Smeby et al., *Cogent Medicine* (2019), 6: 1659746
https://doi.org/10.1080/2331205X.2019.1659746

cogent medicine

EVIDENCE-BASED MEDICINE & MEDICAL INFORMATICS | RESEARCH ARTICLE

Improving assessment quality in professional higher education: Could external peer review of items be the answer?

Susanne Skjervold Smeby[1*], Børge Lillebo[2,3], Vidar Gynnild[4], Eivind Samstad[1,5], Rune Standal[6], Heidi Knobel[1,7], Anne Vik[8,9] and Tobias S. Slørdahl[1,10]

Received: 18 January 2019
Accepted: 20 August 2019
First Published: 27 August 2019

*Corresponding author: Susanne Skjervold Smeby, Olav Kyrres gate 9, PLUS-senteret, Medisinsk-teknisk forskningssenter, Trondheim 7030, Norway
E-mail: susanne.s.smeby@ntnu.no

Reviewing editor:
Anne Spurkland, University of Oslo, Norway.

Additional information is available at the end of the article

Abstract: Summative assessment in professional higher education is important for student learning and making sound decisions about advancement and certification. Despite rigorous pre-test quality assurance procedures, problematic assessment items are always discovered post-test. This article examines the implementation of external peer review of items by clinicians in a six-year undergraduate medical programme. The purpose of the article is to identify to what extent clinicians consider multiple choice items to be acceptable for use in examinations, and what comments they provide on items they believe should be revised or not be used at all. 170 clinicians were recruited and reviewed 1353 multiple choice questions. Results showed that one out of five items reviewed were not approved. There were three main reasons for not approving items: (i) relevance of item content, (ii) accuracy of item content and (iii) technical item writing flaws. The article provides insight into a promising quality assurance procedure suitable for in-house examinations in professional higher education.

Subjects: Higher Education; Assessment; Medical Education

Keywords: assessment quality; item relevance; medical education; multiple choice questions

1. Introduction

Professional higher education strives to teach students the competencies they will need in their future professions. This encompasses both subject-specific and generic competencies that prepare students for the complex problems of today's workplace, as well as life-long learning and development (Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2006; Van der Vleuten, Schuwirth,

ABOUT THE AUTHOR

Susanne Skjervold Smeby is a medical doctor and PhD-student at the Norwegian University of Science and Technology. Her research interests include assessment quality and the relation between assessment and learning.

PUBLIC INTEREST STATEMENT

In professional higher education, the link between test content and professional practice is especially important to make sound certification decisions. This research uses clinicians to review examination content in a medical undergraduate programme. One in five examination questions were not approved by clinicians, and relevance and accuracy of content were two of the main reasons. Consulting practitioners in the field may lead to more relevant and accurate content, increasing the validity of examinations.

Scheele, Driessen, & Hodges, 2010). Developing high quality summative assessment is important for both student learning and sound advancement decisions. In the field of medicine, both employers and patients rely on medical schools' ability to certify that students have the knowledge, skills and attitudes necessary to practice medicine safely.

Summative assessment in undergraduate medical education can be in-house examinations prepared by academic staff involved in teaching, or national examinations generally prepared by licensing organisations. Test items in national examinations are usually written and extensively reviewed by subject-specific test committees trained in item writing. National examinations also provide an arena for relevant stakeholders to engage in the process of assessment design, content and standards for entry into practice (Melnick, 2009). Such measures typically result in high quality assessment, but they come at a high cost (Melnick, 2009). Although the use of national licensing examinations in medicine is likely to increase, the majority of examinations are in-house (Swanson & Roberts, 2016). Therefore, developing quality assurance procedures that can be implemented for in-house settings with fewer resources is important.

