Guidelines on Writing Multiple Choice Questions: A Well-Received and Effective Faculty Development Intervention

Piotr Przymuszała¹, Katarzyna Piotrowska¹, Dawid Lipski¹, Ryszard Marciniak¹, and Magdalena Cerbin-Koczorowska¹

Abstract
Multiple choice questions (MCQs) are commonly used for assessing students, but medical teachers may lack training in writing them. MCQs often have imperfections called item-writing flaws (IWFs) that can affect students' results and impede objective evaluation of their knowledge. This study aimed to evaluate a guideline-based faculty development intervention on writing MCQs at Levels 1 and 2 of the Kirkpatrick Model. MCQs written by teachers prior and after the intervention were analyzed with the Shapiro–Wilk test, Student’s t-test, and Wilcoxon signed-rank test as appropriate. In addition, the phenomenological approach was chosen to describe experiences of 10 teachers in semi-structured in-depth interviews. Results showed satisfaction of teachers with the document. They found it helpful in writing MCQs and noticed the requirement for another guidelines. They appreciated it for its briefness and clarity. The statistical analysis of the quality of MCQs before and after the intervention showed that the document contributed to a statistically significant reduction of IWFs. To conclude, our results showed that a guideline document on writing MCQs may serve as a well-received and effective faculty development intervention. The document may be used as a flexible, time-saving, and just-in-time learning method, fitting needs of medical teachers.

Keywords
multiple choice questions, item-writing flaws, faculty development, Kirkpatrick Model, medical education

Introduction
Multiple choice questions (MCQs) are a common method of knowledge verification (Nedeau-Cayo et al., 2013), and there are many reasons for their frequent use in medical education. They allow to verify a vast amount of knowledge in a relatively short period of time (Epstein, 2007), which is important given number of students enrolled to medical schools every year and the amount of theoretical knowledge they have to assimilate during their studies. Moreover, a lot of examinees can be assessed simultaneously and their answers can be graded by a computer system (Epstein, 2007). MCQs are also viewed as time efficient, easy to grade, and as long as they are well written, they can be an objective, trustworthy, and adequate means of assessment with a potential to evaluate also higher levels of thinking (Case & Swanson, 2002; Epstein, 2007; Pho et al., 2015; Tarrant & Ware, 2012).

However, MCQs frequently have imperfections called item-writing flaws (IWFs). Although superficially they may appear harmless, they can profoundly affect the way students understand and answer questions (Jozefowicz et al., 2002; Tariq et al., 2017). As a result, IWFs can affect final test scores of students and lower the reliability of their assessment. For instance, studies suggest that they tend to disadvantage better students and lower their scores (Tarrant & Ware, 2008). Contrastingly, IWFs can improve grades of weaker students, who are not familiar with the subject (Nedeau-Cayo et al., 2013; Tarrant & Ware, 2008). Some test-wise students can actually use them as cues to deduce correct answers (Tarrant & Ware, 2008).

The occurrence of IWFs in questions may arise from inadequate training and knowledge of academic teachers in the subject of MCQ writing (Tarrant & Ware, 2008), lack of their engagement (Malau-Aduli & Zimitat, 2012), and restricted time due to other academic obligations (Nedeau-Cayo et al., 2013). Therefore, it seems valuable to promote
different interventions aiming at improving the quality of MCQs and making the faculty members aware that fair and reliable assessment of students does not only constitute some form of work-related duty. Especially in medical education, its importance should be understood more broadly as an ethical obligation to students and the society (Tarrant & Ware, 2008, 2012). After all, students who pass a poorly designed exam, although they do not possess adequate knowledge of the topic, may constitute a real threat for their future patients.

Meanwhile, medical teachers are specialists in their disciplines as health care providers, but often lack training in academic skills (Prashanth, 2013). Faculty development can help them to fill these gaps and acquire relevant competencies. However, the success of individual faculty development programs is dependent on local conditions and there are no best, one-size-fits-all models (Lancaster et al., 2014). Faculty development programs may take different forms and their informed choice should be based on careful analysis of their benefits and potential costs (Lancaster et al., 2014). The factors that should be taken into the account to choose the most appropriate intervention include the following: time constraints and educational needs of potential participants, available human and financial resources, and institutional support, among others (Lancaster et al., 2014; Prashanth, 2013).

Evaluation of proposed interventions is important, especially in medical education (Dorri et al., 2016). The methodological approach to faculty development evaluation can vary; however, an increased use of qualitative and mixed methods was suggested to capture their complexity (Steinert et al., 2006, 2016). Many evaluation methods have been developed over the years and one of the most popular is the Kirkpatrick Model, even despite its flaws pointed out by some critics (Dorri et al., 2016; Moreau, 2017). It originally consists of four levels: 1—Reaction, 2—Learning, 3—Behavior, and 4—Results. To respond to criticism and requirements of the modern world, a revised version of the model has been developed—the New World Kirkpatrick Model (Moreau, 2017). It maintained existing division into four parts while expanding and upgrading their descriptions (Kirkpatrick & Kirkpatrick, 2016). The Kirkpatrick Level 1 measures reactions of participants—their satisfaction, engagement, involvement to the learning experience, and relevance of new knowledge to their work. The knowledge and skills gained by the participants as a result of the intervention are measured at Kirkpatrick Level 2. At Kirkpatrick Level 3, the on-job application of what had been learned can be assessed. The Kirkpatrick Level 4 evaluates the occurrence of targeted outcomes of the training.

Most research conducted so far has focused on workshops or training sessions on writing good-quality test questions (Abdulghani et al., 2015, 2017; Dellinges & Curtis, 2017; Tenzin et al., 2017). Although they were proven to be effective in terms of raising the quality of MCQs, the literature describes problems with long-time retention of knowledge from them, which tends to decrease when not systematically repeated (Iramaneerat, 2012). Moreover, the participation of medical teachers in workshops may be problematic given their busy schedule and many other clinical, scientific, and academic duties. Meanwhile, creation of certain tools aiming at enhancing the quality of teaching process is a quite common practice in medical education (Kumar et al., 2014).

