





Experiences from Two Ways of Integrating Pre- and Post-course Multiple-choice Assessment Questions in Educational Events for Surgeons

Monica Ghidinelli ^a, Michael Cunningham ^b, Isobel C. Monotti^b, Nishma Hindocha ^c, Alain Rickli^a,
Iain McVicar^d and Mark Glyde ^b

^aAO Education Institute, AO Foundation, Duebendorf, Switzerland; ^bCollege of Veterinary Medicine/School of Veterinary and Life Sciences, Murdoch University, Murdoch, Australia; ^cDepartment of Oral and Maxillofacial Surgery, Rigshospitalet, University Hospital, Copenhagen, Denmark; ^dMaxillofacial Unit, Queen's Medical Centre, Nottingham University Hospitals NHS Trust, Nottingham, UK

ABSTRACT

To examine how to optimise the integration of multiple-choice questions (MCQs) for learning in continuing professional development (CPD) events in surgery, we implemented and evaluated two methods in two subspecialties over multiple years. The same 12 MCQs were administered pre- and post-event in 66 facial trauma courses. Two different sets of 10 MCQs were administered pre- and post-event in 21 small animal fracture courses. We performed standard psychometric tests on responses from participants who completed both the pre- and post-event assessment. The average difficulty index pre-course was 57% with a discrimination index of 0.20 for small animal fractures and 53% with a discrimination index of 0.15 for facial trauma. For the majority of the individual MCQs, the scores were between 30%-70% and the discrimination index was >0.10. The difficulty index post-course increased in both groups (to 75% and 62%). The pre-course MCQs resulted in an average score in the expected range for both formats suggesting they were appropriate for the intended level of difficulty and an appropriate pre-course learning activity. Post-course completion resulted in increased scores with both formats. Both delivery methods worked well in all regions and overall quality depends on applying a solid item development and validation process.

ARTICLE HISTORY

Received 18 February 2021

Revised 9 April 2021

Accepted 13 April 2021

KEYWORDS

MCQ; assessment; cpd; cme; psychometric test; difficulty index; validation

Introduction


Multiple-choice questions (MCQs) are one of the most widely used categories of instrument for assessing learning outcomes for physicians at the postgraduate level [1]. Although it is not possible to assess everything that a physician is expected to know, learning can be driven through a validated assessment system, or blueprint, that deliberately samples representative knowledge and skills [2]. Shumway et al. recommend that items (test questions) should be set in the context of patient scenarios and that validity, reliability, impact on learning, and feasibility, including cost, should be considered when selecting instruments for assessing learning outcomes [1]. MCQs indeed offer a cost-efficient testing format with high validity and reliability, if the questions meet appropriate quality criteria [3]. They can be a reliable form of testing theoretic knowledge and clinical reasoning and can therefore form a component of clinical competency assessment [4]. In

addition, MCQs are an efficient and objective approach for assessing a broad range of topics, making this an appealing method for driving learning in the workplace [2], for assessing certified programmes [5], for test-enhanced learning [6], and as a predisposing activity before an educational event [7].

Test-enhanced learning is being increasingly incorporated into continuing medical education (CME) and continuing professional development (CPD) courses due to its demonstrated impact on knowledge retention and its value in assessing the efficacy of the course in imparting prescribed learning outcomes [3,6,8,9]. The introduction of pre-course assessment identifies gaps in knowledge and enhances learning at the course [6,10]. Similarly, feedback from post-course assessment allows the participant to correct conceptual misunderstandings and promotes informed self-assessment [6]. In surgical education, the problem-solving skills involved in the clinical reasoning process need shaping and perfecting through repeated practice

CONTACT Monica Ghidinelli  monica.ghidinelli@aofoundation.org  Stettbachstrasse 6, 8046 Duebendorf, Switzerland, 41 79 5764433

*These authors contributed equally to this work

 Supplemental data for this article can be accessed [here](#).

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

and feedback to become effective and efficient. Education seems most efficient when it is undertaken in the context of future tasks and results in better retention of knowledge[11]. Finally, the perceived relevance of working with medical problems and the challenge of solving problems provide a strong motivation for learning[11].

Although MCQs are widely used, quality varies. Creating high-quality items is challenging and time-consuming. Many examples in the literature report poorly constructed MCQs that, for example, provide prompts to the correct answer or only superficially test knowledge [4,12]. To avoid these errors, guidelines for producing effective items have been developed and training faculty is recommended [13,14]. In addition, psychometric analysis of the items helps to identify which questions are performing well and those that need improvement. The difficulty index (also called *p*-value or power) describes the percentage of participants who correctly answer an MCQ and ranges from 0% to 100%. The lower the percentage, the harder the item and vice versa[15]. Random guessing should result in 25% of the participants answering an item with four options correctly; therefore, it is reasonable to have a difficulty index between the recommended range of 30%–70%. The discrimination index of an item (also called point biserial) describes the relationship between getting an MCQ correct and the overall score. It ranges between -1.00 and $+1.00$ with values greater than $+0.10$ indicating that the item is performing well. An item with a high value of discrimination indicates that participants who had high test scores got the item correct while participants who had low test scores got the item incorrect [15,16]. Distractor efficiency is another parameter to consider. A distractor (an incorrect option) is not efficient if it is selected by less than 5% of the participants thus it should be reviewed [17]. The presence of non-functioning distractors (NFDs) affects difficulty and discrimination indices and therefore should be avoided. Reliability describes the ability of an instrument to measure consistently. The internal consistency of the items in a test is calculated by Cronbach's alpha coefficient. This ranges from 0 to 1, where 0 represents no correlation between the items and 1 represents a significant covariance. For high-stakes exams, an alpha coefficient of 0.8 or more is desired [18] and can be increased by expanding the number of items given in an exam [3]. In summary, an excellent MCQ has an average difficulty index (between 30% and 70%), a high discrimination index (≥ 0.30), 3 functioning distractors, and a high Cronbach's alpha coefficient (≥ 0.8)[15]. However, for low stake assessment with 10–12 items, a discrimination index above 0.10 and a Cronbach's alpha coefficient above 0.4 are acceptable[19].

