Research Article

Rating Scale Measures in Multiple-Choice Exams: Pilot Studies in Pharmacology

Andreas Melzer,¹ Ulrich Gergs ,² Josef Lukas,¹ and Joachim Neumann ²

¹*Institute of Psychology, Martin Luther University Halle-Wittenberg, 06097 Halle, Germany*

²*Institute for Pharmacology and Toxicology, Medical Faculty, Martin Luther University Halle-Wittenberg, 06097 Halle, Germany*

Correspondence should be addressed to Joachim Neumann; joachim.neumann@medizin.uni-halle.de

Received 19 March 2018; Revised 1 June 2018; Accepted 6 June 2018; Published 10 July 2018

Academic Editor: Aleksander Aristovnik

Copyright © 2018 Andreas Melzer et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Multiple-choice questions are widely used in clinical education. Usually, the students have to mark the one and only correct answer from a set of five alternatives. Here, in a voluntary exam, at the end of an obligatory pharmacology exam, we tested a format where more than one alternative could be correct ($N = 544$ students from three year groups). Moreover, the students were asked to rate each item. The students were unaware how many correct answers were contained in the questions. Finally, a questionnaire had to be filled out about the difficulty of the new tests compared to the one out of five tests. In the obligatory final exam, all groups performed similarly. From the results, we conclude that the new rating scales were a better challenge and could be adapted to assess student knowledge and confidence in more depth than previous multiple-choice questions.

1. Introduction

Written examinations using multiple-choice questions (MCQ) are now widely used all over the world in medical education. These tests were developed in psychology for research purposes by E. L. Thorndike around 1900 [1]. Frederick J. Kelly was, likely, the first to use such items as part of a large-scale assessment for educational purposes [1]. Probably, the first massive use of MCQ occurred in the US armed forces during the First World War [2]. MCQ offered the possibility to assess the recruits for combat use with limited involvement of human resources. Later in the 1950s [3], the national board of examiners in the USA, based on military experiences, used these tests to assess medical students and foreign medical doctors.

MCQ-based examinations have the advantages of being fast, cheap, objective, and of making it possible to test a broad sampling of the curriculum. They can be given in a very standardized form and can have high statistical validity. Ideally, tests should also be sensitive and specific. MCQ-based examinations are now often electronically marked, and the results can then be mathematically processed to give the examiners and the students a quantitative feedback on the quality, reliability, and validity of the tests. A concern in this

test format was always how valid it really is. In other words, does it test the skill and knowledge used in clinical practice? Is it sensitive enough to differentiate which students are fit for clinical practice and which should be kept out of practice? Hence, the national board of medical examiners in the USA and also other testing institution as well as medical faculties have continuously altered, adapted, and tried to improve their multiple-choice formats.

There is controversy whether MCQ only assess lower levels of knowledge like recall of isolated facts, encourage trivialization of knowledge, or lend to rote learning [4]. One main concern, when using these MCQ-based examinations, is whether really knowledge is assessed or other patterns like structure of the personality, the experience with the test format (“test sophistication” [5]), or the decision to take an educated guess. Others support MCQ and argue that better constructed MCQ can measure and perhaps improve problem-solving skills in examinees [6, 7]. Moreover, it has been claimed (in an US introductory science course) that there is a gender bias in answering MCQ [8]. Possibly, females are less likely to guess than males [9] or exhibit less “test wiseness” [10]. In the NBME (national board of medical examiners) part I (basic medical sciences) in the USA, men (number of examinees: 7234) had

FIGURE 1: Systematization of question types.

significantly higher grades than women (number of examinees: 4090). The relative difference in scores was 5.9% in total. Of interest, here, the mean score of men on pharmacology was 496 (standard deviation: 113) and in women, it was 469 (standard deviation: 107); hence, they were significantly different. These differences were much smaller if the results in science parts of the admission test (MCAT) were taken into account [11]. However, women scored better in part II (clinical sciences) than men.

A-type questions were the first and are the “one correct item out of five items” format. While this simple form is commonly termed a multiple-choice question [12], we will consider this as single choice (SC) or single response (SR), due to only one single alternative being correct which needs to be marked. It is important to note that examinees are well aware of the fact that all alternatives but one are false and presumably make use of this information.

The first aim of our study was to examine the usefulness of a new type of question with the following properties: (1) more than one (k ; $0 \leq k \leq 5$) alternative can be correct requiring the examinee to give multiple responses (MR) and (2) the number of correct alternatives for a given question is unknown to the examinee. We used multiple correct alternatives per question because the SC format as the predominantly used question type in medical education and examination involves some methodological problems. The most serious of which is, that if examinees choose the correct alternative, it is inferred, they would have also known, that all other four are incorrect; a conclusion, which cannot hold critical contemplation.

The second aim of our study was to test different implementations of an answering key for this MR format (Figure 1), namely, the already known key, where true alternatives are checked and false alternatives are left blank (only that more than one checkmark might be needed; MC), a multiple true-false (MTF) format, where the examinee indicates for every alternative whether it is correct or incorrect, and two different rating keys, where every alternative is rated by the examinee on a scale with confidence ratings with either four (R4) or five (R5) categories ranging from “definitely false” to “definitely true.” A progress report of our studies has been published in an abstract form [13]. The overall aim of the study was to test new test instruments that facilitate feedback for the examiners on deeper knowledge and concept on pharmacology formed in their students.

