Development of knowledge tests for multi-disciplinary emergency training: a review and an example

J. L. Sørensen[1], L. Thellesen[2], J. Strandbygaard[1], K. D. Svendsen[3], K. B. Christensen[3], M. Johansen[2], P. Langhoff-Roos[2], K. Ekelund[4], B. Ottesen[1] and C. Van der Vleuten[5]

[1]Juliane Marie Centre for Children, Women and Reproduction, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark
[2]Department of Obstetrics, Juliane Marie Centre for Children, Women and Reproduction, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark
[3]Department of Biostatistics, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[4]Department of Anesthesiology, Juliane Marie Centre for Children, Women and Reproduction, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark
[5]Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands

Correspondence
J. L. Sørensen, Juliane Marie Centre for Children, Women and Reproduction, Rigshospitalet, University of Copenhagen, Blegdamsvej 9, 2100 Copenhagen, Denmark
Email: jette.led.soerensen@regionh.dk

Conflicts of interest
The authors confirm that there are no conflicts of interest.

Funding
Development of the MCQ test was part of a trial[7] based on departmental funding and on non-profit funding from the Danish Regions Development and Research Fund, and a minor amount of non-profit funding from the Laerdal Foundation for Acute Medicine and the Aase and Ejnar Danielsens Foundation. None of the foundations played a role in the design or conduct of the study.

Submitted 15 August 2014; accepted 21 September 2014; submission 5 April 2014.

Citation
Sørensen JL, Thellesen L, Strandbygaard J, Svendsen KD, Christensen KB, Johansen M, Langhoff-Roos P, Ekelund K, Ottesen B, Van der Vleuten C. Development of knowledge tests for multi-disciplinary emergency training: a review and an example. Acta Anaesthesiologica Scandinavica 2014

doi: 10.1111/aas.12428

Background: The literature is sparse on written test development in a post-graduate multi-disciplinary setting. Developing and evaluating knowledge tests for use in multi-disciplinary post-graduate training is challenging. The objective of this study was to describe the process of developing and evaluating a multiple-choice question (MCQ) test for use in a multi-disciplinary training program in obstetric-anesthesia emergencies.

Methods: A multi-disciplinary working committee with 12 members representing six professional healthcare groups and another 28 participants were involved. Recurrent revisions of the MCQ items were undertaken followed by a statistical analysis. The MCQ items were developed stepwise, including decisions on aims and content, followed by testing for face and content validity, construct validity, item–total correlation, and reliability.

Results: To obtain acceptable content validity, 40 out of originally 50 items were included in the final MCQ test. The MCQ test was able to distinguish between levels of competence, and good construct validity was indicated by a significant difference in the mean score between consultants and first-year trainees, as well as between first-year trainees and medical and midwifery students. Evaluation of the item–total correlation analysis in the 40 items set revealed that 11 items needed re-evaluation, four of which addressed content issues in local clinical guidelines. A Cronbach's alpha of 0.83 for reliability was found, which is acceptable.

Conclusion: Content and construct validity and reliability were acceptable. The presented template for the development of this MCQ test could be useful to others when developing knowledge tests and may enhance the overall quality of test development.

Multi-professional training is expected to lead to better patient care and is increasingly recommended as essential in the context of continuous professional development.[1,2] Evidence on its effectiveness is nevertheless limited[3–6] and only few studies on multi-disciplinary and multi-professional education present any kind of outcomes.[3] Guidance on how to develop written tests for a post-graduate multi-disciplinary setting is sparse, but principles for written test development are universal. The post-graduate multi-disciplinary context, however, must be taken into account in the process of test development.

Acta Anaesthesiologica Scandinavica
AN INTERNATIONAL JOURNAL OF ANAESTHESIOLOGY AND INTENSIVE CARE, PAIN AND EMERGENCY MEDICINE

This paper presents the developmental process of a written knowledge test applied in a post-graduate multi-disciplinary training program to determine the training effect on knowledge.[7] Other outcomes studied in the same training program[7] are attitudes toward patient safety, team performance, level of stress, and motivation. The training program was conducted in a randomized controlled trial comparing in situ simulation with off-site simulation, but this article, however, does not describe these outcome data. The knowledge test target group comprised auxiliary nurses, midwives, nurse anesthetists, operating room nurses, trainees, and consultants in obstetrics and anesthesiology.[7]

The optimal approach to assessing clinical competences in a post-graduate setting is disputable.[1] Performance-based tests are currently the most preferred type of test for assessing clinical competences, but they are expensive and less practical than written tests, which is why written knowledge testing may be preferred. Previous research on assessment suggests that knowledge-based written assessments can predict the results of performance-based tests applied to the same test group.[8–10] Several types of written assessment tests exist, including multiple-choice question (MCQ) tests, which have been found to be effective for cognitive assessment.[11–15] Development of a MCQ test requires testing for validity and reliability. The validity of a test is the extent to which the test measures what it intends to measure, and reliability pertains to the accuracy with which a score on a test is determined.[11]