Assessment is a component of competency-based education at the AO Foundation and its clinical

divisions. We integrated formative assessment (measurement for the purpose of improvement) with the goals of showing participants their current level of knowledge and to enhance their learning through reflection on their gaps in our standard evaluation and assessment system that supports curriculum implementation [20,21]. Each learner coming to the educational experience takes a pre-assessment that provides feedback and motivation for the course. The event chairperson receives the results of the MCQs as well as self-assessed ability and competency gap scores 2 weeks before each event begins and this helps fine-tune the content [10,22]. Post-event evaluation and assessment data are provided to the event chairperson and curriculum planning committee to help the design of future events.

Our MCQs are typically vignette format or recall items as classified by the National Board of Medical Examiners (NBME) approach, with 4 answer options (3 distractors and 1 correct/preferred option) [14,23]. Feedback is provided after each question, covering the correct and incorrect (or less preferred) options, and the references for the rationale. Supplementary figures 1 and 2 show examples of MCQ items with an extract from one pre-event and one post-event report.

CME/CPD courses are delivered by AO CMF and AO VET for surgeons and veterinarians in the principles and techniques of bone fracture fixation. In 2016, the courses on the principles of small animal fracture management (a 3-day event) and the management of facial trauma (a 2-day event) were modified using a competency-based curriculum approach[24]. Pre- and post-course MCQs were integrated as a method of formative assessment in both courses. All of the items developed were of the vignette type, except for one question in the pre-course assessment for small animal fracture management[14]. The vignette questions were case-based and assessed clinical reasoning, which is why we preferred this type of item.

The two planning committees introduced MCQs differently, with the facial trauma group utilising one set of 12 identical question before and after the course and the veterinary group using two different sets of 10 questions of similar difficulty level in the pre- and post-course assessment. Both techniques have been shown to achieve an improvement in post-course scores [25,26].

The purpose of this study was to report the experiences in developing and implementing MCQs in the two curricula, to evaluate the quality of the MCQs, determine whether participant scores improved after the course, to compare the two ways of implementing MCQs in the two curricula for enhancing learning in

CME/CPD in fracture management, and to share our recommendations for integrating MCQs into CME/CPD activities.

Materials and Methods

Educational intervention

The AO CMF management of facial trauma course is a 2-day event. Before the course, participants received a set of preparation readings, online materials, and a pre-course assessment (7 profiling/self-assessment questions and 12 MCQs). The course is divided into three modules. Module 1 (8 hours) focuses on mid and upper face trauma. Module 2 (5.5 hours) focuses on mandibular trauma. Module 3 (1.5 hours) addresses paediatric fractures and panfacial fractures. Content is delivered through 12 lectures, 4 small group discussions (13 cases), and 6 hands-on practical exercises, with almost equal time allocation. Participants complete a post-course evaluation (8 questions and 12 MCQs).

The AO VET principles in small animal fracture management course is a 3-day event. Before the course, participants received a set of preparation readings, online materials, and a pre-course assessment (7 profiling/self-assessment questions and 10 MCQs). The course is divided into five modules. Module 1 (8 hours) focuses on the basics of fracture treatment. Module 2 (6.5 hours) addresses diaphyseal fractures. Module 3 (2 hours) explores articular and juxta-articular fractures. Module 4 (4.5 hours) focuses on miscellaneous fracture treatments and module 5 (3.5 hours) on managing complications. Content is delivered through 23 lectures, 3 small group discussions (11 cases), 2 plenary case discussions (6 cases), a skill lab, and 6 hands-on practical exercises. Participants complete a post-course evaluation (8 questions and 10 MCQs).

MCQ development

The AO Assessment Toolkit [10] and the idea of using MCQs to prepare for educational events and to help in post-event evaluation were introduced to the members of the two curriculum planning committees who were developing the small animal fracture (AO Vet) or the facial trauma curricula (AO CMF). The approach of creating two MCQs per curriculum competency or event objective (1 easy and 1 difficult) was chosen. Key recommendations for item content and structure were provided: 1) The questions should focus on the patient problems in practice and use case scenarios to present clinical decisions; 2) Images (x-rays, MRIs, photos) should be included if needed in order to answer the question; 3) Each item should have

four answer options with one correct/preferred answer; 4) Each option should be referenced; 5) Each item should include the feedback rationale that explains the correct and incorrect answer options; 6) Best practice question-writing principles should be followed to avoid bias and common mistakes (e.g. avoid phrasing in the negative with “except” and no “all of the above”); 8) All proposed questions should be peer reviewed by a panel of faculty (aiming for 80% agreement on the preferred/correct answer). An example question was shown and discussed, and the planning committees developed one or two items together with an educationalist. The facial trauma committee agreed to create one set of 12 items while the small animal fracture committee two sets of 10 items (one pre- and one post-event). Questions were allocated to committee members and other faculty to develop using a template structure and the above recommendations.

Validation and testing

The questions were submitted to a lead editor who worked with a MCQ expert and educationalist to consolidate the items that were then peer reviewed by the respective planning committee (face validation). For facial trauma, the input and consensus from the committee were integrated and the items were finalised for online administration. All items were pilot tested with participants in several courses. The performance data were reviewed by the lead editor who presented suggested changes to the planning committee.

For the small animal fractures, the items were also sent to a panel of faculty who were asked to answer them and to rate if the question was appropriate, needed small changes, or should be replaced (content validation). The items were adapted where necessary and a second round of faculty review was performed.

