Graduate thesis

NTNU
Norwegian University of Science and Technology
Faculty of Medicine and Health Sciences

Inga Skogvold Rygg

The effects of external peer-review on multiple choice item quality in medical education

Graduate thesis in Programme of Professional Study, Medicine
Supervisor: Tobias S. Slørdahl
Co-supervisor: Susanne Skjervold Smeby Martinsen, Rune Standal, Børge Lillebo

June 2021

NTNU
Norwegian University of Science and Technology

Inga Skogvold Rygg

The effects of external peer-review on multiple choice item quality in medical education

Graduate thesis in Programme of Professional Study, Medicine
Supervisor: Tobias S. Slørdahl
Co-supervisor: Susanne Skjervold Smeby Martinsen, Rune Standal, Børge Lillebo
June 2021

Norwegian University of Science and Technology
Faculty of Medicine and Health Sciences

NTNU
Norwegian University of
Science and Technology

Abstract

Quality assurance measures play an important role in the quality of assessment programs, and several studies have shown positive effects of in-house peer-review of multiple choice questions (MCQs). However, we lack knowledge about the effects of external peer-review. In this study we explore how external, double-blinded peer-review affect the quality of summative examinations in a medical program during a four-year period. We analyzed the number of items changed following peer-review, the psychometric measures and student appeals. Results showed that peer-reviewers are able to identify problematic items, and that item writers and peer-reviewers seem to disagree most on the relevance of basic science items and items that require recall of knowledge, as opposed to higher cognitive reasoning (K1-items). Difficult items and K1-items were less likely to be approved in peer-review, and items that were unchanged following review had more functioning distractors. We found no improvement in psychometric measures after implementation of peer-review; in the following years we found easier and less discriminating items with less functioning distractors. Student appeal numbers indicated that students and peer-reviewers agree when identifying problematic items.

Introduction

Assessment plays an important role in learning. For students, what faculty assess heavily influences what they learn in preparation for the assessments. For medical schools, it is essential to certify that their students have the necessary skills, knowledge, and competency to practice medicine safely [1]. Educational institutions rely on assessment programs to ensure that their students acquire the necessary competencies required by a health care system in need of flexible and adaptive [2].

Written assessment makes up the most significant part of assessment programs in most medical schools, and multiple choice questions (MCQs) is the format that students most often encounter. MCQs have many advantages. An important strength is their ability to test both an extensive breadth of knowledge and higher-level cognitive reasoning [3, 4]. In addition to this, MCQs are easy to administer to a large group of students in a relatively short period of time and are easily

computer-scored. However, a significant disadvantage is that many MCQs given in in-house examinations are flawed [5-7] introducing systematic error into the interpretation of their results. Violation of accepted item-writing guidelines is called item writing flaws (IWFs)[8].

To prevent IWFs, medical schools implement quality assurance procedures such as educating faculty members on item writing, student feedback, psychometric evaluation, and internal review committees. Despite these measures, in-house examinations often fall short to national examinations in terms of quality assessment, in part due to economic costs and resources [9].

In 2015, Norwegian University of Science and Technology (NTNU) introduced external, double-blinded peer-review as an added measure to the assessment program to reduce the number of flawed items discovered post-test and improve the quality of written assessments. Several studies have documented the effect of in-house peer review of MCQs [5, 10, 11]. Smeby et al. [12] found external peer-review of MCQs to be promising for in-house examinations. One out of five items reviewed were not approved, the main reasons being relevance and accuracy of item content, and technical item writing flaws [12]. This study is a follow-up study looking into the effect of external peer-review of MCQ items on psychometric effects and the long-term effect on item quality after four years with this measure. Our main aim was to look at how psychometric measures such as difficulty, discrimination index, point biserial item discrimination (RPB), cognitive level (K1/K2) and distractor statistics differed depending on their review status, as well as how they changed over the years. We also wanted to look at the number of student appeals to see how they correlated with peer-review decisions.

Materials and methods

Assessment program and quality assurance

NTNUs medical program is a six-year program, and uses an assessment program consisting of yearly or semesterly summative examinations, depending on the study year. The program is divided into three stadiums - stadium 1 (year 1 and 2), stadium 2 (year 3 and 4) and stadium 3 (year 5 and 6). The examinations consist of an oral and clinical skills examination, most study

years in the form of OSCE (Objective Structured Clinical Examination), and a written examination consisting of 100-120 MCQs and 2-4 modified essay questions (MEQs). The examination is pass/fail, with the absolute limit for a passing grade of the written examinations set at 65% correct answers [13].

There is an extensive quality assurance procedure (Figure 1) in place in the medical program at the NTNU. MCQs are written by academic and clinical staff, are single best answer, consist of three to five options, and marked without penalty for incorrect answers. The MCQs are reviewed in a process similar to the Maastricht model [14], with departments using a blueprint to write items for examinations. These are then entered into a web-based item bank, and a multidisciplinary review committee and two senior students review the items before they are given on examinations [12].

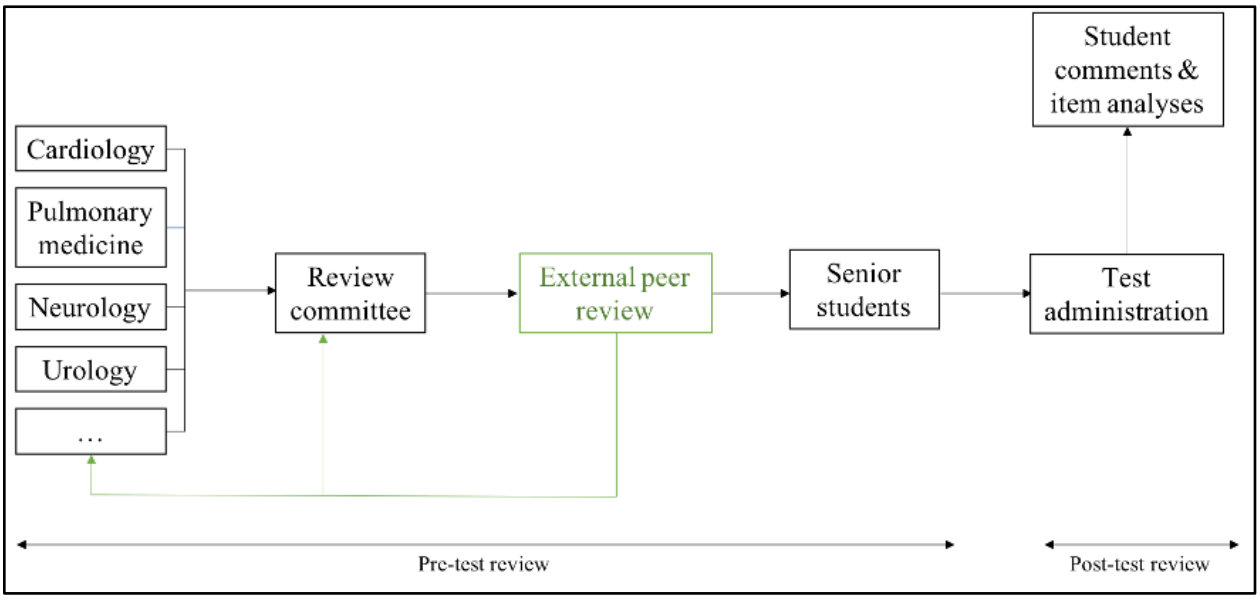
In 2016, external, double-blinded peer review was implemented, with items being submitted to clinicians who meet a set of criteria as described in Smeby et al. [12]. Reviewers received written information on the MCQ format and review process, along with item writing guidelines. Items were randomly selected to be peer-reviewed, with each reviewer receiving ten items within their field of expertise to review. Items in subjects which lack external reviewers with corresponding field of expertise, were pooled and divided between reviewers.