Many anesthesiologists and nurse anesthetists participate in internationally developed courses such as Advanced Life Support,[16] European Paediatric Life Support and Advanced Trauma Life Support, whereas obstetricians and midwifes participate in courses like Advanced Life Support in Obstetrics. In these courses, knowledge-based tests such as MCQ tests are applied as pre- and post-tests. The literature argues that a cautious approach must be taken when describing MCQ tests as comparable with one another because tests can have varying degrees of difficulty.[17] Many articles on these international courses refer to MCQ tests when describing course results, but articles describing how these tests were developed or validated are sparse.[16,18]

The aim of this study was to present a short review of the literature and a template for developing an MCQ test and to describe the process of developing and evaluating it for use in a multi-disciplinary training program in obstetric-anesthesia emergencies.

Material and methods

Setting

The Departments of Obstetrics and Anesthesiology, Juliane Marie Centre for Children, Women and Reproduction, Rigshospitalet, University of Copenhagen, Denmark, with approximately 6000 deliveries per year.

The management teams at the Departments of Obstetrics and Anesthesiology appointed a multidisciplinary working committee to do the overall planning of the training program[7] and the development of the MCQ test. The committee consisted of 12 members: a chair (author JLS), two consultant obstetricians (which include author MJ), two consultant anesthesiologists (which include author KE), two midwives (which include author PLR), two nurse anesthetists, two operating room nurses, and one obstetric nurse.

Study participants

Twelve representatives from the working committee were involved in designing and reviewing the test. An additional 28 participants were involved in pilot testing and validation and comprised four consultant obstetricians and four consultant anesthesiologists from four university hospitals in Denmark; six first-year obstetric trainees from two university hospitals in Denmark; nine medical students in their sixth and final year of medical school at the University of Copenhagen who had completed 2 out of 4 weeks of clinics at the Departments of Obstetrics and 2 weeks at the Department of Anesthesiology; and five midwifery students in their final year of a 3.5-year program at Metropolitan University College, Copenhagen.

Stepwise test development

Anchored in the current test development literature, the test development method was based on the following stepwise protocol (Table 1).[12,13,19,g]

Table 1 Item writing template. Basic principles for writing 'one-best-answer' items for a multiple-choice question (MCQ) test.[12,13,19,g]

Content
Do the items align with the aims and objectives of the course or training program?
Are the items consistent with the content of the course or training program?
Is the MCQ test relevant for clinical work and practice?
Are trick items, trivial items, and opinion-based items avoided?
Is the MCQ test reviewed by relevant healthcare professionals including those for whom the test is aimed for?
MCQ item structure
Do the items follow the basic rules for 'one-best-answer' format'?
Do the items test application of knowledge and integration of information rather than recall of facts?
Are the items formulated with most of the text in the stem and relatively short options?
Stem (consisting of the vignette or case and the lead-in question)
Are the items well formulated and phrased without abbreviations and jargon?
Do the items follow the 'cover the options' rule, i.e. can an answer be formulated based only on the stem?
Is the setting of the items clear (such as emergency room, operation theater, patient ward, etc.)?
Are the lead-in questions structured as a complete sentence ending with a question mark?
Are the lead-in questions structured as a clear task for the participants?
Are negatively phrased lead-in questions avoided and are words like 'not' and 'except' avoided?
The options (one is the correct answer and the others are distracters)
Are the options homogenous and uniform in content and phrasing and within the same category such as diagnoses, treatments, prognoses, etc.?
Develop as many options as possible; however, research shows that three is adequate.
Use typical errors from clinical work or errors from previous tests when designing the distracters.
Are the most common technical faults avoided, such as:
- Are options excluding each other?
- Are distracters made plausible?
- Are similar words in the lead-in question and the correct answer avoided?
- Are absolute terms in options as 'always' and 'never' avoided?
- Are terms for choosing options as 'none of the above' and 'all of the above' avoided?
- Are clues, such as 'clang associations', where options are identical to or resembling words in the stem avoided?
- Are conspicuously correct options or absurd ridiculous options avoided?
Are options technically correct written, such as:
- Options follow grammatically and logically from the lead-in?
- Options are made in a logical or numerical order?
- Options have similar length and are parallel in structure?
- Options are reviewed by relevant healthcare professionals including those for whom the test is aimed for?