Implementation and data collection

Following validation and revision, the MCQs were integrated and administered online using SurveyMonkey as part of our standard pre- and post-event evaluation and assessment process in facial trauma courses from January 2017 and in small animal fracture courses from January 2018. Pre-event assessment administration starts 30 days before each event and reminders may continue until 3 days before the event starts. Each participant is invited via a unique link and is informed that responses are pooled and shared with the faculty and may also be used for research publications. Post-event administration starts 1 day after the event and remains open for 16 days. All the responses were saved in our central management information system.

To ensure consistent data collection in the many courses worldwide and to gather adequate data for

a thorough review of item performance, the planning committees decided that changes to the questions would be considered based on data and findings after 2 or 3 years of implementation. The MCQs were translated by experienced faculty into French, German, Italian, and Chinese for small animal fractures and Spanish and French for facial trauma.

During the first year, the items were also offered to faculty at the courses in order to gather more feedback and performance data.

Analysis

A retrospective analysis was performed on anonymised data from 21 small animal fracture events in English (2018–2019) and from 66 facial trauma courses in English or Spanish (2017–2019). Standard psychometric tests (difficulty index, discrimination index, non-functioning distractors, Cronbach's alpha coefficient) were performed using the Lertap 5 software (Curtin University, Perth, Western Australia) on a dataset including only participants who had completed both the pre- and post-event assessments (a total of 422 participants from small animal fracture courses and 723 participants from facial trauma courses). Differences between pre-test and post-test scores were investigated using paired t-tests. In addition, responses from faculty were analysed and compared to participants using unpaired t-tests. Data from both courses were compared and analysed and conclusions were drawn regarding what worked well and what best practices could be identified.

Ethical approval

Ethics exemption was granted from Murdoch University, Western Australia (number 2020–183)

Results

To evaluate the quality of the developed MCQs we performed a standard psychometric analysis from matching pre- and post-responses from 723 participants in the facial trauma courses and 422 participants in the small animal courses.

For the small animal course, participants take two different sets of 10 MCQs pre- and post-event with items matched based on the content covered. The difficulty index (percentage of correct answers) was below 30% for one pre-course MCQ (C5-Q2) and above 70% for four items. Six post-course items had a difficulty index above 70%. The remaining items performed within the 30–70% range (Table 1).

To determine if the pre- and post-event questions have a matched difficulty index, we additionally compared the results of the two sets administered

“inverted” to similar groups of participants before one course in 2017 and 2018. Only two items were similar (less than 5% difference in difficulty), five were easier in the current pre-event set while three were more difficult (Table 2). The average difficulty of the two sets, however, was similar. Therefore, while an item-by-item comparison is not appropriate, it was reasonable to look at overall changes in performance on the complete sets. The average difficulty indices were 57% pre-course and 75% post-course with a gain of 18% overall ($p = 0.063$) (Table 1).

For the facial trauma courses, participants take the same 12 MCQs pre- and post-event, however rationale feedback is given only after the event. The difficulty index was below 30% for two pre-course MCQs (C5-Q1 and C5-Q2) and above 70% for two items (C1-Q1 and C3-Q1) (Table 1). The remaining eight items performed within 30–70%. Post-event difficulty indexes are expected to increase compared with pre-event one. C5-Q2 had a difficulty index below 30% also after the course, suggesting that this is a difficult or a confusing question. One item, C1-Q2 showed a decrease in the difficulty index of 3% (opposite of what is expected), three items increased less than 5% (C1-Q1, C5-Q2, C6-Q2), while the remaining eight items showed modest increases, with item C6-Q1 showing the greatest improvement at 29%. The average difficulty index was 53% pre-course and 62% post-course with a significant gain of 9% overall ($p < 0.01$) (Table 1).

The average discrimination indices (point biserial correlation between the correct answer and overall exam score) pre-course were 0.20 for small animal and 0.15 for facial trauma while post-course were 0.11 for small animal and 0.16 for facial trauma (Table 1). The majority of the MCQs had a positive discrimination index as expected although three fell below 0.10, the usual threshold for acceptable discrimination (C1-Q2 in facial trauma, and C1-Q1 and C5-Q2 in small animal post-event).

We also analysed the number of non-functioning distractors (NFDs, incorrect options chosen by less than 5% of the participants) for each item. Small animal had five MCQs with one NFD in the pre-event set, and two items with three NFDs, three with two NFDs, and two with one NFD in the post-event set (Table 1). Items with two or more NFDs should be revised. Facial trauma had three MCQs with two NFDs and five with one NFD.

The average reliability test (Cronbach's alpha coefficient) result was above 0.27 for small animal post-event and above 0.42 for all the other sets (Table 1) indicating that the small animal post-event test was performing less well than the others.

Table 1. Performance of assessment MCQs in facial trauma (2017-2019, n = 723) and small animal fracture management events (2018-2019, n= 422). C1 - Q1 = Competency 1 - Question 1