Reviewers were asked to give feedback on whether the item should pass, fail or whether they suggested changes to the item. In addition, they gave feedback on the relevance of the item, whether it required higher cognitive reasoning, if the chosen alternative was clearly the best, if there was an explanation for the correct answer, and the quality of this explanation. They were also asked what percentage of students they expected to answer correctly, and how many of the borderline students they expected would answer the question correctly, both given in 5-percent intervals.

The review and feedback from external reviewers were then shared with the item writer, who could choose whether to change the item and comment on the feedback given. The final decision

on including the item in the examination is taken by the multidisciplinary examination committee based on input from the external reviewer and the item writer.

Figure 1. Item review process at the medical program, NTNU.



From Smeby, S. S. et al. (2019). Improving assessment quality in professional higher education: Could external peer review of items be the answer?. Cogent Medicine, 6(1), 1659746. (https://doi.org/10.1080/2331205X.2019.1659746). Reproduced with the author's permission.

At NTNU, the medical students are allowed to send formal feedback on items within one week after each examination - this is called the student appeal. It must be sent as a collective document from all the students who have taken the examination. For an item appeal to be included, at least 15% of the students have to support the appeal. This is not a formal complaint, but a measure to improve the quality of examinations and items. The examination committee contacts the item writer for feedback on the appeal before making the final decision: to leave the item as is, delete the item from the examination, or accept more than one alternative as a correct answer.

Sample description

Summative MCQs administered to medical students at NTNU for the academic years of 2016-2019 were retrieved from our MCQ database for analysis, and all peer-reviewed examinations

were included. Examinations from 2020 were not included in the data due to the Covid-19 pandemic, making all examinations home examinations where students could use supporting materials. One examination set from 2016 was also not included, as it was deemed an outlier - it was significantly more difficult than the rest of the examinations. 25 individual items in one examination were also removed due to technical error revealing the correct answer to the students during the examination.

Psychometric measurements and statistical analysis

The following data were collected and analyzed in IBM SPSS Statistics Version 27: Review decision, whether the item was changed or deleted by the item writer/examination committee prior to the exam, student appeals, whether the item was changed or deleted by the examination committee after the student comments, and psychometric data of the items (Difficulty, discrimination index, mean corrected point biserial item discrimination (CRPB), cognitive level (K1/K2), % of respondents that each distractor elicited, number of distractors and distractor frequency). A cognitive level of K1 indicates an item that requires recall of facts, whereas a K2-item requires reasoning in order to identify the correct answer.

Descriptive statistics were found using frequencies and custom tables, and then analyzed. When comparing a categorical, independent variable with a continuous, dependent variable we used Kruskal-Wallis Test, and when comparing two categorical variables we used Chi-Square Test for Independence. When a table contained one or more cell with value < 5, we used Fisher's Exact Test.

Ethics

The data material was extracted anonymously from the Faculty's MCQ-database. This meant the research was done on data material not containing identifying information about the students or the peer-reviewers of the items. Therefore no approval from the Norwegian Center of Research Data was necessary to conduct this study.

Results

Number of items changed following peer review

A total of 2930 items were analyzed. Of these, 2017 items (68.8%) were peer-reviewed, while 913 items (32.2%) were not peer-reviewed. Of the 2017 items that were reviewed, 1740 items (86.3%) were judged as approved ("accepted for use"). An additional 244 items (12.1%) were judged as needing revision ("requested revised"), and 33 items (1.1%) did not pass the review ("rejected"). Of the items not considered approved ("requested revised" or "rejected"), 132 (47.7%) were then changed by the item author or the examination committee, while 145 (52.3%) were left unchanged.

Table 1. Peer review decisions and changes made to MCQs

	Accepted for use	**Requested revised**	**Rejected**	**Total**
Changed, ***n***	n/a	125	7	132
Unchanged, ***n***	n/a	119	26	145
Total, ***n***	1740	244	33	**2017**

Peer review and psychometric measures of items

We were interested in examining any differences in items depending on their peer review status, to look for potential patterns and biases in the peer-review process.

Items that were judged as approved in peer-review were easier than items that were not considered approved and left unchanged by the item writer ($p = .002$). When dividing item difficulty into subgroups of easy, middle range and difficult, we find that fewer difficult items were considered approved, and there were fewer easy items among items that were unchanged following review ($p = .009$). There were no significant differences in discriminating power or CRPB across the groups (Table 2).

There was a significant difference ($p < .001$) in the distribution of cognitive level, with more K1-items and fewer K2-items in the unchanged category. Correspondingly, items that were judged as approved had fewer K1-items and more K2-items (table 2). Unchanged items had more

functioning distractors and fewer non-functioning distractors when compared to the other groups, while the opposite was true for items that were accepted for use (p = .004).