Taking all of the above into consideration, we decided to create a guideline document with good practices on writing MCQs for faculty members of the Poznan University of Medical Sciences (PUMS). The aim of this study was to evaluate the effectiveness of the guideline document as a faculty development intervention on improving the quality of MCQs at Levels 1 and 2 of the Kirkpatrick Model. Furthermore, an additional objective was to determine previous experiences of the teachers with MCQs that could potentially affect their views on the tool.

Materials and Methods

Researchers’ Characteristics

When the study was conducted, main researchers were medical students of PUMS—currently they are medicine graduates. During their studies, they often encountered poorly constructed MCQs at their exams, for instance, very complicated in structure and difficult to understand or, on the contrary, ones that could be answered instinctively without adequate knowledge of the topic (logical and grammatical inaccuracies). Their observations led to the idea to create a set of tips that could be useful for the academic staff in raising the quality of MCQs and thus the quality of assessment. Due to aforementioned characteristics of the researchers, which potentially could affect the results, the study was supervised by a senior researcher who is an academic teacher of PUMS employed at the Chair and Department of Medical Education. The tool presented in the study is supposed to become a part of a bigger series, so the motivation of the senior researcher was additionally to gain knowledge on the effectiveness and opinions of the teachers on the document to formulate any future guideline documents in even more efficient manner.

The Tool

The guideline document was formulated in compliance with standards provided by the National Board of Medical Examiners (NBME), which specializes in creating high-quality assessment (Case & Swanson, 2002). Unlike in the case of courses or workshops, the document can be used later by teachers whenever needed (just-in-time learning). It is important due to aforementioned problems with long-time retention of knowledge from them (Iramaneerat, 2012). The main concept behind creating the document was to make it relatively short in order not to discourage the teachers from using it. Finally, it amounted to four pages, beginning with a
brief explanation of the importance of maintaining high MCQ quality and thus employing principles of andragogy—
giving a reason for learning and increasing internal motivation. The tool was designed to guide a teacher step by step
through the MCQ writing process from the choice of topic to final improvements. In addition, a checklist was added at the
end for fast and efficient proofreading. Therefore, the tool can be applied for both writing new MCQs and correcting
the ones that already exist.

Prior to disseminating the guideline document to the faculty members, its draft was reviewed by three experts on the
subject, which allowed to form the final version of the document (Supplemental Appendix 1).

**Study Settings**

Invitations to participate in the study were sent by e-mail to all academic teachers of PUMS by the administrative office.
They contained a detailed description of the study protocol and discussed rights and obligations of the participants, including
the right to resign from further participation at any stage of the study. Potential participants were also assured that all data
would be anonymized and later analyzed and processed only in that form. To minimalize the influence of selection mecha-
nisms on the results of the study, the sole criterion of inclusion was the active status of an academic teacher.

**Data Collection**

Academic teachers who agreed to partake in the study were asked to send in a sample of at least 15 MCQs written by
them (Phase 1). Then the intervention took place—the aforementioned guideline document on MCQ writing was sent to
the participants via e-mail and they were asked to get familiarized with it. In Phase 2, teachers were asked to evaluate
and appropriately correct their old questions or write new ones and send them in for re-evaluation. This part of the
study intended to evaluate the quality of MCQs written by the academic teachers before and after reading the guideline
document (Kirkpatrick Level 2).

Subsequently, participants of previous stages of the study were invited for semi-structured in-depth interviews to
express their opinions on the document, including its perceived advantages and disadvantages (Kirkpatrick Level 1).
They were also asked about their previous experiences with MCQs. Each interview was conducted according to the stan-
dardized template and lasted 10 to 20 min. The outline of the template used during the interviews is presented in Table 1.
To ensure high quality of the data obtained, every interview was recorded and emphasis ought to be put on the fact that
every participant had a right to refuse it and resign from further participation.

**Data Analysis—Kirkpatrick Level 1**

Researchers decided to use phenomenology as their qualitative approach to attain detailed comprehension of personal
experiences, perspectives, and motivations of the teachers on the topic of MCQs rather than clichéd assumptions and con-
ceptions (Cleland, 2017). Therefore, the results are not to be viewed as objective and representative for wider population.
Instead, they are very individual and subjective insights on ways the participants think about MCQ creation, usefulness
of the tool, and quality of questions they encountered in the past. Due to their diversified educational backgrounds, fields
of medical science, and years of work as the teacher, their attitudes and mentality are also different and we tried to cap-
ture them all to get an uncontaminated and thorough picture. Efforts were made to hide personal views of the researchers
in order not to suggest answers and let the participants to speak freely in a welcoming environment.

Qualitative data obtained from each interview were separately analyzed and encoded with Atlas.ti Software (ATLAS.
ti 8 EDU). Their subsequent analysis revealed that several opinions and experiences were mentioned by more than one
respondent. To ensure the best objectivity of the results, data were analyzed and encoded by two independent researchers.
During the data analysis, each researcher identified fragments of interviews that seemed crucial and important. They
were later compared with each other and debated until

---

**Table 1. The Outline of the Interview Template.**

| Motivation for participation in the project |
| Questions on the tool |
| 1. What is your opinion on the presented document to evaluate and standardize multiple choice questions? |
| 2. Would you like to suggest any changes to the document? |
| 3. What are the positive and negative aspects of the document? |
| 4. How do you rate the usefulness of the document? |
| 5. How do you rate the attractiveness of the document? |
| 6. What is your suggested way of its popularization? |

| Questions on multiple choice questions |
| 1. In your opinion, is writing multiple choice questions an easy or difficult task? Why? |
| 2. What is your opinion on the quality of multiple choice questions you have encountered in the past? |
| 3. Do you think it is necessary to implement actions to improve the quality of multiple choice questions? If so, what should be done? |
consensus was reached. To achieve high reporting standards in qualitative studies, researchers followed the Standards for Reporting Qualitative Research (O’Brien et al., 2014).