Small animals - Pre-event MCQs							Small animals - Post-event MCQs						
MCQ	Difficulty index	Discrimination index	Non-functioning distractors	MCQ	Difficulty index	Discrimination index	Non-functioning distractors	MCQ	Difficulty index	Discrimination index	Non-functioning distractors	Change	
C1 - Q1	0.73	0.24	1	C1 - Q1	0.91	- 0.01	2	C1 - Q1	0.91	- 0.01	2	0.18	
C1 - Q2	0.73	0.18	0	C1 - Q2	0.71	0.16	1	C1 - Q2	0.71	0.16	1	-0.03	
C2 - Q1	0.69	0.25	0	C2 - Q1	0.89	0.21	2	C2 - Q1	0.89	0.21	2	0.20	
C2 - Q2	0.74	0.15	1	C2 - Q2	0.96	0.21	3	C2 - Q2	0.96	0.21	3	0.22	
C3 - Q1	0.47	0.21	1	C3 - Q1	0.51	0.10	0	C3 - Q1	0.51	0.10	0	0.04	
C3 - Q2	0.32	0.19	0	C3 - Q2	0.91	0.08	3	C3 - Q2	0.91	0.08	3	0.60	
C4 - Q1	0.50	0.24	0	C4 - Q1	0.63	0.05	0	C4 - Q1	0.63	0.05	0	0.13	
C4 - Q2	0.42	0.16	1	C4 - Q2	0.94	0.15	3	C4 - Q2	0.94	0.15	3	0.52	
C5 - Q1	0.87	0.13	2	C5 - Q1	0.55	0.02	1	C5 - Q1	0.55	0.02	1	-0.32	
C5 - Q2	0.25	0.24	0	C5 - Q2	0.45	0.14	1	C5 - Q2	0.45	0.14	1	0.20	
Average	0.57	0.20	0	0.75	0.11	0.11	1	0.75	0.11	0.11	1	0.17	
Std.Dev.	0.20	0.04		0.19	0.07			0.19	0.07				
Reliability (coefficient alpha) = 0.486							Reliability (coefficient alpha) = 0.277						
Facial trauma - Pre-event MCQs							Facial trauma - Post-event MCQs						
MCQ	Difficulty index	Discrimination index	Non-functioning distractors	MCQ	Difficulty index	Discrimination index	Non-functioning distractors	MCQ	Difficulty index	Discrimination index	Non-functioning distractors	Change	
C1 - Q1	0.74	0.14	2	C1 - Q1	0.75	0.25	2	C1 - Q1	0.75	0.25	2	0.01	
C1 - Q2	0.45	0.02	1	C1 - Q2	0.42	0.10	1	C1 - Q2	0.42	0.10	1	-0.03	
C2 - Q1	0.47	0.12	1	C2 - Q1	0.55	0.15	1	C2 - Q1	0.55	0.15	1	0.08	
C2 - Q2	0.66	0.25	0	C2 - Q2	0.75	0.24	0	C2 - Q2	0.75	0.24	0	0.10	
C3 - Q1	0.82	0.17	2	C3 - Q1	0.89	0.17	2	C3 - Q1	0.89	0.17	2	0.06	
C3 - Q2	0.63	0.13	2	C3 - Q2	0.79	0.20	2	C3 - Q2	0.79	0.20	2	0.16	
C4 - Q1	0.50	0.15	0	C4 - Q1	0.64	0.12	1	C4 - Q1	0.64	0.12	1	0.15	
C4 - Q2	0.57	0.16	1	C4 - Q2	0.70	0.08	2	C4 - Q2	0.70	0.08	2	0.14	
C5 - Q1	0.27	0.14	1	C5 - Q1	0.35	0.15	1	C5 - Q1	0.35	0.15	1	0.08	
C5 - Q2	0.26	0.10	1	C5 - Q2	0.29	0.13	1	C5 - Q2	0.29	0.13	1	0.04	
C6 - Q1	0.58	0.26	1	C6 - Q1	0.86	0.19	2	C6 - Q1	0.86	0.19	2	0.29	
C6 - Q2	0.42	0.18	0	C6 - Q2	0.45	0.16	0	C6 - Q2	0.45	0.16	0	0.04	
Average	0.53	0.15	0	0.62	0.16	0.16	0	0.62	0.16	0.16	0	0.09	
Std.Dev.	0.17	0.06		0.19	0.05			0.19	0.05				
Reliability (coefficient alpha) = 0.421							Reliability (coefficient alpha) = 0.433						

Table 2. Comparison of the difficulty index of the two sets of 10 MCQs for small animal fractures.

MCQ	2017 (inverted) (n = 86)	2018 (n = 73)	Difference between 2018 and 2017
C1 – Q1	80.23	79.16	-1.07
C1 – Q2	66.28	84.29	18.01
C2 – Q1	56.98	75.71	18.73
C2 – Q2	56.98	74.29	17.31
C3 – Q1	36.47	47.14	10.67
C3 – Q2	85.88	50.00	-35.88
C4 – Q1	70.59	67.14	-3.45
C4 – Q2	84.71	27.14	-57.57
C5 – Q1	75.29	92.75	17.46
C5 – Q2	62.35	42.03	-20.32
Average	67.576	63.97	-3.61
Standard dev.	15.20	21.19	

C1 – Q1 = Competency 1 – Question 1

The two different MCQ sets, covering the same competences, were administered before the event to two similar groups of participants in 2017 and 2018. Inverted means post-event MCQs administered before the event. .

The MCQs were delivered in different languages based on the course location. We therefore asked if there could be a difference between them. Comparisons between Asia Pacific, Europe and Southern Africa, and Latin America showed similar outcomes for small animal fractures (Supplementary figure 3). Regional breakdowns were also very consistent for facial trauma MCQs (Supplementary Table 1). There was a slightly larger pre- to post-course gain with courses using the Spanish version (pre-53%, post-65%, n = 154) of the facial trauma MCQs compared with the English version (pre = 53%, post = 61%, n = 569) (Supplementary Table 1).

To help validate the items during the first year of implementation, faculty were asked to answer the items and to rate each question in one of three categories: “It’s good (clear and fair) – keep it” (% Good), “It’s average – keep it or make minor changes” (% OK), “It’s unclear, complicated, too unusual etc – replace it with a different question” (content validity) (Table 3). The ratings for each assessment item were variable with question C4-Q1 of the pre-course MCQs being the lowest rated for small animal, and C5-Q1 and C5-Q2 the lowest for facial trauma (Table 3, % Good and % OK). The difficulty index for faculty ranged between 70% and 100% for small animal questions, and between 47% and 95% for facial trauma (Table 3). In addition to determine construct validity, we compared the difficulty indexes for faculty and participants (Figure 1). For all sets, the average difficulty index for faculty was significantly higher than for participants (facial trauma faculty 77%±16%, participants 53%±17%, $p < 0.001$; small animal pre-event faculty 91%±8%, participants 57%±21, $p < 0.01$; small animal post-event faculty 87%±11%, participants 68%±15%, $p < 0.001$; one-tailed unpaired Student's t-test). However, by taking a closer look at the single items

for small animal fracture management, three items were not able to discriminate between faculty and participants (pre-event C5-Q1 (Figure 1(b)) and post-event C1-Q1 and C4-Q2 (Figure 1(c)) and should be revised.