Table 2. Item and distractor analysis by review decision and changes made for MCQs used in ordinary examinations

	Accepted for use	**Requested revised or rejected**	
		Changed	**Unchanged**
Number of items, *n*	1740	132	145
Mean difficulty (SD)	0.789 (0.202)	0.743 (0.240)	0.711 (0.252)*
Easy items (p ≥ 0.80), *n* (%)	1049 (60.3)	72 (54.5)	73 (50.3)*
Middle range items (0.20 < p < 0.80), *n* (%)	669 (38.4)	56 (42.4)	66 (45.5)
Difficult items (p ≤ 0.20), *n* (%)	22 (1.3)	4 (3.0)	6 (4.1)*
Mean discrimination index (SD)	0.178 (0.158)	0.163 (0.147)	0.184 (0.171)
Items with negative discrimination indices, *n* (%)	87 (5.0)	7 (5.3)	13 (9.0)
Items with zero discrimination indices, *n* (%)	177 (10.2)	13 (9.8)	13 (9.0)
Items with discrimination indices 0.1- 0.15, *n* (%)	615 (35.3)	52 (39.4)	43 (29.7)
Total number of non-discriminating items, *n (%)*	879 (50.5)	72 (54.5)	69 (47.6)
Total number of discriminating items (> 0.15), *n* (%)	861 (49.5)	60 (45.5)	76 (52.4)
Mean point biserial item discrimination, CRPB (SD)	0.157 (0.142)	0.134 (0.135)	0.154 (0.150)
Cognitive level			
K1, *n* (%)	547 (31.4)	47 (35.6)	69 (47.6)*
K2, *n* (%)	1193 (68.6)	85 (64.4)	76 (52.4)*
Number of distractors per item			
2, *n* (%)	44 (2.5)	7 (5.3)	4 (2.8)
3, *n* (%)	1680 (96.6)	124 (93.9)	141 (97.2)
4, *n* (%)	16 (0.9)	1 (0.8)	0 (0.0)
Total number of distractors *n*	5191	390	431
Distractors with frequency = 0%, *n* (%)	1423 (27.4)	99 (25.4)	100 (23.2)
Distractors with frequency 1-4%, *n* (%)	1739 (33.5)	135 (34.6)	127 (29.5)
Total number of non-functioning distractors, *n* (%)	3162 (60.9)	234 (60.0)	227 (52.7)*
Total number of functioning distractors, *n* (%)	2029 (39.1)	156 (40.0)	204 (47.3)*

* indicates a significant finding (p < 0.05) compared to Accepted for use.

Changes in psychometric data in the years following implementation of external peer-review

A potential consequence of implementation of external peer-review, in addition to the standard quality assurance process, is a gradual improvement in the items written each year. We aimed to look for longitudinal changes in peer-review decisions and psychometric parameters.

Peer-review decisions were different across 2016 and 2018 ($p < .001$), with more items passing peer review in 2016 and fewer passing in 2018 (table 3). 2016 also saw a higher number of student appeals than the other years ($p = .035$).

2016 had a higher number of functioning distractors ($p < .001$) and lower number of non-functioning distractors ($p = .043$), while both 2018 and 2019 had subcategories of non-functioning distractors that were higher than 2016. ($p < .001$).

Items were more difficult in the first year of peer-review and thereafter they had a similar difficulty level (Table 3). 2018 had a lower CRPB compared to 2016 ($p = .004$), but there was no difference in mean discrimination index. 2018 had more items with negative discrimination indices and fewer items with discriminating power > .015 ($p = .022$, table 3). In 2019 there were more items with zero discrimination indices (table 3).

Table 3. Item and distractor analysis, peer review decision and student appeals by year

	2016	**2017**	**2018**	**2019**
Number of items, *n*	301	835	865	929
Peer review decision				
Accepted for used, *n* (%)	206 (68.4)	491 (58.8)	442 (51.1)	601 (64.7)
Changes suggested, *n* (%)	32 (10.6)	57 (6.8)	71 (8.2)	84 (9.0)
Failed, *n* (%)	4 (1.3)	8 (1.0)	13 (1.5)	8 (0.9)
Number of student appeals, *n (% of total items)*	36 (12.0)	62 (7.4)	61(7.1)	67 (7.2)
Items changed following student appeal, n (%)	10 (27.8)	23 (37.1)	23 (37.7)	21 (31.3)
Items not changed following student appeal, n (%)	26 (72.2)	39 (62.9)	38 (62.3)	46 (68.7)
Mean difficulty (SD)	0.747 (0.221)	0.780 (0.207)	0.782 (0.216)*	.782 (.210)*
Easy items ($p \geq 0.80$), *n* (%)	166 (55.1)	489 (58.6)	525 (60.7)	548 (59.0)
Middle range items ($0.20 < p < 0.80$), *n* (%)	127 (42.2)	332 (39.8)	322 (37.2)	363 (39.1)
Difficult items ($p \leq 0.20$), *n* (%)	8 (2.7)	14 (1.7)	18 (2.1)	18 (1.9)
Mean discrimination index (SD)	0.194 (0.174)	0.180 (0.155)	0.165 (0.161)	0.179 (.161)
Items with negative discrimination indices, *n* (%)	17 (5.6)	41 (4.9)	71 (8.2)*	49 (5.3)
Items with zero discrimination indices, *n* (%)	26 (8.6)	79 (9.5)	77 (8.9)	106 (11.4)*
Items with discrimination indices .01- 0.15, *n* (%)	101 (33.6)	306 (36.6)	324 (37.5)	307 (33.0)
Total number of non-discriminating items, *n (%)*	144 (47.8)	426 (51.0)	472 (54.6)	462 (49.7)
Total number of discriminating items (> 0.15), *n* (%)	157 (52.2)	409 (49.0)	393 (45.4)*	467 (50.3)
Mean point biserial item discrimination, cRPB (SD)	0.167 (0.142)	0.161 (0.145)	0.139 (0.145)*	0.159 (0.152)
Cognitive level				
K1, *n* (%)	98 (32.6)	289 (34.6)	282 (32.6)	303 (32.6)
K2, *n* (%)	203 (67.4)	546 (65.4)	583 (67.4)	626 (67.4)
Number of distractors per item				
2, *n* (%)	6 (2.0)	25 (3.0)	21 (2.4)	23 (2.5)
3, *n* (%)	284 (94.4)	803 (96.2)	838 (96.9)	900 (96.9)
4, *n* (%)	11 (3.7)	7 (0.8)	6 (0.7)	6 (0.6)
Total number of distractors, *n*	908	2487	2579	2770
Distractors with frequency = 0%, *n* (%)	189 (20.8)	655 (26.3)	785 (30.4)*	770 (27.8)
Distractors with frequency 1-4%, *n* (%)	319 (35.1)	852 (34.3)	792 (30.7)*	909 (32.8)
Total number of non-functioning distractors, *n* (%)	508 (55.9)	1507 (60.6)	1577 (61.1)	1679 (60.6)
Total number of functioning distractors, *n* (%)	400 (44.1)	980 (39.4)	1002 (38.9)	1091 (39.4)

* indicates a significant finding ($p < 0.05$) compared to 2016

Changes in psychometric data throughout the study years

At NTNU, year 1 and 2 is focused on the basic sciences, while the rest of the study program has a more clinical focus. In year 5 and 6, students are quite close to life as a newly educated physician, and as such much closer in knowledge and experience to many of the external peer-reviewers. We wanted to see how all of this played into peer-review decisions and psychometric parameters.