2. Materials and Methods

2.1. General. We conducted three studies in three consecutive years (2012–14) for overall design, Table 1. In each year,

TABLE 1: Overview of study design, year, mode of questions, and participation.

Year	General questions	R5	MC	MTF	R4	Bonus
2012	60	$n = 188$	–	–	–	–
2013	60	–	$n = 58$	$n = 56$	$n = 48$	–
2014	60	–	–	–	$n = 194$	+

Year: year where examination was given; bonus: plus and minus indicate whether or not bonus points were given to increase student participation/motivation; n : indicates how many questionnaires could finally be analyzed in each test instrument; test instrument: general questions were obligatory for all students in each year; 30 questions were given in the midterm examinations and 30 questions were given in the final examinations; the format is single choice (one correct answer to be scored out of five possible answers); R5: five possible answers in multiple-choice questions; MC: single best answer in multiple-choice test; MTF: test of false or true; R4: one of four answers of confidence was to be given.

the general concept of the study followed the same pattern as follows: in the pharmacology course, which comprises lectures and seminars, we administered a midterm exam and a final exam. The final exam was administered at the final week of the course. Immediately following this final exam, the students were given additional questions in an MR format, which we changed. We then calculated the correlation between students’ scores in the regular exams and different scoring methods for the additional questions.

Students were always seated individually, and educators were in the lecture hall where the test was given for the whole time. For the regular exams, identical versions were prepared in four groups (different random sequences of questions). Test versions were given out at random in the class, and examinees received only one version throughout the whole exam. The 30 questions in the midterm exam contained only items that should be known from seminars and lectures till that time. The final exam contained 30 questions pertaining to the whole course. Questions in the midterm and final exam were designed following commonly available guidelines for A-type questions [3], and we allotted 90 seconds for answering each question.

For grading, only the results in the regular midterm and final exams were considered. As each exam consisted of 30 equally weighted questions, we calculated simple sum scores (max. 30 points each). Students passed the course when reaching at least 60% of the combined score of these two exams. This meant 36 correct answers from the 30 questions given in the midterm exam and the 30 questions given in the final exam. Thus, passing the course required taking part in both exams.

The additional questions were given out randomly together with the regular exam sheets before the start of the exam. These questions had been used in their A-type format on prior cohorts and were selected as having shown a difficulty from 0.34 to 0.96 and test discrimination (from 0.21 to 0.58), following the classical test theory (e.g., Kubinger and Gottschall [14]). For the MR format, the question stems were adjusted as slightly as possible to allow for multiple correct alternatives. Answering these questions was always voluntary.

2.1.1. Ethics Statement. This study can be understood as a kind of voluntary survey. No health-related data were collected. Therefore, an ethics statement was not necessary for

<p>Which of the following antimycotic drugs would be best to treat systemic fungal infections?</p> <p>☐ Amphotericin B</p> <p>☐ Flucytosine</p> <p>☐ Ketoconazole</p> <p>☐ Fluconazole</p> <p>☐ Griseofulvin</p>	<p>Which of the following antimycotic drugs can be used to treat systemic fungal infections?</p> <p>Amphotericin B ☐ ☐ ☐ ☐ ☐</p> <p>Nystatin ☐ ☐ ☐ ☐ ☐</p> <p>Tolnaftate ☐ ☐ ☐ ☐ ☐</p> <p>Amorolfine ☐ ☐ ☐ ☐ ☐</p> <p>Griseofulvin ☐ ☐ ☐ ☐ ☐</p> <p style="text-align: center;"> ☐ Definitely true ☐ Probably true ☐ I am not certain ☐ Probably wrong ☐ Definitely wrong </p>
(a)	(b)

FIGURE 2: Examples of SC and R5 question types. Sample questions with their five alternatives and the corresponding key used in 2012. (a) Left panel: the question as it was used in prior exams in the SC format requiring to check the one and only true alternative and to leave false alternatives blank. (b) Right panel: the same question with slightly modified stem to allow for multiple true alternatives. Students were asked to rate each alternative on a 5-point scale (R5).

this kind of voluntary survey. Nevertheless, the study complies with any applicable personal data protection regulations. The students were informed about the project several weeks before the examination. Participants gave their consent by filling out the exam. These additional voluntary exams were written directly after the obligatory exams but were not part of routine course procedure, and participation had no influence on passing the obligatory exams.

2.1.2. Statistical Methods. Significances of differences between correlations were tested using Fisher's r -to- z transformation (Cohen and Cohen [15], p. 54).

2.2. Year 2012

2.2.1. Aim. As a first step, we wanted to assess how the MR format with a 5-point rating key (R5) performs versus the traditional SC format. Since all the questions were taken from the same pool used for regular exams, we expected that the performance in both formats should be about the same when each alternative is scored in the MR format. However, we expected substantially lower scores in the R5 format when the questions are scored as a whole because only one incorrectly answered alternative would render the whole question incorrect and cost all other correct answers for this question.