Written assessments make up a large part of assessments in medical education, along with assessments that cover other important competencies, such as communication, professionalism and clinical skills. Despite the fact that multiple choice questions (MCQs) have many advantages and disadvantages, they remain the most frequently used assessment method in medicine (Wallach, Crespo, Holtzman, Galbraith, & Swanson, 2006). They are efficient for use in large groups of examinees as they can be administered in a relatively short period of time and are easily computer scored (Downing & Yudkowsky, 2009). Additionally, MCQs can test a large breadth of knowledge as well as higher-level cognitive reasoning (Downing & Yudkowsky, 2009; Schuwirth & Van Der Vleuten, 2004). There are best practice principles for writing effective MCQs, and violating these standards is termed item writing flaws (IWFs) (Case & Swanson, 1998; Haladyna, Downing, & Rodriguez, 2002). IWFs reduce assessment validity by introducing the systematic error of construct-irrelevant variance, and have been shown to occur frequently in in-house examinations (Downing, 2005; Jozefowicz et al., 2002).

Quality assurance procedures around test item development and administration is necessary for high quality assessment (Van der Vleuten et al., 2010). These include faculty development programmes on proper item writing and blueprinting, review of items through internal review committees and psychometric evaluation, as well as student feedback. Item writing workshops have been shown to improve quality of MCQs in terms of difficulty and item discrimination, and reduces the frequency of IWFs (Abdulghani et al., 2015). Several studies have documented the effect of in-house peer review of MCQs (Abozaid, Park, & Tekian, 2017; Malau-Aduli & Zimitat, 2012; Wallach et al., 2006).

In our medical programme, MCQs are subject to review similar to that of the Maastricht model (Verhoeven, Verwijnen, Scherpbier, & Schuwirth, 1999). The departments write items based on a blueprint for the end-of-year examinations, and are entered into a web-based item bank. A multidisciplinary review committee (examination committee) reviews items for content, clarity and IWFs. In addition, one or two senior students are asked to comment on the examination draft. However, despite rigorous review, we still discover problematic items through post-test item analyses and student comments, as is experienced by other institutions (Verhoeven et al., 1999).

In an attempt to reduce the number of problematic items discovered post-test, we developed an additional review process suitable for in-house examinations in professional higher education. We were interested in consulting front line practitioners in the field, inviting them to share their thoughts on examination items through external, double-blinded peer review of MCQs in an undergraduate medical programme. The aim was to explore the following research questions:

(1) To what extent are items considered approved, needing revision or rejected by clinicians?

(2) What comments are provided by clinicians on items considered needing revision or rejected?

(3) To what extent are items changed by the item writer or examination committee following external peer review?

2. Materials and methods

2.1. The medical curriculum and assessment programme

The six-year undergraduate medical programme at the Norwegian University of Science and Technology (NTNU) is integrated and problem-based, featuring one oral and one written summative examination at the end of each year (Ware & Vik, 2009). Written examinations consist of 100–120 single best answer MCQs and several modified essay questions (MEQs). The examinations are pass or fail with a cut-off score of 65%.

2.2. The intervention: external peer review of MCQs

Clinicians as reviewers were recruited with the following inclusion criteria: (a) at least two years of postgraduate training, (b) not completed postgraduate training, although this did not apply to specialists in general practice, (c) does not teach at or write items for the faculty. These criteria were chosen because we considered junior doctors and general practitioners to be qualified to judge whether the content followed recommended clinical guidelines and practice, and its relevance for medical students. All reviewers were required to sign a research consent form and asked to complete a questionnaire on personal background information (Table 1). The study was approved by the Norwegian Centre for Research Data (project number: 45229).

In all, 172 reviewers were recruited, of which two reviewers later withdrew. Recruitment started among colleagues perceived to be highly professionally competent, and continued as snowball sampling in which reviewers recommended their own colleagues. Clinicians were recruited for a period of three years, and the annual work-load was estimated to be two hours. They received no

Table 1. Characteristics of reviewers

Age	
Mean, *years (min, max)*	32 (27, 63)
Gender	
Female, *n (%)*	75 (49.7)
Male, *n (%)*	76 (50.3)
Position	
General practitioner, *n (%)*	8 (5.3)
Junior doctor, *n (%)*	135 (89.4)
Mean number of months approved in specialist training (SD)	29.2 (16.2)
PhD student or researcher, *n (%)*	3 (2.0)
Other, *n (%)*	5 (3.3)
Workplace	
GP surgery, *n (%)*	19 (12.6)
District hospital, *n (%)*	67 (44.4)
University hospital, *n (%)*	61 (40.4)
Other, *n (%)*	4 (2.6)

Response rate: 151 (88%) responded to the questionnaire.

financial compensation for reviews, but were registered as external employees and given access to the university's resources, including IT facilities. We aimed at recruiting clinicians from multiple hospitals and GP surgeries, from different counties in Norway (n = 18), with a background from various medical schools (n = 13) and in various specialities (n = 33).