Peer review decision differed between the first two years and the last two years ($p = .001$). More items failed peer-review in the first two years of the study program, while fewer items failed in the last two years.

Peer reviewers were asked how they would consider the relevance of an item in medical education, and results showed significant differences between study years ($p < .001$). In the first two years, more items were deemed less relevant ("Irrelevant" or "Acceptable"), and fewer were deemed relevant ("Important" or "Essential"). The opposite trend was true for the final four years, who both had fewer items deemed less relevant ("Irrelevant" or "Acceptable"). Year 3 and 4 had more items deemed "Important", and year 5 and 6 had more items deemed "Essential" (table 4).

Psychometric parameters also differed across the study years. Items in the two final years were easier than in the first two when looking at mean difficulty ($p = .002$, table 4). Items in year 1 and 2 discriminated better than items in year 5 and 6, as reflected in both mean discrimination index ($p < .001$), CRPB ($p < .001$) and number of discriminating and non-discriminating items ($p < .001$). Year 3 and 4 had a higher CRPB than the first two years ($p < .001$), and a higher number of items with negative discrimination indices ($p < .001$, table 4). The last four years had a higher number of K2-items and a lower number of K1-items when compared to the first two. Year 5 and 6 had more non-functioning distractors and fewer functioning distractors when compared to year 1 and 2 ($p < .001$). In year 3 and 4, there were more distractors chosen by only 1-4% of the students or none of the students.

Table 4. Item and distractor analysis and peer review decision by year of study program

	Year 1 and 2	**Year 3 and 4**	**Year 5 and 6**
Number of items, *n*	670	881	466
Peer review decision			
Accepted for used, *n* (%)	577 (73.8)	767 (50.6)	396 (62.8)
Changes suggested, *n* (%)	72 (9.2)	104 (6.9)	68 (10.8)
Failed, *n* (%)	21 (2.7)	10 (0.7)	2 (0.3)*
Relevance of item in medical education, according to peer reviewer			
Essencial, *n* (%)	92 (15.5)	179 (22.7)	108 (28.6)*
Important, *n* (%)	255 (43.1)	421 (53.4)*	199 (52.8)
Acceptable, *n* (%)	209 (35.3)	175 (22.2)*	65 (17.2)*
Irrelevant, *n* (%)	36 (6.1)	13 (1.6)*	5 (1.3)*
Mean difficulty (SD)	0.776 (0.203)	0.774 (0.212)	0.798 (0.216)*
Easy items ($p \geq 0.80$), *n* (%)	386 (57.6)	515 (58.5)	293 (62.9)
Middle range items ($0.20 < p < 0.80$), *n* (%)	276 (41.2)	350 (39.7)	165 (35.4)
Difficult items ($p \leq 0.20$), *n* (%)	8 (1.2)	16 (1.8)	8 (1.7)
Mean discrimination index (SD)	0.195 (0.158)	0.175 (0.162)	0.156 (0.148)*
Items with negative discrimination indices, *n* (%)	24 (3.6)	60 (6.8)*	23 (4.9)
Items with zero discrimination indices, *n* (%)	55 (8.2)	84 (9.5)	64 (13.7)*
Items with discrimination indices 0.01-0.15, *n* (%)	224 (33.4)	306 (34.7)	180 (38.6)
Total number of non-discriminating items, *n (%)*	303 (45.2)	450 (51.1)	267 (57.3)*
Total number of discriminating items (> 0.15), *n* (%)	367 (54.8)	431 (48.9)	199 (42.7)*
Mean point biserial item discrimination, CRPB (SD)	0.187 (0.137)	0.148 (0.151)*	0.124 (0.126)*
Cognitive level			
K1, *n* (%)	318 (47.5)	231 (26.2)*	114 (24.5)*
K2, *n* (%)	352 (52.5)	650 (73.8)*	352 (75.5)*
Number of distractors per item			
2, *n* (%)	27 (4.0)	23 (2.6)	5 (1.1)*
3, *n* (%)	639 (95.4)	849 (96.4)	457 (98.1)*
4, *n* (%)	4 (0.6)	9 (1.0)	4 (0.9)
Total number of distractors *n*	1986	2629	1397
Distractors with frequency = 0%, *n* (%)	374 (18.8)	779 (29.6)*	469 (33.6)*
Distractors with frequency 1-4%, *n* (%)	780 (39.3)	769 (29.3)*	452 (32.4)
Total number of non-functioning distractors, *n* (%)	1154 (58.1)	1548 (58.9)	921 (65.9)*
Total number of functioning distractors, *n* (%)	832 (41.9)	1081 (41.1)	476 (34.1)*

* indicates a significant finding ($p < 0.05$) compared to Year 1 and 2

Peer-review and effect on student appeals

The student appeal is part of the post-test quality assurance at NTNU, and is a possibility for the students to give feedback on items. We wanted to see whether this feedback would fall in line with feedback from peer-reviewers, and looked at the number of student appeals as well as examination committee decision and compared it to the decision of peer-reviewers.

We found that items that were requested revised or rejected by the peer-reviewer, but left unchanged by the item writer, were more subject to a student appeal (Table 5).

Table 5. Student appeal analysis by review decision and changes made for MCQs used in ordinary examinations

	Accepted for use	**Requested revised or rejected**	
		Changed	**Unchanged**
Number of items, *n*	1740	132	145
Student appeals			
Items that students appealed, *n (%)*	135 (7.8)	8 (6.1)	23 (15.9)*
Items that students did not appeal, *n (%)*	1605 (92.2)	124 (93.9)	122 (84.1)*
Examination committee decision			
Student appeals that were accepted, *n (%)*	48 (35.6)	4 (50.0)	8 (34.8)
Student appeals that were not accepted, *n (%)*	87 (64.4)	4 (50.0)	15 (65.2)

* indicates a significant finding ($p < 0.05$) compared to Accepted for use

Peer reviewers prediction of relevance and difficulty

Relevance of an item is an important part of creating good test items, and also a potensial area of dispute, especially in the basic sciences. We wanted to investigate how peer-reviewers decision on item relevance compared to item difficulty.

Peer reviewers were asked to categorize how relevant they saw an item to be in the education of future doctors. They could choose between “Essential”, “Important”, “Acceptable” and “Irrelevant”. We looked at the difficulty level of items in these categories of relevance. The more

relevant the item was conceived by the reviewer the lower the difficulty of the item ($p < .001$, figure 2). When comparing pairwise, there was a significant difference between all groups, except Irrelevant and Acceptable.