2.2.2. Test Instrument. To test the usefulness of our new R5 format, we asked two groups of students to either mark ten questions in the traditional SC format or mark their ten modified counterparts in the R5 rating format following the regular final exam. An example of a typical additional question is shown in Figure 2. As described above, we used previously used A-type questions (single choice) with slightly adjusted question stems for the additional questions. Further, one question was negated, so that all four alternatives were correct. All other nine questions had one correct and four incorrect alternatives.

2.2.3. Test Administration. Half of the students received the additional items in their original SC format and half of the students in the new adjusted MR rating format with a 5-point rating scale (R5). Rating scales are a classical instrument ("Likert scales") in education research [16]. Students received no extra credit when answering the additional questions. Students did not know in advance what types of additional questions they were to expect. However, both groups were informed about the required way of giving a correct answer in an enclosed text. Few (less than 5%) students stayed in the examination room till the end of the allotted time.

2.2.4. Sample. Examinees were, $N = 245$, first-year clinical students in medicine who took part in both the obligatory midterm and final exam in pharmacology. Of this number of students, only the data of $N = 188$ were analyzed as follows. The rest of the students were excluded from the analysis either, because they did not return their sheet with the additional questions (1), they did not provide their student identity number on the sheets with the additional questions and thus could not be matched with their regular exam performance (14) or, in the R5 group, they left blank more than five alternatives (42). These 57 students are further summarized as the dropout group (DO). This resulted in rather unequal group sizes of 129 for the SC group and 59 for the R5 group albeit the random distribution of the same amount of exam sheets of both formats.

2.2.5. Statistics. For the additional questions, different scoring methods were applied in the two groups. In the SC group, a similar sum score was calculated for the additional ten questions as in the regular exam. The correct choice of the one out of five answers received one point (max. ten points), and reaching at least 60% of the score was considered as hypothetical for passing these questions.

In the R5 group, each of the 50 alternatives was scored individually as follows: alternatives were scored as correct

and received one point, if (a) a correct alternative was answered with “definitely true” or “probably true” and (b) an incorrect alternative was answered with “definitely false” or “probably false.” In all other cases, the alternative was scored as incorrect. Sum scores were then calculated for either all alternatives (max. 50 points) or for all questions (points for a question were allotted only if all alternatives for these questions were answered correctly; max. ten points as in the SC group).

Due to the increased guess probability of $p = 2/5$ when scoring individual alternatives, different percentages were applied for determining whether examinees have hypothetically passed these additional questions. When scoring alternatives, at least 70% of the score was required, and when scoring questions as a whole, 16.8% of the score was required to hypothetically pass these additional questions.

The ten additional items were not graded and used only for our research purposes. The students did not receive bonus points for good performance in this test.

2.3. Year 2013

2.3.1. Aim. As a second step, we wanted to assess how the three different keys for the MR format (Figure 1) perform versus each other. We expected that each key yields about the same performance, but the keys are being judged differently in difficulty and acceptance by the examinees.

2.3.2. Test Instrument. For the additional questions, we varied the key for the same questions in three groups: multiple choice (MC), multiple true-false (MTF), and a rating format (R4). An example of a typical additional question is shown in Figure 3. As described above, we used previously used A-type questions. While transforming them to the MR format, the question stems were as slightly as possible adjusted to allow for multiple correct alternatives. Further, some alternatives were modified for each question, so that always two questions possessed one, two, three, four, or five correct alternatives. Moreover, only positive correct alternatives were given in the test instruments.

2.3.3. Test Administration. One-third of the students received the additional items in the MC implementation, one-third in the MTF implementation, and the last third in the R4 implementation. Marking the additional questions was voluntary, and students received no extra credit when answering the additional questions. Students did not know in advance what types of additional question they were to expect. However, all groups were informed about the required way of giving a correct answer in an enclosed text.

Together with the exam questions, we distributed a questionnaire about the experiences the students had while answering the additional questions. Students were asked to answer these questions after completing the additional questions. Few (less than 5%) students stayed in the examination room till the end of the allotted time.

Which of the following drugs can be used for the treatment of myasthenia gravis?

(a)	(b)	(c)
☐ Diazepam	☐ ☐	☐ ☐ ☐ ☐
☐ Atropine	☐ ☐	☐ ☐ ☐ ☐
☐ Neostigmine	☐ ☐	☐ ☐ ☐ ☐
☐ Scopolamine	☐ ☐	☐ ☐ ☐ ☐
☐ Pyridostigmine	☐ ☐	☐ ☐ ☐ ☐
	True Wrong	True Probably true Probably wrong Wrong

FIGURE 3: Examples of MC, MTF, and R4 question types. A sample question with its five alternatives and the three implementations of the key used in 2013 (all) and 2014 (only panel). (c): Students were asked to check true alternatives and leave false alternatives blank (MC, panel (a)), to indicate for each alternative whether it is true or false (MTF, panel (b)), or to rate each alternative on a 4-point scale (R4, panel (c)).

2.3.4. Evaluation Instrument. Student’s experiences and opinions about the new questions were collected immediately after the test using a questionnaire. The questionnaire contained ten questions about how easy to answer the additional questions had been, whether they are suitable for exams or online preparation tests, if students wished for a more frequent use of them, and if one of the other not self-experienced keys might be better.