A. Aims and objectives: The working committee defined aims and objectives for the multidisciplinary training program,[7] which constituted the content to be included in the MCQ test. The departmental management teams in the Departments of Obstetrics and Anesthesiology approved the aims and objectives.
B. Blueprint and content of the MCQ test: The term blueprint describes subcategories and subclassifications of content in the MCQ test, precisely specifying the proportion of test questions in each category.[20] The working committee determined the blueprint configuration, which was based on the occurrence of various obstetric-anesthesia emergencies and the aims and objectives from step A. The blueprint was divided into four topics and the items within each topic were distributed according to importance: management of postpartum bleeding: approximately 35%; preeclampsia: approximately 35%; Cesarean section: approximately 15%; emergency obstetrics, including resuscitation of the pregnant woman: approximately 15%. The

content was in accordance with national guidelines.[a,b,c] The Danish Society of Obstetrics and Gynecology in cooperation with representatives from the Danish Society of Anesthesiology and Intensive Care Medicine have appointed groups to develop the national guidelines based on a comprehensive review of the international literature and the British guidelines. The national guidelines were then adjusted and made into local guidelines reviewed by the Departments of Obstetrics and Anesthesiology, Juliane Marie Centre, Rigshospitalet.[d,e,f]

C. Items in the MCQ test: The one-best-answer principle was applied in designing each test items included in the MCQ test and is a method acknowledged by the American National Board of Medical Examiners[g,19] and in the test literature.[11–15] Based on the literature, we developed an item-writing template (Table 1),[12,13,19,g] which was used to create items in our MCQ test. Each item consists of a stem, e.g. a clinical case or vignette and a lead-in question. The vignette or case represents a relevant clinical problem and the setting (e.g. an emergency room, operating room, patient ward) for the clinical problem must be clear. The 'cover-your-options' rule helps to control the text in the stem, which means that it must be possible to answer appropriately based solely on the stem (vignette and lead-in question). The stem is followed by a variety of options comprising only one correct answer and various distracters, the latter of which must be homogenous and uniform regarding content and phrasing and within the same category, such as diagnosis, treatment, or prognosis etc. The principle is that the correct answer is the most likely answer.[19] See Table 1 for more details. Previous obstetric knowledge tests were used for inspiration in designing our MCQ test.[21,22] Table 2 presents four examples of MCQ test items. The entire MCQ test has not been published, because future studies and training that involve the use of the MCQ test are planned. A copy of the entire test can be obtained from the corresponding author (JLS) by request from departments or organizations that would like to use the MCQ test. The first author (JLS) wrote the first 50 items, and the third author (JS), who also has experience in test development, was responsible for editing them.[23]

D. Face and content validity I: Validity can be differentiated into face and content validity, representing the acceptance by experts that the test actually tests what it intends to test, whereas construct validity represents a tests ability to discriminate between participants with various levels of competence.[11,13,24] To ensure that the items were in accordance with the aims, objectives, and content of the multidisciplinary training program,[7] a midwife also trained as a nurse (PLR), an obstetrician (MJ), and an anesthesiologist (KE) from the working committee performed and reviewed the MCQ test by providing feedback and discussing each item with JLS before she revised the MCQ test.

E. Face and content validity II: Twelve healthcare professionals from six healthcare professionals group on the working committee took the MCQ test and added written comments on each item to ensure the relevance of the entire MCQ test for the training program participants.[7] JLS then designed a revised version of the 50-item test.

F. Face and content validity III: To test the accuracy and generalizability of the content and the clinical relevance of the test inside and outside local settings, four consultant obstetricians and four consultant anesthesiologists

a http://www.dsog.dk/files/postpartum_bloedning.pdf (19.10.2014)
b http://www.dsog.dk/sandbjerg/120403%20PIH%202012%20final.pdf (19.10.2014)
c http://www.dsog.dk/sandbjerg/090405%20Guideline%20Akut%20sectio%20-%20klassifikation%20%20Sandbjerg%202009.pdf (19.10.2014)
d http://vip.regionh.dk/VIP/Admin/GUI.nsf/Desktop.html?open&openlink=http://vip.regionh.dk/VIP/Slutbruger/Portal.nsf/Main.html?open&unid=X2A39F5D1F8F5992CC125795000421E19&dbpath=/VIP/Redaktoer/1301XJ.nsf/&windowwidth=1100&windowheight=600&windowtitle=S%F8g (19.10.2014)
e http://vip.regionh.dk/VIP/Admin/GUI.nsf/Desktop.html?open&openlink=http://vip.regionh.dk/VIP/Slutbruger/Portal.nsf/Main.html?open&unid=X5143CF11C0E062A9C125791500780413&dbpath=/VIP/Redaktoer/1301XJ.nsf/&windowwidth=1100&windowheight=600&windowtitle=S%F8g (19.10.2014)
f http://vip.regionh.dk/VIP/Admin/GUI.nsf/Desktop.html?open&openlink=http://vip.regionh.dk/VIP/Slutbruger/Portal.nsf/Main.html?open&unid=X324B22542088F6A4C125791500786CBC&dbpath=/VIP/Redaktoer/1301XJ.nsf/&windowwidth=1100&windowheight=600&windowtitle=S%F8g (19.10.2014)
g http://www.nbme.org/pdf/itemwriting_2003/2003iwgwhole.pdf (20.6.2014)

Table 2 Four examples of multiple-choice question items. These items were constructed in accordance with the template in Table 1.