Figure 2. Difficulty by peer-reviewers decision on item relevance.

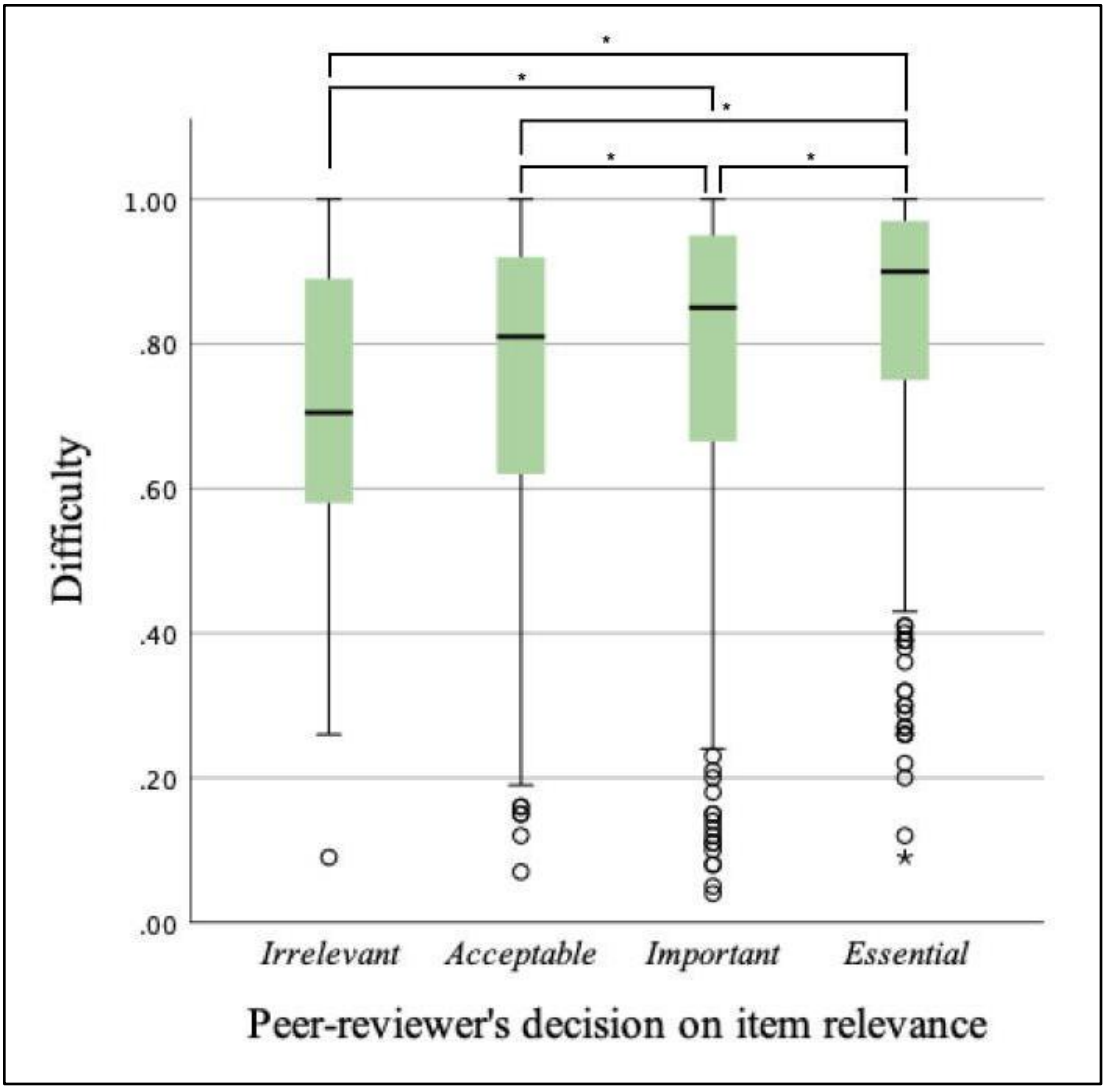


*Figure shows the result of a Kruskal-Wallis Test comparing difficulty and peer-reviewer's decision on item relevance. * along bracket indicates a significant difference ($p < .05$). The black line indicates the median. Green boxes represents interquartile range, and contains 50 percent of cases. Whiskers show smallest and largest value for each category. Circles indicate values deemed as outliers by SPSS. Asterisk below circle indicates values deemed as extreme outliers by SPSS.*

Peer reviewers were also asked to answer what percentage of students they thought would answer an item correctly. This was compared to the actual difficulty on that item. Spearman's correlation showed a significant positive correlation between difficulty and peer reviewers estimate ($p < .001$, rho = .205) (Figure 3).

Figure 3. Peer reviewer's approximation of difficulty compared to actual difficulty