2.3.5. Sample. Examinees were first-year clinical students in medicine of which $N = 218$ took part in the obligatory midterm exam and $N = 211$ took part in the final exam in pharmacology. Students who did not took part in both exams (no midterm: ten of the remainder; no final: two) were excluded, resulting in $N = 209$ examinees with data in both exams.

Of this number of students, only the data of $N = 162$ were analyzed as follows. The rest of the students were excluded from the analysis either, because they did not return their sheets with the additional questions (5), they did not provide or provided an incorrect student identity number on the sheets with the additional questions and thus could not be matched with their regular exam performance (2) or they left blank more than one whole question in the MC group (5) or they left blank more than five alternatives in the MTF group (14) or in the R4 group (21). These 47 students are further summarized as the dropout group (DO). This resulted in more or less equal group sizes of 58 for the MC group, 56 for the MTF group, and 48 for the R4 group.

2.3.6. Statistics. In 2013, the 50 alternatives of the ten additional items were again scored individually. In the MC group, alternatives were scored as correct and received one point if (a) a correct alternative was checked and (b) an incorrect alternative was left blank. In the MTF group, alternatives were scored as correct and received one point if (a) a correct alternative was answered with “true” and (b) an

TABLE 2: Overview of scoring an alternative in the 2014 additional questions. The same table (in German) was made available to students three and a half months prior to the exam.

For one alternative, you receive the points shown here, if...		
	the alternative is correct and...	the alternative is not correct and...
you mark <i>definitely true</i> :	3 points	0 points
you mark <i>probably true</i> :	2 points	1 point
you mark <i>probably false</i> :	1 point	2 points
you mark <i>definitely false</i> :	0 points	3 points

incorrect alternative was answered with “false.” In the R4 group, alternatives were scored as correct and received one point if (a) a correct alternative was answered with “definitely true” or “probably true” and (b) an incorrect alternative was answered with “definitely false” or “probably false.” In all other cases, the alternative was scored as incorrect. Sum scores were then calculated over alternatives (max. 50 points). Due to the increased guess probability when scoring alternatives of $p = 0.5$, at least 75% of the score was required to hypothetically pass the additional questions.

The ten additional items were not graded and used only for our research purposes. The students did not receive bonus points for good performance in this test.

2.4. Year 2014

2.4.1. Aim. As a last step, we wanted to assess if motivation is the key factor to better performance in the additional questions in the MC format. To achieve this, this year, examinees received additional bonus credit for answering the questions in the MC format. We expected that this measure is capable of reducing dropout and/or eliciting better performance.

2.4.2. Test Instrument. This year, only the R4 key was used for the additional questions which were exactly the same as in 2013, but students were informed about three and a half months before the exam about which type of question they had to expect, which scoring method was applied and that they could earn bonus points for good performance. This information was given through a separate instruction sheet enclosed in the announcement of the exam.

2.4.3. Test Administration. All students received the additional items in their MR format with the R4 key. Marking the additional questions was voluntary; however, for the first time, this year, students knew well in advance what types of additional question they were to expect and received up to two bonus points for the regular exam when performing well in the additional questions.

We told students that in order to calculate the bonus points, the following scoring method would be applied: if (a) a correct alternative was answered with “definitely true,” three points were awarded, was answered with “probably true,” two points were awarded, and was answered with “probably false,” one point was awarded, and (b) an incorrect alternative was answered with “definitely false,” three points were awarded, was answered with “probably false,” two points were awarded,

and was answered with “probably true,” one point was awarded (Table 2). In all other cases, the alternative was scored as incorrect and received no points. Sum scores were then calculated for all alternatives (max. 150 points). One bonus point was awarded for reaching at least 60% of the points, and two bonus points were awarded for reaching at least 80% of the score. This information had already been given in the announcement of the exam and was repeated in an enclosed text. Few (less than 5%) students stayed in the examination room till the end of the allotted time.

2.4.4. Sample. Examinees were first-year clinical students in medicine of which $N = 203$ took part in the obligatory midterm exam and $N = 201$ took part in the final exam in pharmacology. The two students who did not take part in both exams were excluded, resulting in $N = 201$ examinees with data in both exams.

Of this number of students, only the data of $N = 194$ were analyzed as follows. The rest of the students were excluded from the analysis only, because they left blank more than five alternatives. These seven students are further summarized as the dropout group (DO).

2.4.5. Statistics. In 2014, only the R4 format was used for all the examinees and again all 50 alternatives were scored individually as before (max. 50 points). Further, we calculated for each alternative the score from zero to three points as described above for awarding bonus points. We then also calculated sum scores over all alternatives (max. 150 points). This scoring method may be considered as a new key which we will report for the sake of completeness and term BP for “bonus points scoring” from here on. Again, at least 75% of the score was required to hypothetically pass the additional questions.

3. Results

3.1. Year 2012. There were no differences between the three groups (SC, R5, and DO) regarding age, $F(2,241) = 0.311$, $p = 0.733$; distribution of gender, $\chi^2 = 1.82$, $p = 0.401$; or total score in the exams, $F(2,242) = 0.213$, $p = 0.808$.