**Data Analysis—Kirkpatrick Level 2**

The quality of MCQs sent in by the teachers was evaluated in search of IWFs, both prior and after the intervention. First, text analysis was performed to identify IWFs in questions obtained from participants. To ensure the high quality of data obtained, the evaluation of questions was performed simultaneously and independently by two researchers. Their results were later compared and analyzed by a third researcher and debated until consensus was reached. Identified IWFs were then encoded and quantified.

Given unequal number of MCQs presented by different teachers, authors defined additional parameter illustrating the mean frequency of IWF made by a particular participant:

\[
IWF\text{ratio} = \frac{\sum IWF}{\sum MCQ},
\]

where \(IWF\text{ratio}\) is the mean frequency of IWF made by a particular teacher, \(\Sigma IWF\) is the total number of identified IWFs in questions written by a particular teacher, and \(\Sigma MCQ\) is the total number of MCQs written by a particular teacher.

Quantitative data received were later statistically analyzed with the use of Shapiro–Wilk test (to check for normality), Student’s \(t\)-test (for normally distributed data), and Wilcoxon signed-rank test (for non-normally distributed data), being a part of Statistica Software (StatSoft).

**Ethical Considerations**

To ensure the highest ethical standards of the research project, the study protocol was presented to the ethical committee at PUMS for opinion and authors paid especial attention to meet the Ethical Guidelines for Educational Developers (British Educational Research Association, 2012). In particular, all data obtained that could be used to identify the participants were encoded and further proceeded anonymously. Moreover, every participant was informed about the aims and full protocol of the study. They were also asked to sign a written consent prior to participation and had a right to resign from further participation at every stage of the study. Every interview was recorded and the recordings were encoded, kept in a password-protected folder and disclosed only to the researchers.

**Results**

**Kirkpatrick Level 1—Reaction**

Respondents provided different reasons for their participation in the study, including the opportunity for self-development, curiosity, and complaints from students in the past—resulting in a decision to confront their knowledge with the guideline document and improve their test questions.

All participants expressed their satisfaction with the document and recognized it as a useful and helpful tool in everyday work of an academic teacher. They pointed to the paucity of similar initiatives and general lack of knowledge on the topic among the faculty members:

P04: There is a need to collect all these tips on writing questions in one place, because no one teaches us how to do it and everybody only expects us to write them.

In the opinion of respondents, the presented guideline document fulfilled their educational needs. They emphasized its usefulness in acknowledging their mistakes and saw a potential for creating similar tools in the future:

P08: The guide you sent me is a very concrete and factual help while designing test questions and in my personal opinion it is on very high level, especially given that at PUMS there is nowhere to find information on the subject.

P04: I can feel that my knowledge is now systematized. Now I know what was good and what should be improved.

P06: You should also make guidelines for practical exams and other things. I had no idea it could be done in such way—it is very necessary.

The document was valued on the basis of its merit (as it provides readers with essential information on MCQ writing), easily understandable language, shortness, and step-by-step instructions:

P01: I think that everything is very transparent, clearly written, and contains concrete guidelines.

P04: Concrete, sensible advices, user-friendly in volume.

Some respondents had already participated in workshops dedicated to MCQ writing, but none of them ever acquainted a tool or a document summing up the most important rules that could be used in everyday practice. In addition, two respondents stated that the estimated time of writing MCQs was shorter with the use of the tool, while the rest thought that it did not change significantly, but the task was easier:

P06: Earlier, [...] I attended different workshops, but this tool helped me to consolidate and systematize my knowledge. [...] Despite the workshops I was still making mistakes and with this tool I can once more take a look at the questions and rewrite them. [...] I am very pleased that something like this was made. [...] I forwarded it to my colleagues to help them.

Another appreciated aspect was the checklist at the end. Respondents found it helpful as a final verification of their questions in a very time-efficient manner:
P05: I really liked the checklist at the end, [...] you remember many of these tips somewhere deep in your head, but you don’t think about them on daily basis. Thanks to the checklist you can repeat them constantly and in a while you won’t have to use it as you will remember all of the points.

P07: First a general explanation and next a checklist—it makes the work easier and later allows to quickly check the questions.

One change proposed by respondents would be adding some examples of high- and low-quality MCQs next to each good practice:

P07: The description is good, but I would add some examples next to each tip. However, I wonder if the document won’t lose its briefness and clear structure.

Respondents were also asked to share their opinions and previous experiences with MCQs and they rated their quality as rather low or mediocre. They noticed the problem of IWFs and that MCQs containing them do not adequately verify the knowledge of examinees:

P06: Lots of questions with negations, tricky ones only to confuse students. They don’t verify knowledge, but rather alertness and this isn’t the way. Instead of useful clinical knowledge they ask about trivia, rare things or percentages.

P07: I find the quality of questions low. They often contain errors and are poorly constructed, e.g., negations in their stem. From a time perspective I see that they don’t bring a lot. [...] They only check the ability of mechanical recollection of knowledge.

They also expressed understanding for the students’ frustration and disappointment connected with it, as they also were unable to answer some questions, because they simply did not understand them:

P03: Now I can understand my students and their doubts after the exam—that something didn’t add up or could be rewritten or the content of questions suggested answers.

P04: There is still a lot of work to be done—the questions are not of best quality and it is a big problem for students. [...] From my personal experience a striking contrast and a source of frustration—after good classes there comes a test that is badly conducted and students feel cheated.

P05: I try to write clear, understandable questions, since [as a student] I often had problems with understanding what the author [of the question] had in mind.

However, some of them also admitted that sometimes they repeated the same flaws in their own questions due to convenience and force of habit, although they knew it was not appropriate:

P04: It’s amazing how it works. Even though I want to write good questions, I have these patterns, imprinted deep inside my mind, of the questions from my past when I was a student and as a result I repeat the same mistakes my teachers made.

According to all respondents, it is very difficult to write a good question and their opinion on the subject has changed significantly as they started their academic carriers—at the beginning they perceived it as an easy task:

P02: As a student I used to think writing questions was easy, but as a teacher I can see it is really difficult not to make mistakes.