Blueprint: Post-partum bleeding (N16)	Stem	Vignette or case	*A previous healthy woman delivered, assisted by a vacuum extractor 2 h ago. She has acceptable vaginal bleeding. She complains of severe pain in the lower abdomen and vagina. She has normal pulse and blood pressure. A midwife asks if she can give morphine injection for pain relief.*
		Lead-in question	What is the most likely explanation for the severe pain?
		Options	a) Rupture of uterus b) Retained part of placenta in the uterus c) Thrombosis in the pelvic veins d) Vaginal hematoma
Blueprint: Preeclampsia (N37)	Stem	Vignette or case	*A pregnant woman in labor ward with severe preeclampsia receives infusion with magnesium sulfate. She is transferred to operation theater for cesarean section. When she arrives in theater, she feels very uncomfortable. Staff in theater discuss whether it is due to side effects of the magnesium infusion*
		Lead-in question	What is the most common side effect of magnesium sulfate that pregnant women complain about?
		Options	a) Tingling in lips and tongue b) Cutaneous flushing c) Ringing in the ears d) Metallic taste
Blueprint: Emergency cesarean section (N1)	Stem	Vignette or case	*A pregnant woman arrives to the operation room for an emergency cesarean section, and needs to be placed on the operating table.*
		Lead-in question	Which is the most likely position for the pregnant women to do during cesarean section?
		Options	a) Trendelenburg position b) Flat on the back c) 15–20 degree in left lateral position d) 15–20 degree in right lateral position
Blueprint: Resuscitation in obstetrics (N40)	Stem	Vignette or case	*The ECG shows ventricular fibrillation, and there is indication for defibrillation.*
		Lead-in question	How will you manage defibrillation in a pregnant patient?
		Options	a) As in non-pregnant patients b) Shall not be used c) Is a risk for the fetus d) Shall be used with increased current flow

from four university hospitals were asked to: (1) take the test; (2) rate the relevance of each item on a 3-point scale [(1): not relevant; (2): relevant; (3): very relevant]; (3) provide written feedback on each item; and (4) suggest new items. Based on this feedback, JLS created a new version of the MCQ test.

G. Construct validity:[11,13,24] This was tested by comparing the test results from groups with expected differentiated level of knowledge and clinical competences comprising: (1) consultant obstetricians and anesthesiologists (from step F); (2) first-year obstetric trainees; (3) medical and midwifery students.

H. Test-time registration: The amount of time it took to take the entire MCQ test was only registered for the first-year obstetric trainees, the medical students, and the midwifery students. In addition to taking the test, the other participants spent time providing feedback, which means that recording how long it took them to take the test was not useful.

Ethics

The study did not involve patients, which means no approval was required under Danish regulations. Responses from medical students and midwifery students were non-traceable data, but responses from the other participants were not anonymized as they were both written and oral. Participants were informed that during the analyses and reporting of the data, all information would be treated as non-traceable.

Data analyses

Data were processed using Microsoft Office Excel 2007 [Microsoft (2007), Microsoft Excel (computer software), Redmond, WA, USA], SAS 9.2 (SAS Institute Inc, Cary, NC, USA), and R version 3.0.1 (R Core Team, Vienna, Austria). Appendix S1 provides supplementary material on the statistical analysis undertaken.

An initial aspect of the analysis process involved a qualitative assessment of the content of the 50 items that entailed revising and excluding items based on feedback, ultimately providing information on face and content validity.[11,13,24]

The second part involved statistical analyses of the remaining 40 items. A Wilcoxon rank sum test was used to analyze the construct validity by comparing the test scores of the three groups of participants (consultants, trainee, and students).[13] This was done to examine the MCQ test's ability to discriminate between participants with various levels of competence.

The statistical validation of the MCQ test investigated the correlation between the MCQ items. To address the function of the individual items, we computed item–total correlations. For each item within the blueprint, Spearman's rank correlation was used to evaluate the total score, while Loevinger H coefficients using non-arametric item response theory, also known as Mokken scale analysis, were used to further examine item quality.[25–28] Mokken scale analysis is particularly convenient if the number of items in a scale is low, as is the case for our MCQ test.[27,28] The latter analysis was done iteratively by removing items until the requirement of values larger than 0.30 was met.[28] A correlation level above 0.30 is generally considered acceptable.