While 95.9% of all students passed the two combined regular exams, in the SC group 78.3% passed the additional questions and in the R5 group only 50.8% passed the additional questions when scoring alternatives or 74.6% passed when scoring the questions as a whole (Table 3 for further statistics on performance). We calculated the correlation between the score in the combined regular exams and the

TABLE 3: Summary of performance in the 2012 regular exams and additional questions. The values given (except N and the correlations) are the respective percentage of the maximum possible value.

		N	Min	Max	Mean	SD	Pass rate	r with combined exam	r with final exam
Midterm exam		245	50.0	93.3	78.9	8.5	97.6	0.774**	0.402**
Final exam		245	16.7	96.7	70.5	11.9	86.9	0.891**	1
Combined exam		245	35.0	91.7	74.7	8.6	95.9	1	0.891**
	SC group	129	0	100	68.3	21.2	78.3	0.359**	0.308**
Additional questions	R5 group (alternatives)	59	0	98.0	64.0	23.0	50.8	0.365*	0.417**
	R5 group (questions)	59	0	90.0	34.1	23.5	74.6	0.253*	0.335*

*Significance with $p < 0.05$ with respective N in each row; **significance with $p < 0.001$ with respective N in each row.

different scorings in the additional questions. The R5 group with scoring alternatives yielded the highest correlation with $r = 0.365$ and $p = 0.004$, closely followed by the SC group with $r = 0.359$ and $p < 0.001$. However, both correlations did not differ significantly, $p = 0.943$, using Fisher's r -to- z transformation [17] and comparing them using formula 2.8.5 from [15]. The correlation in the R5 group with scoring questions was as expected the lowest with $r = 0.253$ but still significant, $p = 0.027$. Scatterplots with a fitted regression line of the performance in the combined regular exams versus the performance in the additional questions are shown in Figure 4.

3.2. Year 2013. There were no differences between the four groups (MC, MTF, R4, and DO) regarding age, $F(3.205) = 0.155$, $p = 0.926$; distribution of gender, $\chi^2 = 1.54$, $p = 0.673$; or total score in the exams, $F(3.205) = 1.61$, $p = 0.188$.

For all further calculations, it is important to note that the overall pass rate in the combined exam was only 69.4%, which is considerably lower than in the 2012 combined exam, where the pass rate was 95.9%. While the pass rate in the midterm exam was already lower than before (78.0% this year versus 93.3% in 2012), the pass rate in the final exam, after which the additional questions were answered, dropped even more (48.8% this year versus 86.9% in 2012). This might have biased all further results. While still 69.4% of all students passed the two combined regular exams, the pass rate for the additional questions was very low in all groups: 5.2% in the MC group, 10.7% in the MTF group, and 10.4% in the R4 group would have passed the additional questions (Table 4 for further statistics on performance). We calculated the correlation between the score in the combined regular exams and the different scores in the additional questions. The MC group yielded the highest and only significant correlation with $r = 0.343$ and $p = 0.004$. The correlation in both of the other groups was about zero and not significant. Scatterplots with a fitted regression line of the performance in the combined regular exams versus the performance in the additional questions are shown in Figure 5. Further, we asked the examinees about their experiences with the additional questions. There were two questions directly related to the questions just answered: whether the examinees could deal well with the format and key they had experienced and whether they could easily answer the questions. For both questions, the MC group agreed most followed by the MTF group and the R4 group (Figure 6(a)). Both effects were significant, $F(2.158) = 9.11$, $p < 0.001$ and $F(2.156) = 5.56$, $p = 0.005$, respectively. In another two questions we asked,

whether the examinees deemed the format and key they had experienced were useful for either teaching or exams. Again, the MC group agreed most, again followed by the MTF and R4 group in this order (Figure 6(b)). Both effects were again significant, $F(2.157) = 6.09$, $p = 0.003$ and $F(2.156) = 9.81$, $p < 0.001$, respectively. Astonishingly, when asked to think of the two other keys they had not experienced themselves and what they think how easy these would be to answer, both the MC and MTF groups chose the R4 key over the other possibility while the R4 group chose the MC key over the MTF key (Figure 7). In the MC and R4 groups, this effect is significant, $t(108) = 3.33$, $p = 0.001$ and $t(88) = 2.11$, $p = 0.037$, respectively, but not in the MTF group, $t(104) = 1.96$, $p = 0.053$.

3.3. Year 2014. There were no differences between the R4 and DO groups regarding age, $t(198) = 1.203$, $p = 0.231$, distribution of gender, $\chi^2 = 1.02$, $p = 0.312$, or total score in the exams, $t(199) = 0.663$, $p = 0.508$.

For all further calculations, it is important to note that while the overall pass rate in the combined exams once again increased compared to the year 2013 (84.6% this year versus 78.0% in 2013 and 93.3% in 2012), the pass rate in the final exam was only 44.8%, which is even lower than in 2013 (48.8%) and considerably lower than in 2012 (86.9%). This is likely to have influenced all further results, as the additional questions were answered after the final exam (Table 5 for a summary of all performances). Still 44.8% of all students passed the two combined regular exams, while the pass rate for the additional questions was low with only 24.2% in the normal R4 scoring. However, this still is an increase of about 14% compared to the year before, where students did not know in advance about the type of questions and the scoring methods applied. We again calculated the correlation between the score in the combined regular exams and the score in the additional questions, but once again, these were around zero and not significant. Scatterplots with a fitted regression line of the performance in the combined regular exams versus the performance in the additional questions are shown in Figure 8.