P07: Making up adequate, plausible distractors and correct stems of the questions is a challenge [...] as well as choosing their content and establishing good proportions between difficult, medium and easy questions—especially from the perspective of someone who knows the topic very well.

In their opinion, low quality of questions may arise from inadequate knowledge and education of examiners, their mentality, and lack of time and motivation to improve their MCQs. One respondent proposed offering teachers additional financial compensation to raise their motivation to improve their skills. Moreover, respondents felt that the control of MCQs’ quality was not adequate and should be increased. Some suggested they would like to receive feedback on their questions:

P10: There is still a lot to be done on the subject. First, you have to raise the awareness that there are guidelines on writing questions and second, you have to encourage teachers to use them.

P08: Obligatory introduction of such guidelines and subsequent random control of their use. We should also ask students for their opinions after exams.

P04: Motivation can be boosted by being controlled—everyone should receive feedback on their questions.

Respondents suggested that the implementation of the tool and changing the thinking process of the academic staff should be multidirectional. They proposed placing it on the website of the university, sending it to the teachers as a newsletter via official communication channels and using it during workshops dedicated to writing questions. They also expressed a strong opinion that preferably every academic teacher or at least one representative of each department should be familiarized with it:

P06: I think there should be a compulsory training in it. On the other hand, it can cause protests—maybe at least one person from every department.

P08: Every new employee should receive these guidelines.
Kirkpatrick Level 2—Learning

Initially, 20 academic teachers agreed to participate in the study and sent in 353 test questions that underwent data analysis in Phase 1. Among 326 questions matching MCQ criteria, 203 (62.27%) questions contained no IWFs and 123 (37.73%) questions contained at least one IWF. Detailed results are presented in Figure 1. Despite the fact that all teachers were subjected to the intervention provided in the study protocol, only 10 of them managed to send new or corrected questions and took part in Phase 2. Detailed characteristics of the whole study group are presented in Table 2.

Interestingly, the number of identified IWFs was higher among teachers who resigned from participation in Phase 2. Out of 149 MCQs provided by them, only 76 (51.70%) were free of IWFs with the total of 106 IWFs identified. These values for the rest of the study group (179 MCQs) were 127 (70.95%) and 77, respectively. Although the IWF ratio values also differed (0.73 ± 0.47 vs. 0.44 ± 0.38, respectively), the difference between both groups did not turn out to be statistically significant in Student’s t-test ($p > .05$).

Given that the aim of the study was to evaluate whether participants’ skills would change as a result of the intervention, the quality of their MCQs was compared before and after they became acquainted with the guideline document. A total of 10 academic teachers who decided to participate in Phase 2 of the study sent in 148 new or corrected questions after receiving the document. Of the 179 MCQs prior the intervention sent in by them, 127 (70.95%) MCQs contained no IWFs while 52 (29.05%) had at least one IWF. On the contrary, among 148 MCQs sent by the same teachers after receiving the guideline document, 139 (93.92%) MCQs contained no IWFs and only 9 (6.08%) MCQs had one and more IWFs. Detailed comparison of MCQs’ quality prior and after the intervention is presented in Table 3.

The incidence of most frequent IWFs at different stages of the study is presented in Table 4. In our sample, the most common IWFs were as follows: presence of negative stems, problem in option not in stem and unfocused stem, phrases like “all of the above” or “none of the above,” and dissimilar length options, although their prevalence differed among different groups. As it can be observed in Table 4, some IWFs were completely eliminated as a result of the intervention (e.g., use of absolute or vague terms, use of phrases like “all of the above” or “none of the above,” and grammar or logical cues). However, some IWFs seem more persistent and due to reduced overall number of IWFs, after the intervention their percentages even increased (e.g., presence of negative stems or problem in option not in stem and unfocused stem).

Discussion

Participants’ Satisfaction With the Tool

The study delivered a lot of information on the usefulness and role of the document and other similar initiatives yet to be made. Our respondents were pleased that such a tool was created and found it helpful in their everyday work. Therefore, it may be stipulated that there is a requirement for compacted sources of information on related topics to be distributed in a form of handouts among academic teachers in the future. Moreover, we hope that as the tool satisfied educational needs of teachers and they found it useful in their work, they will use it while writing new MCQs. However, further studies on use of similar tools are required.

The respondents valued the document for its briefness, clear layout, and structure, making it quick and approachable to read. They also appreciated the checklist at the end as a template for final evaluation of test questions. Some respondents suggested accompanying each tip with examples of MCQs with given IWF and some good-quality MCQs. Ali (1981) also underlined the importance of putting positive and negative examples next to a set of instructions. On the contrary, teachers noticed that it would diminish one of document’s best assets—its shortness and a sole provision of examples do not necessarily assure increase in the
effectiveness (Ali, 1981). Unfortunately, as research on the topic showed, lack of time among academic teachers constitutes a serious problem (Nedeau-Cayo et al., 2013).

**Participants’ Previous Experiences With MCQs**

According to the teachers, the problem of MCQs’ quality is worth mentioning as convenience of use and other benefits still make them a very popular form of knowledge evaluation (Epstein, 2007; Kumar et al., 2014). The respondents acknowledged the fact that questions with IWFs do not adequately evaluate the knowledge of examinees. Teachers also showed an understanding for the frustration and disappointment of students after receiving flawed questions at their exams. Based on their previous experiences with test questions, respondents graded the quality of MCQs as low or mediocre. They often contained IWFs and some were even difficult to understand. Nevertheless, some teachers admitted to subconsciously repeating those IWFs, as a force of habit, in their own questions. Previous experiences of adult learners may hinder forming new habits (McCray, 2016). Mastering new skills should involve critical reflection and formative feedback to challenge them (McCray, 2016; Taylor & Hamdy, 2013). Sadly, as numerous studies showed, nonadherence to rules on writing MCQs is a common practice (Tarrant & Ware, 2008), even despite the extensive literature on the topic (Burton et al., 1991; Campbell, 2011; Case & Swanson, 2002). Meanwhile, Coughlin and Featherstone (2017) concluded that only well-written MCQs can be a motivation for students to learn and serve as a tool for evaluation of their knowledge. As Tarrant and Ware (2008) suggested, poorly written MCQs tend to harm more talented students with bigger knowledge.