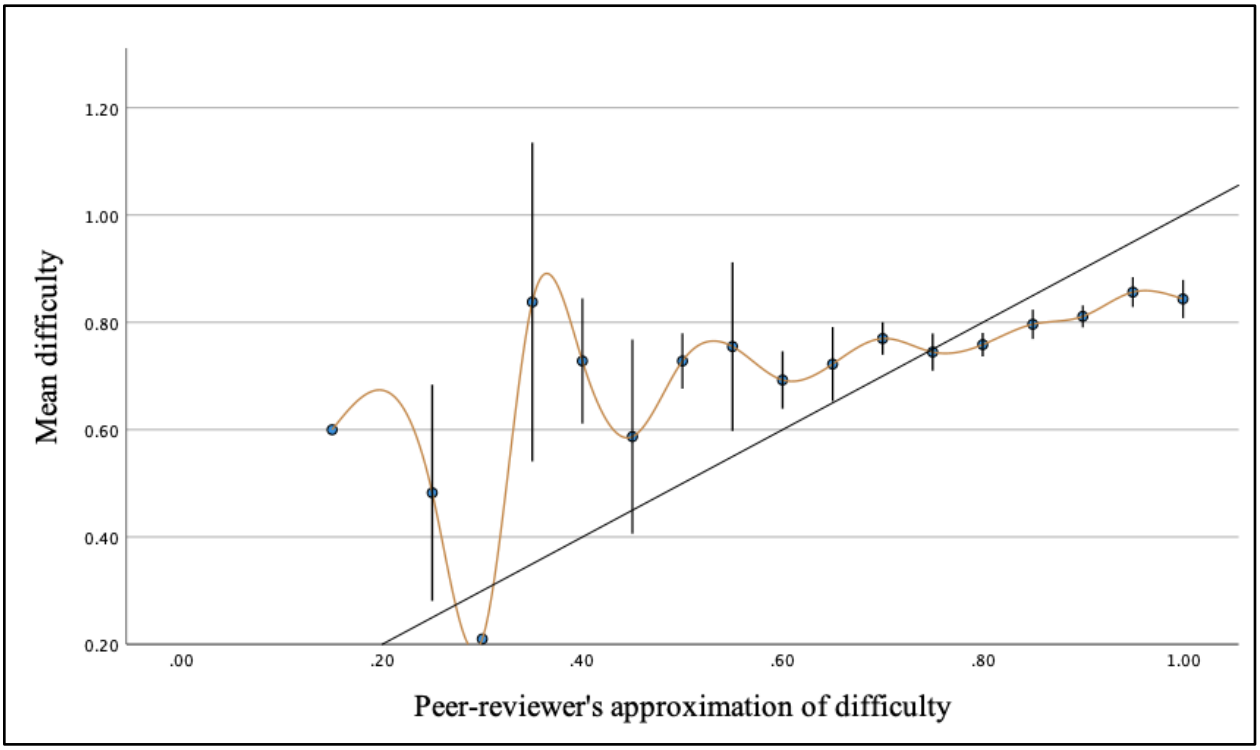


Figure shows the result of a simple error bar graph. Blue dots indicate mean of difficulty. Error bars indicate 95% CI. Yellow line represents interpolation line.

Discussion

The purpose of this study was to explore how external, double-blinded peer review affected the quality of summative examinations, by looking at psychometric measures, longitudinal effects and student feedback. We studied 2930 MCQ items of which 2017 were peer-reviewed. Results showed differences in items depending on peer-review decision, with unchanged items being more difficult, having more functioning distractors and fewer non-functioning distractors than those accepted for use, as well as having a higher number of K1-items and lower number of K2-

items. We found no positive longitudinal effect on psychometric measures after implementing external peer-review. Items in the following years were easier, had more non-functioning distractors and did not discriminate as well as the first year.

When comparing different study years, we found that fewer items failed peer review in the two final years, and that these items were more likely to be deemed relevant by peer-reviewers. The opposite was true for the first two years. Items in year 1 and 2 discriminated better than items in year 5 and 6, and were more difficult. The last four years had a higher number of K2-items and a lower number of K1-items when compared to the first two. Year 5 and 6 had more non-functioning distractors and fewer functioning distractors when compared to year 1 and 2. In year 3 and 4, there were more distractors chosen by only 1-4% of the students or none of the students.

Items requested revised or rejected but not changed received a higher amount of student appeals after the examination. We also found that items deemed more relevant by peer-reviewers were likely to be less difficult. We also examined the correlation between peer-reviewers estimate of difficult and the actual difficulty, and found a positive correlation.

Relevance is a primary area of disagreement between peer reviewers and item writers

In this study we have shown a significant difference in difficulty depending on the review decision, where items that passed peer review were easier than those who did not pass and were not changed in accordance with the peer-reviewers suggestion.

The majority of peer reviewers in the program were young clinicians who had at least two years of postgraduate training but had not yet completed their postgraduate training. This was a planned selection of external peer-reviewers since we train medical students to become ready for their first years as doctors. During post graduate training they will accumulate more knowledge and skills making them experts. Young doctors might have a different viewpoint than that of the item writers, who are largely experts within their field, with many years of clinical, research, and teaching experience. Teachers tend to have unrealistically high expectations of students [15]. The external peer-reviewers could partly compensate for this. Our results indicate that items probably

became easier with the reviewers suggested adjustments. One can argue that the difficulty level in the adjusted items were closer to what to expect of a newly educated doctor within the field. The medical program at NTNU uses a criterium-based approach, meaning the goal of the examination is to set a threshold where students who pass possess the minimum knowledge required to practice medicine safely. An unnecessarily high difficulty level is thus not desirable. At the same time, setting the bar too low would be very problematic as you could risk educating physicians that do not have the skills they need to maintain patient safety.

Difficulty and relevance seem to go hand in hand. When asked about an item's relevance in medical education, peer reviewers were more likely to consider easier items more relevant ("Essential" or "Important") (see figure 2). This seems to be in line with students, who often give feedback about being tested on what they see as "peripheral" knowledge when looking at the curriculum and what they have been taught [15]. In our experience, this is also common reasoning in the student appeals at NTNU.

Easier items being deemed as more relevant ("Essential" and "Important") implies that students are able to identify and learn central aspects of the curriculum, which hints towards medical students and peer reviewers being in line with what they deem as the most relevant knowledge. Smeby et al. [12] found that a main theme in the comments of disapproved items was irrelevance to clinical practice. This is also supported by a higher number of items deemed less relevant ("Irrelevant" or "Acceptable") in the first two years of the study program, where the main focus is basic science, not clinical knowledge.

As medical school is a professional study, the curriculum must be relevant to the clinical life students enter after graduation. Students often use previous examinations to guide their reading and figure out what is essential [2, 7]. Hence, the relevance of what is asked on the examination is essential, especially in a high-stakes setting. However, clinicians and item writers often disagree on what could be considered essential knowledge [16-18].

The number of K1-items in the different peer review categories also points towards a disagreement between clinicians and item writers. Fewer K1-items passed review. This is consistent with Smeby et al. [12], who analyzed peer review comments and found that many

items were rejected because they only tested recall of knowledge. Reviewers argued that this information could easily be looked up or was irrelevant, and as such should not be tested on an examination. However, a higher amount of K1-items were not changed by the item writer following disapproval in peer review. This indicates a disagreement between the item writer and the reviewer on the relevance of the tested knowledge. This is especially true for the first two years, where the amount of K1-items is higher, and as previously discussed - reviewers and item writers potentially stand further apart.

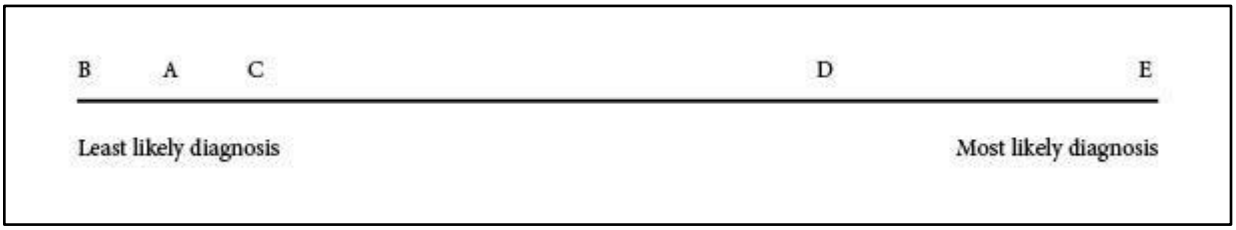
Although clinicians and teachers might agree on important broad concepts to learn, they disagree on the needed depth of knowledge within each concept. Some of this might stem from the difference in the level at which they usually do their work. [17]. One could argue that most medical students will do their work on a clinical level, and as such the peer reviewers' opinion on relevance should be given great value - it could help an item writer correct a focus on a too deep level in their testing. On the other hand, the reviewers and the medical students could be too early in their career to yet know the utility of what may at first seem like irrelevant information. They may see information that goes beyond what they usually need in everyday clinical day life as unnecessary. Further along in their career, they may discover that knowledge they previously saw as useless is now essential, and something they would otherwise not have time to catch up on. However, Koens, Rademakers, & Cate [18] found postgraduate trainees and experienced specialists to agree on what constitutes core knowledge and what does not. This is also compensated for in the peer review process, where the item writer and examination committee have the final say in whether or not to change the item in question.