4. Discussion

There is debate in the literature which kind of written question type is best for medical education. Historically, the first written question items were dichotomic, that is, "true/false." This has been criticized because the language of the question item may be ambiguous (unclear, subject to interpretation, or lacking specifics like age of the patient) or

FIGURE 4: Performance of the year 2012 students. Performance of the 2012 SC group (panel (a)) and the R5 group when scoring alternatives (panel (b)) and whole questions (panel (c)) in the combined regular exams versus their performance in the additional questions. Each dot represents the scores of a single examinee with dots in the upper right rectangle being the ones that passed both exams. Dotted lines indicate scores required for passing. The straight line represents the linear fit to the data.

there may be dissent between graders whether an answer is completely wrong or completely false [3]. Therefore, questions that require one best answer may be superior [3]. This format offers the theoretical possibility to score answers in a continuum between the least correct and most correct. Therefore, they are thought to mirror the clinical reality more closely. Some also recommend not to use negative A-type (i.e., one out of five [3] questions especially types like "each of the following is correct except") because options cannot be rank-ordered on a single continuum, and the examinee cannot determine either the "least" or the "best correct answer" [3]. They rather recommend the Pick "N"

format, in which the examinees are instructed to select "N" responses [3]. In addition, it has been argued that Pick-N format questions might be better than type-A questions because in clinical practice there is often more than one acceptable solution to a given problem [18, 19]. We did, on purpose, not make the examinees aware how many responses they were supposed to pick, in order to make guessing a less useful strategy. Others have studied different scoring scales or scoring algorithms for Pick-N-type MCQ [18]. Firstly, dichotomous scoring: one point if all true and no wrong answers were chosen. Secondly, partial credit scoring: one point for 100% true answers, a half point for

TABLE 4: Summary of performance in the 2013 regular exams and additional questions. The values given (except N and the correlations) are the respective percentage of the maximum possible value.

		N	Min	Max	Mean	SD	Pass rate	r with combined exam	r with final exam
Midterm exam		209	33.3	96.7	68.6	12.6	78.0	0.837*	0.404**
Final exam		209	16.7	83.3	57.2	12.6	48.8	0.839*	1
Combined exam		209	28.3	83.3	62.9	10.6	69.4	1	0.839*
Additional questions	MC group	58	44.0	78.0	58.5	8.1	5.2	0.343*	0.316*
	MTF group	56	38.0	88.0	61.6	10.1	10.7	-0.061	-0.054
	R4 group	48	46.0	80.0	63.7	8.0	10.4	-0.028	0.138

*Significance with $p < 0.05$ with respective N in each row; **significance with $p < 0.001$ with respective N in each row.

FIGURE 5: Performance of the year 2013 students. Performance of the 2013 MC group (panel (a)), MTF group (panel (b)), and R4 group (panel (c)) in the combined regular exams versus their performance in the additional questions. Each dot represents the scores of a single examinee with dots in the upper right rectangle being the ones that passed both exams. Dotted lines indicate scores required for passing. The straight line represents the linear fit to the data.

50% or more true answers, no point deduction for wrong answers, and zero points for less than 50% true answers. Thirdly, a fraction $(1/m)$ of one point was given for each

correct answer given. They concluded that the second and third options had higher statistical reliability and thus recommended to award partial knowledge in these type (Pick-N)

FIGURE 6: Year 2013 evaluation. Mean agreement in the three groups to the statements shown at the bottom of the panels.

FIGURE 7: Year 2013 subjective evaluation. Examinees were asked to think about the other two not self-experienced keys and then to indicate how easy it would be for them to answer the same questions with such a key.

of exams (medical students in Munich, internal medicine tested, Germany [18]). Others came to similar conclusions by showing that the third scoring method also attained higher validity, but the first method exhibited higher reliability (medical students in the USA [19]). It might be also important

that, in contrast to others, we did not detect gender differences in the difficulty of our tests, which might argue for the appropriate construction of our tests. Others noted compared eight different scoring algorithms in secondary school students: they reported a penalizing algorithm that subtracted the portion of marked distractors that was most sensitive for differences in performance between students [20]. In progress test on medical students in the Netherlands, the authors compared number-right scoring system (i.e., incorrect answers in MCQ were not penalized by subtraction from the total score) versus formula scoring (penalizing wrong answers and forcing students to guess the answer). They concluded that number-right scoring exhibited better psychometric properties than formula scoring [21]. In other studies (in a progress test in radiology residents), number-right scores showed lower reliabilities than formula scoring [22].

Another study compared different item formats and calculated reliability coefficients for the study items in each format. They found that 425 A-type questions are needed to obtain a reliability of about 0.90 but only 275 Pick-N questions with partial credit scoring for the same reliability [19]. Hence, fewer questions are needed to obtain reliability, thus saving time for faculty. These Pick-N-type questions were, however, also different from our format in that more than five answers were available [3, 18]. The number of options to be selected is explicitly indicated in Pick-N-type, because otherwise scoring problems can occur. Moreover, sometimes skilled clinicians fared worse with these questions than “test-wise” medical students. However, in the real world, the number of options to select is rarely specified [19]. Hence, it is of merit that we decided not to indicate how many options to select.