Cronbach's alpha was used to examine the reliability of the MCQ test[29] and was applied by iteratively removing items leading to an acceptable value. A Cronbach's alpha of 0.70 or higher is acceptable in a test that does not have any consequences for the participants.[29] If the test involves certification or is a genuine examination, the reliability requirements are higher, which means that Cronbach's alpha should exceed 0.90.[12,29]

We computed the proportion of correct answers for each of the 40 items to evaluate floor and ceiling effects: The floor effect is when a test is too difficult and only a minor number of participants can answer the test questions correctly, whereas the ceiling effect is when a test has a maximum score that can be attained too easily without an outstanding performance.

Results

The study period for developing and testing MCQ items was December 2012 to April 2013.

Fifty items were developed in accordance with the template in Table 1. According to feedback from the eight consultants, five items were excluded, three of them because three to six of the consultants answered them incorrectly. Of the two remaining items, five out of eight consultants considered one to be irrelevant and the other had more than one correct option, thus reducing the number of MCQ test items from 50 to 45.

Only one of the consultants suggested new ideas for obstetric resuscitation items and one sentence was added to two of the existing items. No new items were constructed. When two medical students pointed out that the correct answer for one item was available in the stem of another item, it was revised.

On average, medical students, midwifery students, and first-year obstetric trainees spent 34 min (28–45 min) on the 45-item test. We wanted to minimize testing time to make the test applicable for a multi-disciplinary training program,[7] which is why an additional five items were excluded, reducing the final total to 40 items. The eight specialists considered the last five items left out as relevant but not very relevant.

After this qualitative analysis, the MCQ test consisted of 40 items distributed in accordance with the blueprint: management of post-partum bleeding: 14 items (35%); preeclampsia: 14 items (35%); Cesarean section: 6 items (15%); and emergency obstetrics, including resuscitation: 6 items (15%).

Table 3 presents information on construct validity. The mean test scores for consultant obstetricians and anesthesiologist, first-year obstetric trainees, medical students, and midwifery students are presented and a significant difference in

Table 3 Criterion-related construct validity: Mean test scores in the 40-item multiple-choice question test for consultant obstetricians and anesthesiologists, obstetric first-year trainee, medical students, and midwifery students. In comparison among either consultants or first-year trainees and medical/midwifery students, the two latter groups were merged.

	Consultant obstetricians and consultant anesthesiologists (n = 8)	Obstetric first-year trainee (n = 6)	Medical students (n = 9)	Midwifery students (n = 5)
Mean score [standard deviation (SD)]	35.3 (SD = 2.9)	28.8 (SD = 2.7)	24.8 (SD = 3.5)	20.4 (SD = 4.5)
Mean test score in percentage (range)	89% (78–98%)	72% (60–78%)	62% (43–70%)	50% (38–63%)
	P = 0.0072*			
	P = 0.001*			
	P = 0.02*			

*Wilcoxon rank sum test.

the mean score between the groups was detected, indicating acceptable construct validity.

Table 4 shows the items within each blueprint, the proportion of correct answers, the item–total Spearman's rank correlation, and the H coefficients. Values for the remaining items after omission of those with values below 0.3 are also shown. The item–total correlations indicated that 7 of the 40 items were needed to be re-evaluated. The H coefficients also indicated a misfit of the same seven items and an additional four. Thus, 11 of the 40 items were needed to be re-evaluated. Hence, the criteria based on Spearman's rank correlation and the Mokken scale analysis coincided. The content in 4 of the 11 problematic items was based on local guidelines on management of post-partum bleeding (N12, 26, 29) and preeclampsia (N31). For example, an MCQ item on local guidelines specified when to call for help when managing post-partum bleeding and another item on expected time to wait for blood transfusion in emergency situations.

The computed proportion of correct answers for each of the 40 items revealed no floor effect and a minor and acceptable ceiling effect.

When analyzing all 40 items, the MCQ test revealed a Cronbach's alpha of 0.83. The Cronbach's alpha in each of the four blueprint topics resulted in lower values. In post-partum bleeding, Cronbach's alpha could be improved from 0.45 to 0.65 when removing 4 (N8, N12, N14, and N26) of the 14 items. The Cronbach's alpha for the 14 items in preeclampsia was 0.75, and there was no notable increase when items were removed. Cronbach's alpha for the six items in cesarean section and obstetric resuscitation were 0.60 and 0.54, respectively.

Discussion

The final version of the MCQ test had acceptable reliability, content, and construct validity.

The initial part of the development process consisted of qualitative analyses involving relevant healthcare professionals and representatives from anesthesiology and obstetrics. Feedback from consultant and trainee obstetricians and anesthesiologists, midwives, nurse anesthetists, operating room nurses, medical students, and midwifery students ensured that the content and diction were understandable for a broad group of healthcare professionals.

We followed the method described in the literature for MCQ development.[11–15,19,30] We concluded that the face and content validity and the construct validity of the MCQ test were acceptable, as the test was able to discriminate between groups of participants expected to perform differently due to variations in knowledge and clinical skills.