The faculty data for each question were tabulated beside the psychometric results from participants (from Table 1). This helped to identify MCQs with suboptimal performance and recommendations were added (Table 3).

Discussion

The purpose of our analyses of the two ways of using pre- and post-event MCQs was threefold: to evaluate the quality of the MCQs, to determine whether learning occurred (if participant knowledge and gaps improved after the course), and to provide recommendations for integrating MCQs into CME/CPD activities.

The items developed by the two planning committees were case-based vignettes, except for one question in the pre-course assessment for small animal fracture management, which was a recall item.

The difficulty and discrimination indices, along with the distractor efficiency, were used to determine MCQ quality and to highlight which questions would need further improvement [15,27]. The average difficulty index was in the acceptable range (30–70%) for the facial trauma pre- and post-event sets and for the small animal fracture management pre-event set. This range represents a good target level before the event to motivate participants on the topics and to help them and the faculty to identify gaps. The difficulty index for the small animal fracture management post-event set was 75%, which is above the recommended range of difficulty. Looking at single items within the facial

Table 3. Performance of assessment items (faculty content validation, psychometric scores, recommendations)

	Faculty difficulty index (Dif I) and rating (n=36)				Participants Psychometric tests				Quality and Recommendations	
	Dif I	% Good	% OK		Dif I	Dis I	NFD			
AO Principles in Small Animal Fracture Management - Pre										
C1-Q1	93%	89%	8%		73%	0.24	1		Easy item with 1 NFD – possibly revise	
C1-Q2	95%	85%	15%		73%	0.18	0			
C2-Q1	100%	57%	34%		69%	0.25	0			
C2-Q2	94%	77%	20%		74%	0.15	1		Easy item with 1 NFD - possibly revise	
C3-Q1	92%	82%	12%		47%	0.21	1		Moderately difficult item with 1 NFD - possibly revise	
C3-Q2	97%	88%	9%		32%	0.19	0			
C4-Q1	75%	54%	29%		50%	0.24	0			
C4-Q2	92%	74%	16%		42%	0.16	1		Moderately difficult item with 1 NFD – possibly revise	
C5-Q1	89%	84%	13%		87%	0.13	1		Easy item, similar Dif I of faculty and participants - revise	
C5-Q2	80%	70%	21%		25%	0.24	0		Difficult item - keep	
Average	91%	76%	18%		57%	0.20	-			
AO Principles in Small Animal Fracture Management - Post										
Faculty difficulty index (Dif I) and rating (n=23)										
	Dif I	% Good	% OK		Dif I	Dis I	NFD			
C1-Q1	80%	78%	13%		91%	-0.1	2		Easy item, negative Dis I and 2 NFDs - revise	
C1-Q2	92%	88%	12%		71%	0.16	1		Moderately easy item with 1 NFD – possibly revise	
C2-Q1	96%	96%	4%		89%	0.21	2		Easy item with 2 NFDs - revise	
C2-Q2	100%	92%	8%		96%	0.21	3		Easy item with 3 NFDs - revise	
C3-Q1	96%	91%	9%		51%	0.10	0			
C3-Q2	92%	63%	29%		91%	0.08	3		Easy item, similar Dif I of faculty and participants, poor Dis I, 3 NFDs - revise	
C4-Q1	79%	79%	21%		63%	0.05	0		Average item with poor Dis I - revise	
C4-Q2	70%	83%	8%		94%	0.15	2		Easy item, participants Dif I higher than faculty, 2 NFDs - revise	
C5-Q1	96%	91%	9%		55%	0.02	1		Average item with poor Dis I - revise	
C5-Q2	71%	64%	27%		45%	0.14	0			
Average	87%	83%	14%		75%	0.11	-			
AO Management of Facial Trauma - Pre										
Faculty difficulty index (Dif I) and rating (n=138)										
	Dif I	% Good	% OK		Dif I	Dis I	NFD			
C1-Q1	95%	84%	13%		74%	0.14	2		Moderately easy item with 2 NFDs - revise	
C1-Q2	68%	78%	18%		45%	0.02	1		Moderately difficult item with poor Dis I, 1 NFD - revise	
C2-Q1	71%	69%	20%		47%	0.12	0			
C2-Q2	91%	74%	19%		66%	0.25	0			
C3-Q1	95%	87%	10%		82%	0.17	2		Easy item with 2 NFDs - revise	
C3-Q2	91%	83%	13%		63%	0.13	2		Average item with 2 NFDs - revise	
C4-Q1	81%	73%	17%		50%	0.15	0			
C4-Q2	79%	74%	17%		57%	0.16	1		Average item with 1 NFD – possibly revise	
C5-Q1	47%	42%	35%		27%	0.14	1		Difficult item with 1 NFD – possibly revise	
C5-Q2	51%	50%	28%		26%	0.10	1		Difficult item, poor Dis I, 1 NFD - revise	
C6-Q1	82%	85%	13%		58%	0.26	1		Average item with 1 NFD – possibly revise	
C6-Q2	76%	81%	15%		42%	0.18	0			
Average	77%	73%	18%		53%	0.15	-			

Notes

C1-Q1 = Competency 1 - Question 1

Good = It's good (clear and fair) - keep it, OK = It's average - keep it or make minor changes, Dif I = difficulty index, Dis I = discrimination index, NFD = non-functioning distractor
 Easy items (Dif I >80%), moderately easy (Dif I 71%-80%), average (difficulty index, Dif I 50%-70%), moderately difficult (Dif I 30-50%) and difficult (Dif I <30%).