An interesting study used a different approach to study confidence in medical students. The authors gave A-type MCQ

TABLE 5: Summary of performance in the 2014 regular exams and additional questions. The values given (except N and the correlations) are the respective percentage of the maximum possible value.

		N	Min	Max	Mean	SD	Pass rate	r with combined exam	r with final exam
Midterm exam		201	40.0	96.7	84.2	13.6	93.0	0.848**	0.301**
Final exam		201	23.3	90.0	56.1	11.1	44.8	0.761**	1
Combined exam		201	36.7	91.7	70.2	10.0	84.6	1	0.761**
Additional questions	R4 scoring	194	44.0	88.0	68.4	8.9	24.2	0.031	0.087
	BP scoring	194	47.3	84.0	66.6	7.8	16.5	0.023	0.053

**Significance with $p < 0.001$ with respective N in each row.

FIGURE 8: Performance of the year 2014 students. Performance of the 2014 R4 group with alternative scoring (panel (a)) and bonus point scoring (panel (b)) in the combined regular exams versus their performance in the additional questions. Each dot represents the scores of a single examinee with dots in the upper right rectangle being the ones that passed both exams. Dotted lines indicate scores required for passing. The straight line represents the linear fit to the data.

in a pharmacology course for junior (third year) and senior students (fifth year) and offered a four-point scale for confidence (from I am very sure to I am guessing). They noticed that higher level medical students were less confident in their answer but exhibited a higher correctness. They concluded that towards graduation, medical students gain more knowledge and also more skepticism which may be a valid goal in medical education [23].

Even though all three implementations of our new type of questions require the same set of knowledge, a different set of “test wiseness” seems required to successfully answer the questions. While the groups did not differ in their scores in the graded exam, only the MC implementation yielded a significant correlation with these scores. Further, the students answering the R4 implementation scored highest, although rating this implementation was the most demanding.

Though in many institutions and countries, the central licensing test given to medical students relies on a clear-cut MC single-choice format, and faculties are often free to choose a format of their own liking for examinations in their courses. We suggest that MC questions with a random number of right answers increase the power of the MC to assess the knowledge of students and might also offer the chance to improve their problem-solving skills and in other words make it possible to

reach higher levels in Bloom’s taxonomy, indeed at least in the US biology teaching to undergraduates even higher Bloom levels of knowledge were measurable with MCQ [24]. Others complained that at least in tests being given to first- and second-year medical students, only low competence levels were used and called for improving the MCQ by offering faculty programs on MCQ writing skills [25, 26].

We noted that performance in questions relevant for the exam was much better than that in our additional voluntary tests given for research purposes which were without consequence for final grading. Probably, this is due to student motivation expecting better grading increases the motivation and hence the performance in the test considerably.

When offering graded response categories ranging from “certainly true” to “certainly false” it is apparently better to offer an even number of options. A neutral category (“I don’t know”) is probably not semantically well defined and thus might corrupt the answer selection by students.

Different question formats are confounded by differing guessing probabilities. To make results comparable with respect to overall performance between tests, appropriate correction formulas should be applied. This only works well if the assumptions on guessing probabilities are correct. In

our study 1, we saw both cases: guessing correction worked well in the guessing scoring case but failed with the single scoring case.

Studies 2 and 3 demonstrated the importance of familiarity with the question formats. Obviously, our multiple-response formats and in particular the confidence scales were initially unfamiliar and this fact posed difficulties to students. It seems to be therefore advisable to introduce any new question format before giving the new examination style (in other words: at the beginning of classes). Thus, students can get acquainted to the new format, can devise optimal answering strategies, and might get familiar with the logic behind the new scoring systems given to them.

In summary, true multiple-response questions are more difficult to answer than single-choice questions. Multiple-response questions offer the examiner more flexibility to test the knowledge of the student, and in some topics, this format makes it easier to construct questions with clinical relevance. Rating scales (I am confident and I am not confident) offer the possibility for the examiner to question the curriculum, detect false concept, and detect which topics have not been covered in the course work. Hence, rating scale questions might be a useful tool in formative testing to improve teaching curricula.

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.

Acknowledgments

The authors want to acknowledge the eLearning development group of the Faculty of Medicine "HaMeel-Hallesches Medizinisches eLearning" and the center for multimedia-enhanced teaching and learning (@LLZ) of the Martin Luther University Halle-Wittenberg. The authors also acknowledge the financial support of the Open Access Publication Fund of the Martin Luther University Halle-Wittenberg.