Participants emphasized the usefulness of the document while preparing new questions and acknowledging flaws in

---

**Table 2.** Characteristics of the Study Group.

| Type of participation in the study | Gender | Type of classes given | Academic degree or title | Profession | Code |
|-----------------------------------|--------|-----------------------|--------------------------|------------|------|
| Teachers who participated in the whole study | Female | Basic | MSc | Pharmacist | P04 |
| | Female | Basic | MSc | Pharmacist | P07 |
| | Female | Basic | MSc | Physiotherapist | P05 |
| | Female | Clinical | MSc | Nurse, biologist | P02 |
| | Female | Clinical | MSc | Pharmacist | P10 |
| | Female | Clinical | MD | Physician | P06 |
| | Female | Clinical | PhD | Nurse | P03 |
| | Female | Clinical | PhD | Physician | P01 |
| | Male | Basic | MSc | Pharmacist | P08 |
| | Male | Clinical | PhD | Pharmacist, paramedic | P09 |
| Teachers who participated only in Phase 1 | Female | Basic | MSc | Nurse, paramedic |
| | Female | Basic | PhD | Pharmacist |
| | Female | Basic | Professor | Physician |
| | Female | Clinical | PhD | Nurse |
| | Female | Clinical | PhD with “habilitation” | Physician |
| | Male | Basic | PhD | Paramedic |
| | Male | Basic | PhD with “habilitation” | Physician |
| | Male | Clinical | PhD | Physician |
| | Male | Clinical | Professor | Physician |

*PhD with “habilitation”—a post-doctoral degree.

**Table 3.** Quality of MCQs Prior and After the Intervention.

| MCQs overview | Prior intervention | After intervention |
|----------------|--------------------|--------------------|
| **ΣMCQs**      | 179                | 148                |
| **IWFRatio**   | 0.44 ± 0.38        | 0.09 ± 0.15        |
| **Upper quartile** | 0.53               | 0.10               |
| **Median**     | 0.40               | 0.00               |
| **Lower quartile** | 0.21               | 0.00               |

Note. MCQ = multiple choice question; IWF = item-writing flaw; IWFRatio = mean frequency of IWF made by particular teacher.

Statistical analysis was performed using the Wilcoxon signed-rank test.

Participants emphasized the usefulness of the document while preparing new questions and acknowledging flaws in
the existing ones. They expressed the notion that writing good MCQs and avoiding mistakes was more difficult as they had initially thought. Contrastingly, they believed that too extensive knowledge of the subject could negatively affect the quality of MCQs. Their opinion is mirrored by Kheyami et al. (2018) who also viewed writing MCQs as a difficult task. Interestingly, Abdulghani et al. (2015) showed no correlation between the experience in the work of an academic teacher and the number of flaws in MCQs.

**Propositions on Popularization of the Tool**

Inadequate knowledge and education of the examiners was often stated as a reason for the low quality of MCQs. Jozeffowicz et al. (2002) concluded that the quality of questions constructed by authors who had been trained on the topic by the NBME was considerably higher than of those who had not participated in such training sessions. Interestingly, although many studies proved that organizing workshops on MCQ writing positively influences their quality (Abdulghani et al., 2015; Jozeffowicz et al., 2002; Naeem et al., 2012), some authors pointed to the problem of inadequate long-term retention of knowledge from them, which seems to decay without regular repeating (Iramaneerat, 2012). Some of the respondents of this study also participated in workshop sessions dedicated to writing MCQs. They admitted that their recollection of knowledge on the topic was insufficient and were open to an idea of collecting most crucial tips in the form of a short hand-out. Another problem related to the quality of MCQs often recalled by the participants was the mentality of the academic staff. In their opinion, the implementation of the tool should be multidirectional and the most common means of its popularization were as follows: the website of the university, official communication channels, and using it during workshops dedicated to writing questions. Some teachers expressed the opinion that familiarizing with it should be compulsory for all new employees and at least one representative from each department. According to the theory of planned behavior by Ajzen (1985), the way in which an environment perceives given behavior is crucial. As a result, putting high emphasis on raising MCQs’ quality and promoting the tool by the university may additionally motivate teachers to self-improvement in this area as a kind of positive social pressure.

**General Overview of MCQs’ Quality Prior the Intervention**

Before the intervention, over 62% of all MCQs did not contain any IWFs while approximately 23% and 15% had one and more than one IWF, respectively. For comparison, in similar studies from other countries, the number of test questions with one IWF varies between approximately 30% and 50% and those with more than one IWF around 10% to 35%.