The statistical analyses provided information on the quality of each item in the MCQ test and

Table 4 Proportion of correct answers, item–total correlation, and H coefficient in the multiple-choice question test.

Blueprint topic	MCQ item	Proportion of correct answers (%)	Item–total correlation (Spearman rank correlation)	H coefficient	MCQ items that need to be re-evaluated
Post-partum bleeding	N8	86			To be re-evaluated
	N11	71	0.33	0.30	
	N12	75			To be re-evaluated (locally relevant)
	N13	46	0.74	0.57	
	N14	57			To be re-evaluated
	N15	25	0.44	0.40	
	N16	93	0.32	0.36	
	N18	79	0.43	0.32	
	N25	50	0.6	0.36	
	N26	57			To be re-evaluated (locally relevant)
	N27	46	0.56	0.35	
	N29	64			To be re-evaluated (locally relevant)
	N30	75	0.51	0.39	
	N32	75			To be re-evaluated
Preeclampsia and eclampsia	N5	71	0.54	0.34	
	N6	79	0.51	0.30	
	N7	86	0.42	0.37	
	N9	57	0.51	0.27	
	N10	54	0.75	0.46	
	N23	68	0.42	0.27	
	N24	57	0.51	0.34	
	N28	79	0.47	0.24	
	N31	75	0.33		To be re-evaluated (locally relevant)
	N33	43	0.31		To be re-evaluated
	N34	64	0.43		To be re-evaluated
	N36	93	0.4	0.56	
	N37	54	0.63	0.39	
	N38	75	0.51	0.33	
Emergency cesarean section	N1	71	0.65	0.33	
	N17	61	0.71	0.42	
	N19	82	0.54	0.40	
	N20	71	0.58	0.27	
	N21	64	0.59	0.28	
	N35	89			To be re-evaluated
Resuscitation in obstetrics	N2	79	0.47		To be re-evaluated
	N3	96	0.33	0.69	
	N4	79	0.63	0.41	
	N22	64	0.67	0.44	
	N39	89	0.44	0.23	
	N40	86	0.57	0.34	

indicated which items needed further discussion and perhaps re-evaluation. The Mokken scale analysis identified 11 misfit items, seven of which were also identified by the Spearman's rank correlation. This means that an expected correct answer to one of these 11 problematic items did not necessarily correlate with the probability of answering other items on the MCQ test correctly. The content in 4 of the 11 problematic items was based on local guidelines, and this might explain why these items turned out to be problematic in the statistical analysis when testing participants who were from other hospitals. These four items can be considered relevant for our MCQ test when applied to the local setting. If the test is to be used at other hospitals, these four items will need to be revised or excluded. The remaining seven items need to be either excluded or reana-

lyzed in a larger population that fully matches the participants for whom the MCQ test targets.[7] Information on the floor and ceiling effect revealed an acceptable balance between easy and difficult items.

Cronbach's alpha provides information on the reliability of a test. If a test is reliable, it indicates that retest results will be similar. The present test generated a high Cronbach's alpha (0.83) when using all 40 items. When analyzing data from each blueprint separately, the Cronbach's alpha values were lower, which could be due to the small number of items. We considered the Cronbach's alpha values of the present test to be acceptable.

To measure the effect of a training program, ideally we would measure direct clinical outcomes such as neonatal and maternal morbidity and mortality. This, however, is normally not feasible as a high number of deliveries are required to measure patient-relevant outcomes in the wide clinical spectrum covered by our MCQ test.[31] Moreover, some educational studies indicate that performance in a written knowledge test can relate to a clinical performance-based test.[8–10,32]

Studies show that participants that were tested on a specific topic retain knowledge better than if they were not tested: the so-called testing effect.[33,34] Relevant testing has even been shown to lead to more knowledge gain than teaching without testing.[34] Well-designed written tests combined with other assessment tools may therefore be used as an integrated learning strategy in a training program. In specialist training in anesthesiology in Denmark, several knowledge tests that are used as formative testing, i.e. tests given for feedback purposes, such as the MCQ test, are currently being implemented. Presently, whether to integrate knowledge testing as a part of several post-graduate training programs in gynecology and obstetrics, e.g. basic laparoscopy training[23] in Denmark and Norway and cardiotocography training programs in Sweden and Denmark, is being discussed.[35] In these post-graduate training programs, a specific test is integrated into the training program, and testing is not isolated from training, e.g. learning and testing are applied as part of a program integrated in the clinical context.

Test development is often considered a simple task; however, designing a valid and reliable test is a complex process. In the process of test development for a multi-disciplinary setting, involving representatives from all the relevant healthcare professional groups and from all the relevant medical specialties is essential. Enhancing the generalizability of our MCQ test to other institutions required incorporating feedback on the content of the test from consultants from other hospitals. Substantial insight in test development literature from textbooks and original scientific work is a necessary pre-requisite when embarking on test development.[11–15,19,20,29,30] Developing valid and reliable items requires competences within both test development and test statistics combined with in-depth knowledge on the content of the test. The template presented for item writing and the examples of MCQ items may prove useful for others who would like to develop a test and can potentially enhance the quality of tests by improving validity, correlation, and reliability.

