol. 91, no. 4, pp. 583–599, 2016.

[28] M. Feldman, O. Fernando, M. Wan, M. A. Martimianakis, and K. Kulasegaram, "Testing test-enhanced continuing medical education," *Academic Medicine*, vol. 93, no. 11S, pp. S30–S36, 2018.

[29] H. L. Roediger and J. D. Karpicke, "The power of testing memory: basic research and implications for educational practice," *Perspectives on Psychological Science*, vol. 1, no. 3, pp. 181–210, 2006.

[30] H. L. Roediger and J. D. Karpicke, "Test-enhanced learning," *Psychological Science*, vol. 17, no. 3, pp. 249–255, 2006b.