---

**Table 4. The Most Common IWFs at Different Stages of the Study.**

| Type of IWF                              | Phase 1                  | Phase 2                  |
|------------------------------------------|--------------------------|--------------------------|
|                                         | Prior intervention       | After intervention       |
| Contains absolute terms (e.g., always, never) | $N = 183$ IWFs           | $N = 77$ IWFs            | $N = 13$ IWFs            |
| Contains grammatical cues                | $n = 3$; 1.64% [8]       | $n = 1$; 1.30% [9]       | $n = 0$; 0.00%          |
| Contains K-type items                     | $n = 10$; 5.46% [5]      | $n = 3$; 3.90% [7]       | $n = 0$; 0.00%          |
| Contains logical cues                    | $n = 2$; 1.09% [9]       | $n = 1$; 1.30% [9]       | $n = 0$; 0.00%          |
| Contains long correct answer             | $n = 9$; 4.92% [6]       | $n = 6$; 7.79% [5]       | $n = 1$; 7.69% [4]      |
| Contains negative stem                   | $n = 47$; 25.68% [1]     | $n = 15$; 19.48% [2]     | $n = 6$; 46.15% [1]     |
| Contains opportunities to use convergence strategy | $n = 3$; 1.64% [8]       | $n = 0$; 0.00%           | $n = 0$; 0.00%          |
| Contains options with dissimilar lengths | $n = 23$; 12.57% [3]     | $n = 17$; 22.08% [1]     | $n = 4$; 30.77% [2]     |
| Contains the phrase                       | $n = 23$; 12.57% [3]     | $n = 10$; 12.99% [3]     | $n = 0$; 0.00%          |
| “all of the above”                       | $n = 16$; 8.74% [4]      | $n = 8$; 10.39% [4]      | $n = 0$; 0.00%          |
| “none of the above”                      | $n = 3$; 1.64% [8]       | $n = 2$; 2.60% [8]       | $n = 0$; 0.00%          |
| Contains vague frequency terms (e.g., rarely, usually) | $n = 2$; 1.09% [9]       | $n = 1$; 1.30% [9]       | $n = 0$; 0.00%          |
| Fill-in-blanks                           | $n = 4$; 2.19% [7]       | $n = 1$; 1.30% [9]       | $n = 0$; 0.00%          |
| type of questions                        | $n = 25$; 13.66% [2]     | $n = 8$; 10.39% [4]      | $n = 2$; 15.38% [3]     |
| Language in options is nonparallel       | $n = 2$; 1.09% [9]       | $n = 0$; 0.00%           | $n = 0$; 0.00%          |
| Numeric data are not stated consistently | $n = 4$; 1.30% [9]       | $n = 0$; 0.00%           | $n = 0$; 0.00%          |
| Problem in option not in stem and unfocused stem | $n = 2$; 1.09% [9]       | $n = 0$; 0.00%           | $n = 0$; 0.00%          |

**Note.** IWF = item-writing flaw.

*Numbers in square brackets indicate position of given IWF in the frequency ranking of their occurrence in a given group. K-type items: combinations of more than one options (e.g., A and C are correct).
For example, in a study by Nedeau-Cayo et al. (2013), only 15.5% questions had no IWFs, 49.9% questions had one IWF, and 34.6% more than one IWF. Tarrant et al. (2006) found out that 33.9% of examined MCQs had one IWF and over 12% of MCQs contained more than one IWF. Similarly, almost half of 185 MCQs evaluated by Kowash et al. (2019) had one or more IWF. Although our results seem to suggest better quality of MCQs at our institution, the authors remain cautious with their interpretation given the low number of academic teachers who participated in the study. Furthermore, it can be stipulated that teachers who kindly agreed to participate in our study were more positively inclined toward the topic and therefore could already have some experience in it. Finally, this was only supposed to be a pilot study aiming to verify whether the proposed intervention could in fact improve the quality of MCQs and not to assess the quality of MCQs at our institution in general.

Analysis of the incidence of individual IWFs in examined test questions allowed to create a list of most common ones throughout different parts of the study (Table 4). The most common types of IWFs identified in Phase 1 were similar for teachers who participated in the whole study and teachers who refrained from further participation. These were as follows: “negative stem,” “all of the above,” “dissimilar length options,” “problem in option not in stem and unfocused stem,” and “none of the above.” Interestingly, similar IWFs are common also in other studies. Detailed comparison of the most common IWFs described in the literature and those identified in this study (Table 5) suggests that academic teachers from around the world seem to have similar difficulties with writing MCQs.

### Table 5. Comparison of the Most Common IWFs in This and Other Studies.

| Type of IWF                                                                 | This study | Nedeau-Cayo et al. | Baig et al. | Tarrant et al. | Kowash et al. |
|----------------------------------------------------------------------------|------------|--------------------|-------------|----------------|---------------|
| Contains negative stem                                                    | 1          | 9                  | 6           | 4              | 1             |
| Problem in option not in stem and unfocused stem                          | 2          | 12                 | 4           | 2              | 1             |
| Contains options with dissimilar lengths                                  | 3          | 5                  |             |                |               |
| Contains the phrase “all of the above”                                    | 3          | 1                  |             |                | 9             |
| Contains the phrase “none of the above”                                   | 4          | 6                  |             | 1              | 11            |
| Contains K-type items                                                     | 5          |                    |             |                |               |
| Contains grammatical cues                                                 | 6          |                    |             |                | 15            |
| Contains long correct answer (more details)                               | 6          |                    | 6           | 6              |               |
| Numeric data are not stated consistently                                   | 7          |                    |             |                |               |
| Contains absolute terms (e.g., always, never)                             | 8          | 11                 |             |                | 13            |
| Contains opportunities to use convergence strategy                         | 8          |                    |             |                | 3             |
| Contains vague terms (e.g., rarely, usually)                              | 8          |                    |             |                | 10            |
| Contains logical cues                                                     | 9          |                    |             | 5              | 7             |
| Fill-in-blanks type of questions                                           | 9          |                    |             |                | 12            |
| Language in options is nonparallel                                         | 9          |                    |             |                |               |
| Stems are tricky or overcomplicated                                       | 9          |                    |             |                |               |
| Others                                                                     |            |                    |             |                |               |
| Implausible distractors                                                   | 3          |                    |             |                | 5             |
| Lost sequences in data                                                    | 7          |                    |             |                | 14            |
| More than one or no correct answers                                       | 2          |                    |             |                | 4             |
| Repeating words                                                           | 4          |                    |             | 8              | 8             |
| Unclear information                                                       | 14         |                    |             |                |               |
| Unnecessary information in stem                                            | 8          |                    |             | 3              | 3             |

Note. Numbers in square brackets indicate the position of given IWF in the frequency ranking of their occurrence in a given group. IWF = item-writing flaw.

Evaluation of the Effectiveness of the Intervention

Comparison of mean frequencies of IWFs ($IWFratio$) made by teachers who participated in the whole study demonstrates that presented intervention may significantly contribute to reducing the number of IWFs in test questions and thus raising their quality. Prior to the intervention, at least one IWF was present in 29.05% and more than one IWF in 10.06% of examined MCQs (Table 3). When the academic teachers were given the opportunity to get acquainted with the guideline document, the number of IWFs in their new MCQs reduced to 6.08% and 2.03%, respectively. Similarly, median of values of $IWFratio$ was reduced from 0.40 before the intervention to 0.00 after the tool was presented to the teachers ($p = .008$).