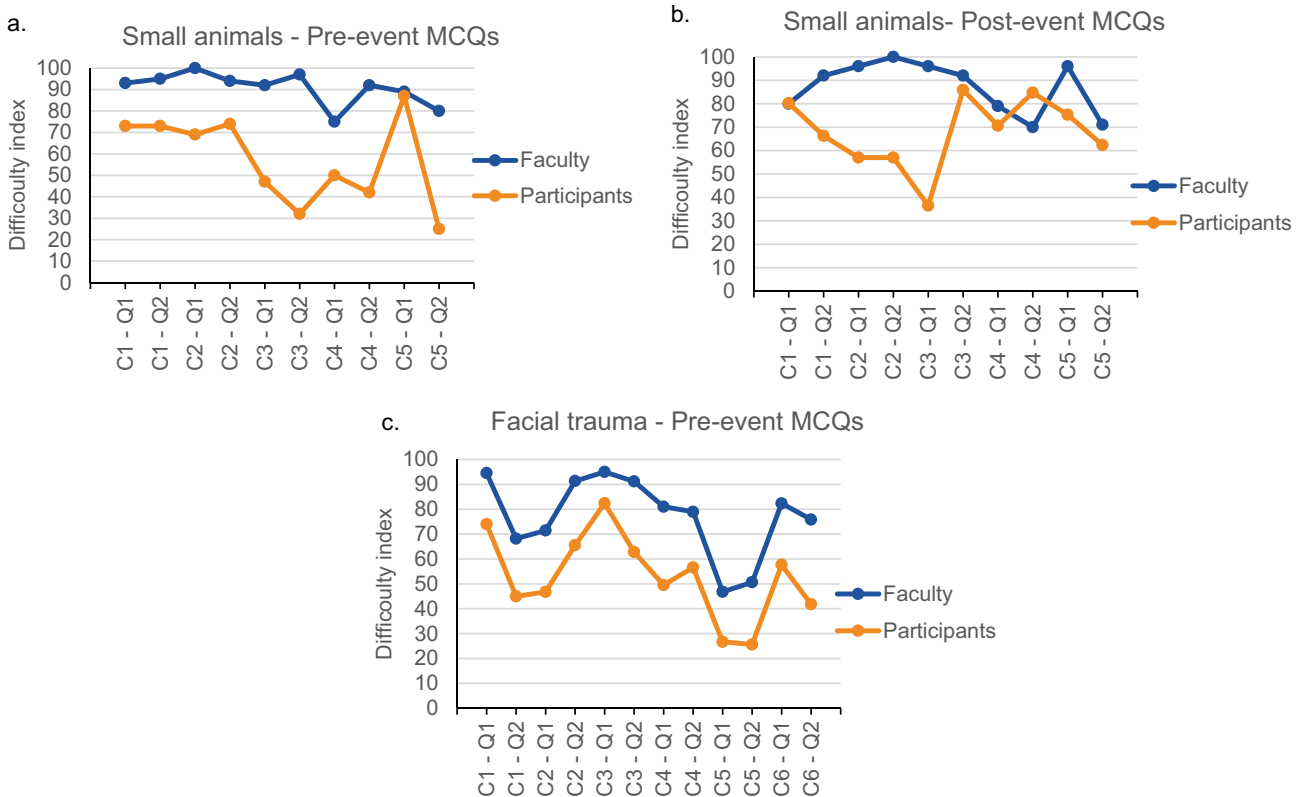


Figure 1. Validation of the MCQs by comparing faculty and participants results (construct validation). C = Competency, Q = Question.

a. Small animal fractures – Pre-event MCQs. $n = 38$ faculty and 422 participants. b. Small animal fractures – Post-event MCQs. The participant’s results from taking the inverted sets during one course in 2017. $n = 23$ faculty and 86 participants. Facial trauma – Pre-event MCQs. $n = 138$ faculty and 723 participants.

trauma and small animal pre-event sets, we found a balance of question difficulty close to the recommended range for an “ideal difficulty balanced exam” (50% average items (difficulty index, Dif I 50%-70%), 20% moderately easy items (Dif I 71%-80%), 20% moderately difficult items (Dif I 30–50%), 5% easy items (Dif I > 80%), and 5% difficult items (Dif I < 30%)) [27,28]. The small animal post-event had instead a high number of easy items (5 items out of 10 with Dif I above 80%). In addition, the intent of having one easier and one more difficult item for each competency was achieved in the facial trauma MCQ set excluding competency five, where both questions were difficult.

Item discrimination, which describes the relationship between getting an MCQ correct and the overall score, was acceptable for most of the items with a few exceptions (items that should be revised). In particular, one item had a negative value that suggests participants who had high tests scores got the item wrong (the opposite of what we expect). Overall, we found 62% of the items with at least 1 non-functioning distractor (NFD), i.e. an obviously wrong or “throw away” option

that even participants with low knowledge can easily avoid. In general, the presence of an NFD should be avoided; however, this should be evaluated in relation to the difficulty index. For example, if an item is within the expected range of difficulty regardless of the presence of one NFD, the distractor could be kept. Nevertheless, MCQs with two or more NFDs should be revised. A strategy to decrease the number of NFDs might be to reduce the number of answer options to three[29]. However, a concern for educators might be the increased odds that guessing has with 3-option versus 4-option MCQs (i.e. 33% vs. 25%).

Well-designed MCQs can effectively measure the learning that takes place as a result of attending a learning activity as well as the participant’s abilities and efforts[30]. On the other hand, results of poorly performing items might be difficult to interpret.

Knowledge gain, considered as an increase between post-test and pre-test scores, was significantly increased for facial trauma (9% gain, average difficulty index 53% pre- and 62% post-event) and for small animal fracture management (18% gain, 57% pre- and 75% post-event). These results suggest that participants’ knowledge

improved in the combined content domains covered by the course and are consistent with results previously reported for surgical residents [10] and other disciplines [17,31–34].