There is also the possibility that a peer reviewer might see an item as relevant or irrelevant depending on whether they are able to answer it correctly themselves [19]. In our peer-review process the reviewer had to answer the item in the reviewer portal prior to seeing the answer and write their review. This could introduce a bias of deeming more items irrelevant.

Items that were not changed following peer review had more functioning distractors

Items that were not changed following peer review had a higher amount of functioning distractors. The peer-review process itself could explain this. If a reviewer sees a distractor as equally good to the correct answer and rejects or suggests changes to an item, and the item writer decides to change the item they will most likely exchange what could have been a well-functioning distractor with a poor distractor. If the item is not changed, however, the students are likely to see this as a good alternative as well, and it will be a well-functioning distractor. A distractor has to be selected by a certain amount of students to serve its function, which is to identify weaker candidates - a critical function in a high stakes examination. At the same time, if many students choose a distractor, that could suggest the item or alternatives are flawed or misleading [20].

Figure 4. Distribution of distractors. With one-best-answer items, distractors do not need to be wrong, and could for example be diagrammed as here.



Panigua, M.A, & Swygert, K.A, 2016. Constructing Written Test Questions For the Basic and Clinical Sciences. NBME.

The number of functioning distractors is more rarely used as a measure of quality than item discrimination, but could be an equally good measure [21]. Teachers tend to focus on an item's stem instead of the distractors, even though plausible distractors play a crucial role in developing high-quality items and tests [8, 22]. The possibility of losing good distractors is consequently something to be highly cautious of.

As the students progressed in their studies, items were easier and less discriminating with fewer distractors

As the students progressed in their studies, items were easier and less discriminating with fewer distractors. Students improving during their course of study is expected, but having items with worse discriminating statistics in the final years is not consistent with current findings.

Colberg et al. [13] looked at standard-setting at the medical program at NTNU and found that the failure rate declined as students progressed in the study program. They suggested several reasons for this, including spiral learning, difficulties adjusting from upper secondary school to higher education, and the fact that 73% of those quitting the study or having their admission revoked because of repeated examination failures do so in the first two years. This leaves behind highly motivated students who have mastered this form of study, which helps explain the progress.

Less discriminating items with fewer well-functiong distractors as students progress in the study is unexpected. As the students move in a more clinical direction, item writers can draw from their own clinical experience and present clinical problems as they often present themselves - unclear and without an obvious diagnosis. From this perspective, it is easy to write multiple items from the same problem, as a patient needs to receive both the proper investigations, diagnosis, and treatment. [23]. However, this finding can be explained by the decreased difficulty as the years progress, as decreased difficulty will affect the discriminating power of the item.

Difference in psychometric measures when comparing academic years

2016 was the first year of examinations after implementing substantial measures to improve the quality of examinations, including a written guide on item writing, training, and peer review. When comparing the academic years, 2016 had more difficult items than the other years, a higher cRPB, and a higher number of functioning distractors, as well as fewer non-functioning distractors. The 2016 statistics may represent the work put into training and guiding, and that the knowledge and motivation among staff have somewhat diminished. There have also been staff replacements, and some knowledge on item writing might have been lost and not replaced in this

process. The Hawthorne Effect is also a possible explanation - when item writers knew they were being observed they would do better. It would be logical for this effect to diminish over the years.

There is a possibility that the statistics are a reflection of peer reviewers helping to identify flawed items, which would be in line with several other studies [5, 10, 11]. At the same time, our study found that items that were not changed following peer review had a higher number of functioning distractors, with the downsides discussed previously in this article. An unintended consequence of this could be easier items.

Items that were not changed following peer review received more student appeals

Others have also implemented post-test student feedback as part of their test framework [24, 25]. At the Maastricht Medical school, 7% of items are withdrawn from the test following student comments [25].

The process immediately seems to be in the students' favor, but there are several aspects to discuss. The appeal can be seen as negative feedback on the item writer's work, and is often quite detailed with extensive argumentation and criticism. As a consequence, it is worth discussing how doctors react to negative feedback. A systematic review by Baines et al. [26] found that patient feedback to doctors could impact medical performance. However, main critical factors needed to be present in both the collection of feedback, the presentation of feedback, and the overall feedback culture of the workplace. This included the feedback being specific and from a credible source, good quality feedback facilitation, and the formulation of actionable behavior change.

Neither the students nor the item writers involved in the student appeal process are trained in giving and receiving feedback, meaning there is no guarantee for any of these factors to be present in each setting. If the item writer perceives the feedback as helpful, specific, and not an attack on how they perceive themselves, it might have a positive effect in many ways. For example, the students could get feedback on their appeal and either be affirmed in their research

or receive a learning opportunity, depending on whether they are judged to be right or wrong. The item writer can use the feedback to improve their item writing further.

If the feedback process is not ideally set up, negative feedback could initiate greater emotional reactions in the item writer, which does not necessarily lead to a behavioral change. Facilitation of reflections seems to be key in this [27]. As there is no facilitation of reflection in the current setup of this process, many item writers may react to this as criticism, with their answer being a defense of their item writing rather than a reflection on what could be done differently.

The students do not have the possibility of responding to the item writer and examination committee. Hence, there is no dialogue or possibility of correcting interpretations on the item writer's part. This could be a disadvantage to the students if the item writer and examination committee misses important points or if the appeal is not clear enough in the first place, they have no means of correcting this.

The finding that items that were not changed following peer review received more student appeals indicates that students and peer reviewers often align on what is perceived as problematic items, and that possibly more time should be spent on quality assurance of these items.

Peer reviewers were able to estimate the difficulty of items

Examinations should strive to include items with a range in difficulty. Item writers make assumptions about the difficulty beforehand, but students often respond to questions in unexpected ways, giving a different difficulty than expected [23].