References

- [1] J. Mathews, *Just Whose Idea Was All This Testing?*, The Washington Post, Washington, DC, USA, 2006, <http://www.washingtonpost.com/wp-dyn/content/article/2006/11/13/AR2006111301007.html>.
- [2] C. Yoakum and R. Yerkes, *Army Mental Tests*, H. Holt and Company, New York, NY, USA, 1920.
- [3] S. M. Case and D. B. Swanson, *Constructing Written Test Questions for the Basic and Clinical Sciences*, National Board of Clinical Examiners, Philadelphia, PA, USA, 3rd edition, 2002.
- [4] D. Newbie, "A comparison of multiple choice and free response test in examination of clinical competence," *Medical Education*, vol. 13, pp. 263–268, 1979.
- [5] G. L. Rowley, "Which examinees are most favored by use of multiple choice tests," *Journal Educational Measurement*, vol. 11, no. 1, pp. 15–23, 1974.
- [6] S. M. Case and D. B. Swanson, "Extended-matching items: a practical alternative to free-response questions," *Teaching and Learning in Medicine*, vol. 5, no. 2, pp. 107–115, 1993.
- [7] S. P. Coderre, P. Harasym, H. Mandin, and G. Fick, "The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts," *BMC Medical Education*, vol. 4, no. 1, p. 23, 2004.
- [8] K. F. Stanger-Hall, "Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes," *CBE Life Sciences Education*, vol. 11, no. 3, pp. 294–306, 2012.
- [9] G. Ben-Shakar and Y. Sinai, "Gender differences in multiple-choice tests: the role of differential guessing tendencies," *Journal of Educational Measurement*, vol. 28, no. 1, pp. 23–35, 1991.
- [10] D. W. Zimmerman and R. H. William, "A new look at the influence of guessing and the reliability of multiple choice tests," *Applied Psychological Measurement*, vol. 27, no. 5, pp. 357–371, 2003.
- [11] S. M. Case, D. F. Becker, and D. B. Swanson, "Performances of men and women on NBME Part I and Part II: the more things change," *Academic Medicine*, vol. 68, no. 2, pp. S25–S27, 1993.
- [12] L. J. Cronbach, "Note on the multiple true-false test exercise," *Journal of Educational Psychology*, vol. 30, no. 8, pp. 628–681, 1939.
- [13] A. Melzer, U. Gergs, J. Neumann, and J. Lukas, "Rating scale methods in multiple choice exams: a pilot study in pharmacology," *Naunyn-Schmiedeberg's Archives of Pharmacology*, vol. 386, no. 1, p. S53, 2013.
- [14] K. D. Kubinger and C. H. Gottschall, "Item difficulty of multiple choice tests dependant on different item response formats: an experiment in fundamental research on psychological assessment," *Psychology Science*, vol. 49, no. 4, pp. 361–374, 2007.
- [15] J. Cohen and P. Cohen, *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Erlbaum, Hillsdale, NJ, USA, 1983.
- [16] G. Norman, "Likert scales, levels of measurement and the 'laws' of statistics," *Advances in Health Science Education*, vol. 15, no. 5, pp. 625–632, 2010.
- [17] R. A. Fisher, "Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population," *Biometrika*, vol. 10, no. 4, pp. 507–521, 1915.
- [18] D. Bauer, M. Holzer, V. Kopp, and M. R. Fischer, "Pick-N multiple choice-exams: a comparison of scoring algorithms," *Advances in Health Sciences Education*, vol. 16, no. 2, pp. 211–221, 2011.
- [19] D. R. Ripkey, S. M. Case, and D. B. Swanson, "A 'new' item format for assessing aspects of clinical competence," *Academic Medicine*, vol. 71, no. 10, pp. S34–S36, 1996.
- [20] A. Domnich, D. Panatto, L. Arata et al., "Impact of different scoring algorithms applied to multiple-mark survey items on outcome assessment: an in-field study on health-related knowledge," *Journal of Preventive Medicine and Hygiene*, vol. 56, pp. E162–E171, 2015.
- [21] D. Ceclio-Fernandes, H. Medema, C. Fernando Collares, L. Schuwirth, J. Cohen-Schotanus, and R. A. Tio, "Comparison of formula and number-right scoring in undergraduate medical training: a Rasch model analysis," *BMC Medical Education*, vol. 17, no. 1, p. 192, 2017.
- [22] C. J. Ravesloot, M. F. Van der Schaaf, A. Muijtjens et al., "The don't know option in progress testing," *Advances in Health Sciences Education*, vol. 20, no. 5, pp. 1325–1338, 2015.
- [23] D. Kampmeyer, J. Matthes, and S. Herzig, "Lucky guess or knowledge: a cross-sectional study using the Bland and Altman analysis to compare confidence based testing of

- pharmacological knowledge in 3rd and 5th year medical students,” *Advances in Health Sciences Education*, vol. 20, no. 2, pp. 431–440, 2015.
- [24] A. Crowe, C. Dirks, and M. P. Wenderoth, “Biology in bloom: implementing Bloom’s Taxonomy to enhance student learning in biology,” *CBE Life Sciences Education*, vol. 7, no. 4, pp. 368–381, 2008.
- [25] H. Abdulghani, M. Irshad, S. Haque, T. Ahmad, K. Sattar, and M. S. Khalil, “Effectiveness of longitudinal faculty development programs on MCQs items writing skills: a follow up study,” *PLoS One*, vol. 12, no. 10, Article ID e0185895, 2017.
- [26] A. A. Vanderbilt, M. Feldman, and I. K. Wood, “Assessment in undergraduate medical examination: a review of course exams,” *Medical Education Online*, vol. 18, no. 1, p. 20438, 2013.