Limitations

While we analysed the data with rigorous psychometric analysis, the number of questions was rather small. In addition, the interpretation of the data should consider that the majority of participants did not complete the MCQs in their first language. Gaining consensus about the correct answer from a panel of international faculty from all parts of the world and varied speciality training was a challenging endeavour. Perhaps the way each question is asked in the clinical scenarios may be affected by the practice setting, clinical experience and education as well as resource availability of the participant (there might be a difference in a reply to “what would you do next?” based on your local resources compared with “what would you do?” based on the current evidence base and equal access to all treatment options (local health system, patient’s insurance, and subspeciality areas of strength and expertise)).

Conclusions

We conclude that pre- and post-course MCQs (both methods) are beneficial for learners and faculty to prepare for educational events and to review the outcomes. Providing an explanation behind the correct and incorrect answers helps identify areas for learning and discussion. Table 4 shows a summary of

suggested strengths and weaknesses of the two methods evaluated in this study. The overall quality and validity of the items and the use of the information gathered for each event are critical components irrespective of assessment methodology. All of the data and detailed item statistics will be reviewed with the subject matter experts in the curriculum planning committees with the goal of revising or replacing poorly performing MCQs and adjusting or replacing NFDs. Our experience supports Pugh et al.’s suggestion that despite the increased emphasis on the use of workplace-based assessment in competency-based education, there is still an important role for the use of MCQs in the assessment of health professionals[35].

Our recommendations for integrating MCQs into CME and CPD activities are:

- Clearly articulate the goals of your assessment process (for example, formative assessment as a learning tool to help prepare participants and faculty for educational events)
- Develop the questions based on the curriculum competencies and learning objectives for the event – focus on the most important principles for addressing common and critical patient problems instead of rare scenarios or obscure treatments that may have worked but are not supported well by evidence

Table 4. Strengths and limitations of two approaches for pre- and post-event assessment using MCQs.

Perspective	Same pre- and post-MCQs	Different pre- and post-MCQs
Learners (participants)	<p>Strengths:</p> <ul style="list-style-type: none"> • More focused on key points • Less time required to complete pre-event <p>Limitations:</p> <ul style="list-style-type: none"> • No feedback before the event (answers and rationale hidden) • Fewer topic, fewer opportunities to receive key messages 	<p>Strengths:</p> <ul style="list-style-type: none"> • More topics covered • More feedback received <p>Limitations:</p> <ul style="list-style-type: none"> • More items to complete • Potential for repetition in areas where some learners have no gaps
Faculty	<p>Strengths:</p> <ul style="list-style-type: none"> • Helps identify any areas where participants have post-event gaps • Less time required to develop items <p>Limitations:</p> <ul style="list-style-type: none"> • Fewer opportunities to deliver key messages • Less overall information about participants 	<p>Strengths:</p> <ul style="list-style-type: none"> • More chance to identify topics where learners have gaps • More objective data to review <p>Limitations:</p> <ul style="list-style-type: none"> • More items to develop and reach consensus • More items to be familiar with before each event
Assessment strategy and system (and depending on the overall goals)	<p>Strengths:</p> <ul style="list-style-type: none"> • More accurate data for showing changes between pre and post • Fewer items to validate and manage <p>Limitations:</p> <ul style="list-style-type: none"> • No pre-event learning opportunity for participants • Less data for faculty to review for future enhancements 	<p>Strengths:</p> <ul style="list-style-type: none"> • More learning offered to participants – feedback before and after event • Possibly less focus on “increases in scores” in a non-exam setting <p>Limitations:</p> <ul style="list-style-type: none"> • Less accurate for measuring pre-post change (or requires more validation) • More items to validate and manage

- Define standards/values for your quality criteria and work with all faculty and committees to meet these (implement a quality checklist for development and review all outcome data after pilots and each year of use to plan changes)
- Validate all items by applying the minimal and appropriate processes and tests (we suggest a faculty difficulty index (agreement on the correct answers) of 80% for a topic with international scope and multi-speciality involvement and adequate pilot testing with typical target audiences)
- Take a pragmatic approach to implement a reproducible system and process that all faculty everywhere will be able to implement (ensure all faculty are aware of the core messages and consider time barriers and language and connectivity issues)

Disclosure Statement

Monica Ghidinelli, Michael Cunningham, and Alain Rickli are employees of the AO Foundation. No potential conflict of interest was reported by the other author(s).

Funding

This work was supported by the AO Foundation. The AO Foundation receives funding for education from Synthes GmbH.

ORCID

Monica Ghidinelli  <http://orcid.org/0000-0002-7378-6273>
 Michael Cunningham  <http://orcid.org/0000-0002-4275-0454>
 Nishma Hindocha  <http://orcid.org/0000-0002-7691-0751>
 Mark Glyde  <http://orcid.org/0000-0003-1433-7694>