Acknowledgements

We would like to thank everyone who participated in taking and providing feedback on the MCQ test: consultant anesthesiologists: Helle Thy Østergaard and Lone Fuhrmann, Department of Anesthesiology, Herlev Hospital; Mette Gøttge Madsen, Department of Anesthesiology, North Zealand Hospital; Søren Heltbo, Department of Anesthesiology, Odense University Hospital. Consultant obstetricians: Nina Colov Palmgren, Department of Obstetrics, Rigshospitalet; Lone Krebs, Department of Obstetrics and Gynecology, Holbæk Hospital; Birgit Bødker, Department of Obstetrics and Gynecology, North Zealand Hospital; Morten Beck Sørensen, Department of Obstetrics and Gynecology, Odense University Hospital. We would also like to thank these first-year obstetrics trainees: Flemming Bjerrum, Tanja Roien Jacobsen, Astrid Kolte, Mette Petri, and Camilla Wullf. We are also grateful to the Rigshospital working committee: consultant anesthesiologist Charlotte Krebs Albrechtsen, consultant obstetrician Berit Woetman Pedersen, midwife Kristine Sylvan Andersen, nurse anesthetists Charlotte Glob Frandsen and Marianne Sand Flindt, operating room nurses Pernille Baagøe Schou and Birgitte Otzen, and obstetric nurse Vibeke Ladefoged. We also extend our

thanks to the medical and midwifery students who took the MCQ test and provided feedback.

Author contributions

JLS created the idea for this article with support from CVdV and BO. JLS was responsible for acquiring funding in cooperation with BO. JLS, LT, JS, MJ, KE, and PL-R made substantial contributions to the practical issues involved in developing the MCQ test. KDS and KBC jointly performed the statistical analysis with JLS. JLS wrote the draft manuscript. All of the authors provided critical review of this paper and approved the final manuscript.

References

1. Norman GR, Shannon SI, Marrin ML. The need for needs assessment in continuing medical education. BMJ 2004; 328: 999–1001.
2. Choudhry NK, Fletcher RH, Soumerai SB. Systematic review: the relationship between clinical experience and quality of health care. Ann Intern Med 2005; 142: 260–73.
3. Reeves S, Zwarenstein M, Goldman J, Barr H, Freeth D, Koppel I, Hammick M. The effectiveness of interprofessional education: key findings from a new systematic review. J Interprof Care 2010; 24: 230–41.
4. Zwarenstein M, Goldman J, Reeves S. Interprofessional collaboration: effects of practice-based interventions on professional practice and healthcare outcomes. Cochrane Database Syst Rev 2009; (3): CD000072.
5. Headrick LA, Wilcock PM, Batalden PB. Interprofessional working and continuing medical education. BMJ 1998; 316: 771–4.
6. Nancarrow SA, Booth A, Ariss S, Smith T, Enderby P, Roots A. Ten principles of good interdisciplinary team work. Hum Resour Health 2013; 11: 19.
7. Sorensen JL, Van der Vleuten C, Lindschou J, Gluud C, Ostergaard D, Leblanc V, Johansen M, Ekelund K, Albrechtsen CK, Pedersen BW, Kjaergaard H, Weikop P, Ottesen B. 'In situ simulation' versus 'off site simulation' in obstetric emergencies and their effect on knowledge, safety attitudes, team performance, stress, and motivation: study protocol for a randomized controlled trial. Trials 2013; 14: 220.
8. Remmen R, Scherpbier A, Denekens J, Derese A, Hermann I, Hoogenboom R, Van der Vleuten C, Van Royen P, Bossaert L. Correlation of a written test of skills and a performance based test: a study in two traditional medical schools. Med Teach 2001; 23: 29–32.
9. Kramer AW, Jansen JJ, Zuithoff P, Dusman H, Tan LH, Grol RP, Van der Vleuten C. Predictive validity of a written knowledge test of skills for an OSCE in postgraduate training for general practice. Med Educ 2002; 36: 812–9.
10. Van der Vleuten C, Van Luyk SJ, Beckers HJ. A written test as an alternative to performance testing. Med Educ 1989; 23: 97–107.
11. Schuwirth LW, van der Vleuten C. ABC of learning and teaching in medicine: written assessment. BMJ 2003; 326: 643–5.
12. Downing SM. Written tests. Constructed-response and selected-response formats. In: Downing SM, Yudkowsky R eds. Assessment in health professions education, 1st edn. New York, NY: Routledge, 2009: 149–84.
13. Downing SM. Twelve steps for effective test development. In: Downing SM, Haladyna TM eds. Handbook of test development. Mahwah: Lawrence Erlbaum Associates, Inc. Publishers, 2011: 3–25.
14. Schuwirth LW, van der Vleuten C. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ 2004; 38: 974–9.
15. Schuwirth LW, van der Vleuten C. Written assessments. In: Dent JA, Harden RM eds. A practical guide for medical teachers. Edinburgh: Churchill Livingstone Elsevier, 2009: 325–32.
16. Yang CW, Yen ZS, McGowan JE, Chen HC, Chiang WC, Mancini ME, Soar J, Lai MS, Ma MH. A systematic review of retention of adult advanced life support knowledge and skills in healthcare providers. Resuscitation 2012; 83: 1055–60.
17. Ringsted C, Lippert F, Hesselfeldt R, Rasmussen MB, Mogensen SS, Frost T, Jensen ML, Jensen MK, Van der Vleuten C. Assessment of advanced life support competence when combining different test methods-reliability and validity. Resuscitation 2007; 75: 153–60.
18. Lorello GR, Cook DA, Johnson RL, Brydges R. Simulation-based training in anaesthesiology: a systematic review and meta-analysis. Br J Anaesth 2014; 112: 231–45.
19. Case SM, Swanson DB eds. Constructing written test questions for the basic and clinical sciences, 3rd edn. Philadelphia, PA: National Board of Medical Examiners, 2002.