There was a significant positive correlation between difficulty and peer-reviewer estimate, but with a poor calibration, indicating that peer reviewers would need substantially more training to set the difficulty correctly. Peer reviewers were not able to calibrate their answers beforehand by getting any feedback on mock questions, and also did not receive any feedback on whether they estimated correctly on the actual questions.

An estimate of 70% of students answering correctly seems to be a breaking point. When estimating that fewer students than this would answer correctly, reviewers tended to underestimate the students, and vice versa when estimating that more than 70% of students would answer correctly. This could partly be explained by the majority of items analyzed having a difficulty of .60 or higher.

Strengths and weaknesses

We did not use a control group in this study. Using earlier, not peer reviewed examinations as a control group seems like an obvious choice; however, the peer review process was just one of many quality improvement measures implemented simultaneously. It would have been challenging to separate the impact of peer review from other measures.

Peer reviewers receive limited training in item writing, and are asked to review the items as clinicians. They are sent faculty guidelines with information on important aspects of item writing, such as item writing flaws and psychometric measures.

Another limitation is that although the number of items analyzed is high, when subcategorized, some categories end up with few items, leading to weaker analyses. Several more examinations would be needed to compensate for this and give more robust answers.

In summary

This study looked at changes and associations in psychometric measures and student feedback following the implementation of external peer-review of MCQs in a medical school. The analyses showed several ways peer-review can affect item quality, and establishes important discussion points that can be used to further improve the quality assessment program. Despite the extensive material, using the data of more examinations would be helpful in order to get a better picture of the effect of peer-review over time.

Conclusions

The study showed that peer-reviewers are able to identify items with weaknesses, which can lead to a revision of the item. In those instances where item writers and peer-reviewers disagree, they seem to disagree most on the relevance of basic science items and K1-items. Student appeal numbers indicated that students and peer-reviewers agree when identifying items of concern.

References

1. van der Vleuten, C.P., et al., *The assessment of professional competence: building blocks for theory development.* Best Pract Res Clin Obstet Gynaecol, 2010. **24**(6): p. 703-19.
2. Baartman, L.K.J., et al., *The wheel of competency assessment: Presenting quality criteria for competency assessment programs.* Studies in Educational Evaluation, 2006. **32**(2): p. 153-170.
3. Downing, S.M. and R. Yudowsky, *Assessment in health professions education*. 2009, New York: Routledge.
4. Schuwirth, L.W. and C.P. van der Vleuten, *Different written assessment methods: what can be said about their strengths and weaknesses?* Med Educ, 2004. **38**(9): p. 974-9.
5. Wallach, P.M., et al., *Use of a committee review process to improve the quality of course examinations.* Adv Health Sci Educ Theory Pract, 2006. **11**(1): p. 61-8.

6. Downing, S.M., *The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education.* Adv Health Sci Educ Theory Pract, 2005. **10**(2): p. 133-43.
7. Jozefowicz, R.F., et al., *The quality of in-house medical school examinations.* Acad Med, 2002. **77**(2): p. 156-61.
8. Tarrant, M. and J. Ware, *Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments.* Med Educ, 2008. **42**(2): p. 198-206.
9. Melnick, D.E., *Licensing examinations in North America: is external audit valuable?* Med Teach, 2009. **31**(3): p. 212-4.
10. Abozaid, H., Y.S. Park, and A. Tekian, *Peer review improves psychometric characteristics of multiple choice questions.* Med Teach, 2017. **39**(sup1): p. S50-S54.
11. Malau-Aduli, B.S. and C. Zimitat, *Peer review improves the quality of MCQ examinations.* Assessment & Evaluation in Higher Education, 2012. **37**(8): p. 919-931.
12. Smeby, S.S., et al., *Improving assessment quality in professional higher education: Could external peer review of items be the answer?* Cogent Medicine, 2019. **6**(1).
13. Colberg, A.B., et al., *How can the examination failure rate be stabilised?* Tidsskriftet Den Norske Legeforening, 2017. **20**.
14. Verhoeven, B., et al., *Quality assurance in test construction: The approach of a multidisciplinary central test committee.* Education For Health: Change in Learning and Practice, 1999. **12**(1): p. 49-60.
15. Cohen-Schotanus, J. and C.P.M. van der Vleuten, *A standard setting method with the best performing students as point of reference: Practical and affordable.* Medical Teacher, 2010. **32**(2): p. 154-160.
16. Janssen-Brandt, X.M.C., A.M.M. Muijtjens, and D.M.A. Sluijsmans, *Toward a better judgment of item relevance in progress testing.* BMC Med Educ, 2017. **17**(1): p. 151.
17. Koens, F., E.J. Custers, and O.T. ten Cate, *Clinical and basic science teachers' opinions about the required depth of biomedical knowledge for medical students.* Med Teach, 2006. **28**(3): p. 234-8.
18. Koens, F., J.J. Rademakers, and O.T. Ten Cate, *Validation of core medical knowledge by postgraduates and specialists.* Med Educ, 2005. **39**(9): p. 911-7.
19. Harris, J.A., H.C. Heneghan, and D.W. McKay, *The rating of pre-clerkship examination questions by postgraduate medical students: an assessment of quality and relevancy to medical practice.* Med Educ, 2003. **37**(2): p. 105-9.
20. *Guidelines for the Development of Multiple-Choice Questions*, M.C.o. Canada, Editor. 2010.
21. Ware, J. and T. Vik, *Quality assurance of item writing: during the introduction of multiple choice questions in medicine for high stakes examinations.* Med Teach, 2009. **31**(3): p. 238-43.
22. Haladyna, T.M. and S.M. Downing, *How Many Options is Enough for a Multiple-Choice Test Item?* Educational and Psychological Measurement, 1993. **53**(4).
23. *Constructing Written Test Questions For the Basic and Clinical Sciences*. 2016, Philadelphia: National Board of Medical Examiners.
24. Wrigley, W., et al., *A systemic framework for the progress test: Strengths, constraints and issues: AMEE Guide No. 71.* Medical Teacher, 2012. **34**: p. 683-697.
25. Van der Vleuten, C., G.M. Verwijnen, and W.H.F.W. Wijnen, *Fifteen years experience of progress testing in a problem-based curriculum.* Medical Teacher, 2009. **18**(2): p. 103-109.
26. Baines, R., et al., *The impact of patient feedback on the medical performance of qualified doctors: a systematic review.* BMC Med Educ, 2018. **18**(1): p. 173.
27. Sargeant, J.M., et al., *Reflection: a link between receiving and using assessment feedback.* Advances in Health Sciences Education, 2009. **14**(3): p. 399-410.

NTNU

Norwegian University of
Science and Technology