Improvement was also noted with regard to incidence of particular IWFs. Out of five most common IWFs before the study, two (phrases “all of the above” and “none of the above”) were completely eradicated as a result of the intervention and the occurrence of remaining three (“negative stem,” “dissimilar length options,” and “problem in option not in stem and unfocused stem”) was significantly reduced.
unfocused stem”) was decreased. Moreover, it seems that less common IWFs were easier to eradicate as MCQs after the intervention contained only one of them—“long correct answer.” It can be stipulated that some of the popular IWFs are rooted more deeply in teachers’ habits. As a result, more attention has to be paid to them and teachers may need more time and training to adjust. Interestingly, it was also observed by our respondents during the interviews.

It should be emphasized that authors of the present study do not expect a single tool to be enough to achieve high-quality test questions. Writing good test questions is a difficult task that requires not only theoretical knowledge but also practice (Tarrant et al., 2006). We do, however, believe that it may constitute an important part of the process, considering promising results of the presented pilot study.

What Are the Alternatives?

Interventions aiming to improve the quality of teaching or assessment process are quite a common practice in medical education. However, to the best of authors’ knowledge, this is the first study presenting that a short guideline document can serve as one of them, contributing to significant reduction of IWFs and thus improving the quality of MCQs. As it was mentioned above, most of the studies conducted so far have focused on workshops or training sessions dedicated to writing test questions (Abdulghani et al., 2015, 2017; Dellinges & Curtis, 2017; Tenzin et al., 2017). Although they are effective in terms of raising quality of questions, it can be only true if teachers take their time to attend them. Meanwhile, medical teachers are often also active health care workers, making their time limited due to numerous obligations. Moreover, organization of workshops requires financial resources for salaries of personnel leading them, classroom rental, equipment and educational materials, among others. In a perspective of couple of workshops, these costs do not seem high, but it should be remembered that they have limited number of places available for participants. In case of larger institutions, the number of academic teachers that need to be trained grows significantly and so do the costs of their organization. Finally, given the aforementioned problems with the long-time retention of non-repeated knowledge (Iramaneerat, 2012), it seems necessary to organize additional refresher courses. On the contrary, the presented intervention offers teachers more flexibility in their learning process. As it takes a relatively short period of time to read the document, teachers can do it whenever they find fit. Moreover, they can read it or repeat it every time they start to write new test questions (just-in-time learning), thus resolving the problem of knowledge retention associated with workshops. Finally, the presented document can be simultaneously presented to all academic teachers at given institution and this process does not generate any additional costs.

No faculty development initiative can be successful if academic teachers are not willing to participate in it. Unfortunately, we noted low interest of teachers in participating in the study. Although invitations were sent to all academic teachers of our institution, only 20 of them agreed to send in their MCQs. Moreover, only 10 of them decided to take part in the whole study. Comparison of percentages of IWFs in MCQs presented initially by teachers seemed to suggest higher quality of MCQs from teachers participating in the whole study. As a result, authors originally suspected that teachers who resigned felt overwhelmed upon realizing their mistakes. As mentioned above, every participant had a right to resign from further participation at any point in the study without giving reasons. Still, some teachers provided an explanation for their resignation and the most common was lack of time due to other obligations. Some teachers even admitted that they would expect an expert in the field to correct their questions for them.

Study Limitations

We acknowledge that our study had several limitations. First, it was a single-center study involving only the academic teachers of PUMS and extension of its results on wider population would be possible only after further studies. In addition, the number of male participants in the study was low. However, given that the aim of the study was only to evaluate the proposed intervention, we decided not to introduce any limitations with regard to the gender of participants. What is more, an occurrence of the selection bias to the study cannot be excluded as those who decided to partake in the study were likely to be more interested in improving their skills and quality of MCQs and as a result could have been more fond of the tool. Furthermore, as the interviews were conducted by authors of the document, it cannot be excluded that some participants, trying to please the interviewer, could avoid bringing up negative aspects of the intervention and highlight only its positive aspects. However, to compensate for that, a separate question in the interview template was dedicated to negative aspects of the intervention. During the interviews, the importance of respondents’ critical comments for improving the document was also emphasized. These last two potential biases should also be contrasted with promising results of the evaluation of the quality of MCQs before and after the intervention. We are also aware of the risk of the author bias given that described tool was our work. However, our goal was to perfect the document to make it the most useful for academic teachers. As a result, we attempted to take a critical look at it and analyze the data more objectively.

Conclusion

This pilot study showed that a short guideline document distributed among academic teachers may serve as an intervention to improve the quality of MCQs by significantly reducing number of IWFs in them. Participants of the study were open to its concept and saw the potential for creating similar tools.
in the future. Medical teachers may find guidelines helpful and useful in their work, especially when they are brief and simply written. Contrary to workshops or training sessions, they seem to offer more flexibility to medical teachers, which is important due to their busy schedule and numerous obligations. Teachers can use the guideline whenever they find some time and come back to it while writing new test questions. Meanwhile, constant professional development is required to improve the knowledge and awareness of medical teachers on good practices of MCQ writing and IWFs.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Students’ Scientific Society of Poznan University of Medical Sciences (Grant No. 502-05-04102108-05274).

ORCID iDs
Piotr Przymuszala https://orcid.org/0000-0001-8785-724X
Ryszard Marciniak https://orcid.org/0000-0001-7878-7403

Supplemental Material
Supplemental material for this article is available online.