20. Downing SM, Haladyna TM. Validity and its threats, In: Downing SM, Yudkowsky R eds. Assessment in health professions education, 1st edn. New York, NY: Routledge, 2009: 21–53.
21. Sorensen JL, Lokkegaard E, Johansen M, Ringsted C, Kreiner S, McAleer S. The implementation and evaluation of a mandatory multi-professional obstetric skills training program. Acta Obstet Gynecol Scand 2009; 88: 1107–17.
22. Crofts JF, Ellis D, Draycott TJ, Winter C, Hunt LP, Akande VA. Change in knowledge of midwives and obstetricians following obstetric emergency training: a randomised controlled trial of local hospital, simulation centre and teamwork training. BJOG 2007; 114: 1534–41.
23. Strandbygaard J, Maagaard M, Larsen CR, Schouenborg L, Ottosen C, Ringsted C, Grantcharov T, Ottesen B, Sorensen JL. Development and validation of a theoretical test in basic laparoscopy. Surg Endosc 2013; 27: 1353–9.
24. Wisborg T, Ringsted C. Tools for the assessment of practical skills and performance. Acta Anaesthesiol Scand 2011; 55: 633–4.
25. Streiner LD, Norman GR. Item response theory, In: Streiner L, Norman GR eds. Health measurement scales, 4th edn. Oxford: Oxford University Press, 2008: 299–330.
26. Streiner LD, Norman GR. Selecting the items. In: Streiner L, Norman GR eds. Health measurement scales, 4th edn. Oxford: Oxford University Press, 2008: 77–102.
27. Sijtsma K, Molenaar IW. Introduction to nonparametric item response theory. Thousand Oaks, CA: Sage Publication, 2002.
28. Mokken RJ, Lewis JR. A nonparametric approach to the analysis of dichotomous item responses. Appl Psychol Meas 1982; 6: 417–30.
29. Axelson RD, Kreiter CD. Reliability, In: Downing SM, Yudkowsky R eds. Assessment in health professions education, 1st edn. New York, NY: Routledge, 2009: 57–73.
30. Naeem N, van der Vleuten C, Alfaris EA. Faculty development on item writing substantially improves item quality. Adv Health Sci Educ Theory Pract 2012; 17: 369–76.
31. Mongelli M, Chung TK, Chang AM. Obstetric intervention and benefit in conditions of very low prevalence. Br J Obstet Gynaecol 1997; 104: 771–4.
32. Ram P, van der Vleuten C, Rethans JJ, Schouten B, Hobma S, Grol R. Assessment in general practice: the predictive value of written-knowledge tests and a multiple-station examination for actual medical performance in daily practice. Med Educ 1999; 33: 197–203.
33. Larsen DP, Butler AC, Roediger HL III. Test-enhanced learning in medical education. Med Educ 2008; 42: 959–66.
34. Kromann C, Koefoed M, Jensen M, Ringsted C. Test af viden og færdigheder øger indlæring [Test of knowledge and skills enhances learning]. Ugeskr Laeger 2012; 174: 716–9.
35. Millde-Luthander C, Hogberg U, Nystrom ME, Pettersson H, Wiklund I, Grunewald C. The impact of a computer assisted learning programme on the ability to interpret cardiotochography: a before and after study. Sex Reprod Healthc 2012; 3: 37–41.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Appendix S1. This appendix outlines the statistical methodology used in the validation of the present MCQ test.