References

- [1] Shumway JM, Harden RM. Association for Medical Education in E. AMEE Guide No. 25: the assessment of learning outcomes for the competent and reflective physician. *Med Teach*. 2003;25(6):569–584.
- [2] Humphrey-Murto S, Wood TJ, Ross S, et al. Assessment pearls for competency-based medical education. *J Grad Med Educ*. 2017;9(6):688–691.
- [3] Gerhard-Szep S, Guntsch A, Pospiech P, et al. Assessment formats in dental medicine: an overview. *GMS J Med Educ*. 2016;33. Doc65. DOI:10.3205/zma001064
- [4] Tangianu F, Mazzone A, Berti F, et al. Are multiple-choice questions a good tool for the assessment of clinical competence in Internal Medicine? *Ital J Aerosp Med*. 1994;25(1–2):105–112.
- [5] Bustraan J, Henny W, Kortbeek JB, et al. MCQ tests in Advanced Trauma Life Support (ATLS(c)): development and revision. *Injury*. 2016;47(3):665–668.
- [6] Feldman M, Fernando O, Wan M, et al. Testing test-enhanced continuing medical education: a randomized controlled trial. *Acad Med*. 2018;93(11S):S30–S36.
- [7] Moore DE Jr., Chappell K, Sherman L, et al. A conceptual framework for planning and assessing learning in continuing education activities designed for clinicians in one profession and/or clinical teams. *Med Teach*. 2018;40(9):904–913.
- [8] Rustici M, Wang V, Dorney K, et al. Application of frequent, spaced multiple-choice questions as an educational tool in the pediatric emergency department. *AEM Educ Train*. 2020;4(2):85–93.
- [9] Larsen DP. Planning education for long-term retention: the cognitive science and implementation of retrieval practice. *Semin Neurol*. 2018;38(4):449–456.
- [10] De Boer PG, Buckley R, Schmidt P, et al. Learning assessment toolkit. *J Bone Joint Surg Am*. 2010;92(5):1325–1329.
- [11] Beard JD, Robinson J, Smout J. Problem-based learning for surgical trainees. *Ann R Coll Surg Engl*. 2002;84(4):227–229.
- [12] DiSantis DJ, Ayoob AR, Williams LE. Journal Club: prevalence of flawed multiple-choice questions in continuing medical education activities of major radiology journals. *AJR Am J Roentgenol*. 2015;204(4):698–702.
- [13] Przymuszała P, Piotrowska K, Lipski D, et al. Guidelines on Writing Multiple Choice Questions: a Well-Received and Effective Faculty Development Intervention. *SAGE Open*. 2020;10(3):215824402094743.
- [14] Billings M, DeRuchie K, Haist S, et al. Constructing written test questions for the basic and clinical sciences. Philadelphia: National Board of Medical Examiners (NBME); 2016.
- [15] Sahoo DP, Singh R. Item and distracter analysis of multiple choice questions (MCQs) from a preliminary examination of undergraduate medical students. *Int J Res Med Sci*. 2017;5(5). DOI:10.18203/2320-6012.ijrms20175453.
- [16] Boopathiraj C, Chellamani K. ANALYSIS OF TEST ITEMS ON DIFFICULTY LEVEL AND DISCRIMINATION INDEX IN THE TEST FOR RESEARCH IN EDUCATION. *Int J Of Social Sci & Interdiscip Res*. 2013;2:189–193.
- [17] Kaur M, Singla S, Mahajan R. Item analysis of in use multiple choice questions in pharmacology. *Int J Appl Basic Med Res*. 2016;6(3):170–173.
- [18] Ali SH, Carr PA, Ruit KG. Validity and reliability of scores obtained on multiple-choice questions: why functioning distractors matter. *J Scholarship Teach Learn*. 2016;16(1):1–14.
- [19] The TK. Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Res Sci Educ*. 2018;48(6):1273–1296.
- [20] O'Malley NT, Cunningham M, Leung F, et al. Early Experience in Implementation of a Learning Assessment Toolkit in the AOTrauma Geriatric Fracture Course. *Geriatr Orthop Surg Rehabil*. 2011;2(5–6):163–171.

- [21] Ghidinelli M, Cunningham M, Uhlmann M, et al. AO Education Platform, U R. Designing and Implementing a Harmonized Evaluation and Assessment System for Educational Events Worldwide. *J Orthop Trauma*. 2021;35(2):S5-S10.
- [22] Fox R, Miner C. Motivation and the facilitation of change, learning and participation in educational programs for health professionals. *J Cont Educ Health Prof*. 1999;19(3):132–141.
- [23] Jayakumar KL. The motivational and evaluative roles of nbme subject examinations. *Acad Med*. 2017;92(10):1363–1364.
- [24] Younas A, Shah I, Lim T, et al. Evaluating an international facial trauma course for surgeons; did we make a difference? submitted.
- [25] Royster E, Morin DE, Molgaard L, et al. Methods used to assess student performance and course outcomes at the national center of excellence in dairy production medicine education for veterinarians. *J Vet Med Educ*. 2020;47(3):263–274.
- [26] Kumar N, Rahman E. Effectiveness of teaching facial anatomy through cadaver dissection on aesthetic physicians' knowledge. *Adv Med Educ Pract*. 2017;8:475–480.
- [27] Licona-Chávez A, Montiel Boehringer P, Velázquez-Liaño L. Quality assessment of a multiple choice test through psychometric properties. *MedEdPublish*. 2020;9(1):91.
- [28] Backhoff E, Larrazolo N, Rosas M. Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA) 2000, Available from: <http://redie.uabc.mx/vol2no1/contenido-backhoff.html>
- [29] Royal K, Dorman D. Comparing item performance on three- versus four-option multiple choice questions in a veterinary toxicology course. *Vet Sci*. 2018;5. DOI:10.3390/vetsci5020055.
- [30] Dawson SD, Miller T, Goddard SF, et al. Impact of outcome-based assessment on student learning and faculty instructional practices. *J Vet Med Educ*. 2013;40(2):128–138.
- [31] Kheyami D, Jaradat A, Al-Shibani T, et al. Item analysis of multiple choice questions at the department of paediatrics, Arabian gulf university, Manama, Bahrain. *Sultan Qaboos Univ Med J*. 2018;18(1):e68–e74.
- [32] Tenzin K, Dorji T, Tenzin T. Construction of multiple choice questions before and after an educational intervention. *JNMA J Nepal Med Assoc*. 2017;56(205):112–116.
- [33] Hingorjo MR, Jaleel F. Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *J Pak Med Assoc*. 2012;62(2):142–147.
- [34] Nazim SM, Riaz Q, Ather MH. Effect of a two-day extensive continuing medical education course on participants' knowledge of clinical and operative urology. *Turk J Urol*. 2018;44(6):484–489.
- [35] Pugh D, De Champlain A, Touchie C. Plus ça change, plus c'est pareil: making a continued case for the use of MCQs in medical education. *Med Teach*. 2019;41(5):569–577.