References
Abdulghani, H. M., Ahmad, F., Irshad, M., Khalil, M. S., Al-Shaikh, G. K., Syed, S., Aldrees, A. A., Alrowais, N., & Haque, S. (2015). Faculty development programs improve the quality of multiple choice questions’ writing. Scientific Reports, 5, 9556.
Abdulghani, H. M., Irshad, M., Haque, S., Ahmad, T., Sattar, K., & Khalil, M. S. (2017). Effectiveness of longitudinal faculty development programs on MCQs items writing skills: A follow-up study. PLOS ONE, 12, Article e0185895.
Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In J. Kuhl & J. Beckmann (Eds.), Action control (pp. 11–39). Springer.
Ali, A. M. (1981). The use of positive and negative examples during instruction. Journal of Instructional Development, 5, 2–7.
British Educational Research Association. (2012). Ethical guidelines for educational developers.
Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). How to prepare better multiple-choice test items: Guidelines for university faculty. https://testing.byu.edu/handbooks/betteritems.pdf
Campbell, D. E. (2011). How to write good multiple-choice questions. Journal of Paediatrics and Child Health, 47, 322–325.
Case, S. M., & Swanson, D. B. (2002). Constructing written test questions for the basic and clinical sciences. National Board of Medical Examiners.
Cleland, J. A. (2017). The qualitative orientation in medical education research. Korean Journal of Medical Education, 29, 61–71.
Coughlin, P. A., & Featherstone, C. R. (2017). How to write a high quality multiple choice question (MCQ): A guide for clinicians. European Journal of Vascular and Endovascular Surgery, 54, 654–658.
Dellinges, M. A., & Curtis, D. A. (2017). Will a short training session improve multiple-choice item-writing quality by dental school faculty? A pilot study. Journal of Dental Education, 81, 948–955.
Dorri, S., Akbari, M., & Sedeh, M. (2016). Kirkpatrick evaluation model for in-service training on cardiopulmonary resuscitation. Iranian Journal of Nursing and Midwifery Research, 21, 493–497.
Epstein, R. M. (2007). Medical education—Assessment in medical education. The New England Journal of Medicine, 356(4), 387–396.
Iramanecrat, C. (2012). The impact of item writer training on item statistics of multiple-choice items for medical student examination. Siriraj Medical Journal, 64, 178–182.
Jozelewicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in-house medical school examinations. Academic Medicine, 77, 156–161.
Kheyami, D., Jaradat, A., Al-Shibani, T., & Ali, F. A. (2018). Item analysis of multiple choice questions at the department of paediatrics, Arabian Gulf University, Manama, Bahrain. Sultan Qaboos University Medical Journal, 18, e68–e74.
Kirkpatrick, J., & Kirkpatrick, W. (2016). Kirkpatrick’s four levels of training evaluation. Association for Talent Development Press.
Kowash, M., Hussein, I., & Al Halabi, M. (2019). Evaluating the quality of multiple choice question in paediatric dentistry post-graduate examinations. Sultan Qaboos University Medical Journal, 19, 135–141.
Kumar, P., Sharma, R., Rana, M., & Gajjar, S. (2014). Item and test analysis to identify quality multiple choice questions (MCQS) from an assessment of medical students of Ahmedabad, Gujarat. Indian Journal of Community Medicine, 39, 17–20.
Lancaster, J. W., Stein, S. M., MacLean, L. G., Van Amburgh, J., & Dorri, S. (2017). Kirkpatrick evaluation model for in-service training on cardiopulmonary resuscitation. Medical Teacher, 12, Article e0185895.
Malau-Aduli, B. S., & Zimitat, C. (2012). Peer review improves advance teaching and learning within health science programs. American Journal of Pharmaceutical Education, 78, 1–7.
Malau-Aduli, B. S., & Zimitat, C. (2012). Peer review improves the quality of MCQ examinations. Assessment & Evaluation in Higher Education, 37, 919–931.
McCray, K. H. (2016). Gallery educators as adult learners: The active application of adult learning theory. Journal of Museum Education, 41, 10–21.
Moreau, K. A. (2017). Has the new Kirkpatrick generation built a better hammer for our evaluation toolbox? Medical Teacher, 39, 999–1001.
Naeem, N., van der Vleuten, C., & Alfaris, E. A. (2012). Faculty development on item writing substantially improves item quality. Advances in Health Sciences Education, 17, 369–376.
Nedeau-Cayo, R., Laughlin, D., Rus, L., & Hall, J. (2013). Assessment of item-writing flaws in multiple-choice questions. Journal for Nurses in Professional Development, 29, 52–57.
O’Brien, B. C., Harris, I. B., Beckman, T. J., Reed, D. A., & Cook, D. A. (2014). Standards for reporting qualitative research: A synthesis of recommendations. *Academic Medicine, 89*, 1245–1251.

Pho, V., Ligozat, A., & Grau, B. (2015, June 21–25). Distractor quality evaluation in multiple choice questions [Conference session]. International Conference on Artificial Intelligence in Education, Madrid, Spain. https://hal.archives-ouvertes.fr/hal-01631779

Prashanth, H. (2013). Faculty development trends in medical education: A review. *South East Asian Journal of Medical Education, 7*, 11–16.

Steinert, Y., Mann, K., Anderson, B., Barnett, B. M., Centeno, A., Naismith, L., Prideaux, D., Spencer, J., Tullo, E., Viggiano, T., & Ward, H. (2016). A systematic review of faculty development initiatives designed to enhance teaching effectiveness: A 10-year update: BEME guide no. 40. *Medical Teacher, 38*, 769–786.

Steinert, Y., Mann, K., Centeno, A., Dolmans, D., Spencer, J., Gelula, M., & Prideaux, D. (2006). A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education: BEME guide no. 8. *Medical Teacher, 28*, 497–526.

Tariq, S., Tariq, S., Maqsood, S., Jawed, S., & Baig, M. (2017). Evaluation of cognitive levels and item writing flaws in medical pharmacology internal assessment examinations. *Pakistan Journal of Medical Sciences, 33*, 866–870.

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today, 26*, 662–671.

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education, 42*, 198–206.

Tarrant, M., & Ware, J. (2012). A framework for improving the quality of multiple-choice assessments. *Nurse Educator, 37*, 98–104.

Taylor, D. C. M., & Hamdy, H. (2013). Adult learning theories: Implications for learning and teaching in medical education: AMEE guide no. 83. *Medical Teacher, 35*(11), e1561–e1572.

Tenzin, K., Dorji, T., & Tenzin, T. (2017). Construction of multiple choice questions before and after an educational intervention. *Journal of the Nepal Medical Association, 56*, 112–116.