The external peer review was implemented as an additional step in the quality assurance procedure already in place (Figure 1). The items had been reviewed and approved by the multidisciplinary review committee prior to the external peer review. Items were sorted by subject and distributed to reviewers specialising in the topic covered by that item. For subjects that did not have a clear link to a medical specialty, items were pooled and divided between all reviewers. The reviewer and item writer did not know each other's identity. Each reviewer received one to ten items, and each item was assessed by only one reviewer. The whole review process, including distribution of items and completing the review, was carried out by way of a web-based item bank which could be accessed from home. Reviewers had two weeks to complete their review.

The reviewers received limited training with regards to item writing and reviewing. They were sent written information on the MCQ format and the review process, along with the item writing guidelines. Before the correct option was revealed to reviewers, they had to answer each item. Reviewers were asked to consider the questions' relevance, whether the correct option undoubtedly is the best option and the suitability of the explanation of the correct option. They had to indicate whether an item should be approved, revised prior to use, or rejected. If an item was deemed as needing revision or rejected, reviewers were asked to provide a comment. The reviews were disclosed to the item writers, who decided whether to revise or delete the item, or leave it unchanged. If an item was left unchanged, a comment explaining their decision had to be provided to the examination committee. The examination committee made the final decision on whether items should be included in the examination, and could also make changes to or delete items.

2.3. Data collection and analysis

Summative MCQs administered to students in Year 1, 2, 3 and 6 for the academic year of 2015/2016, in addition to previously used MCQs from Year 4, were externally reviewed. This study uses a mixed method approach, with both qualitative and quantitative data to answer the research questions. The following data were registered: Review decision, reviewer comments and whether the item was changed or deleted by the item writer/examination committee. The main part of this study focuses on a qualitative analysis of reviewers' comments to answer the second research question. Reviewers' comments were analysed using Systematic Text Condensation (STC) according to Malterud's description (Malterud, 2012). STC is a descriptive cross-case analysis used to capture significant themes in the empirical material. The analysis started by reading through all

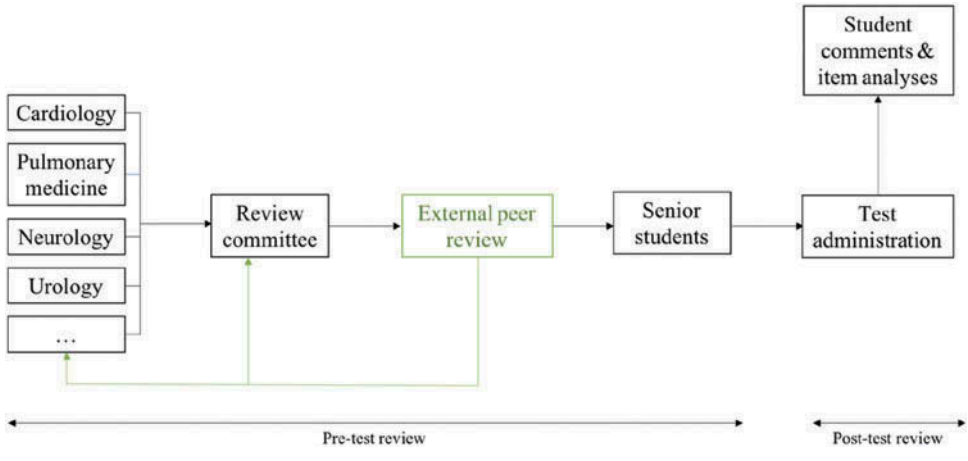

Figure 1. External, double-blinded peer review. The green box indicates where the external peer review was incorporated in the item review process previously in use.

reviewers' comments to get an overall impression of the material and preliminary themes. Meaningful text representing reviewers' reasons for disapproving items was coded into the main themes, adding new themes as they became apparent. In the third step, subthemes within main themes were identified and the contents of each group were condensed into an artificial quote. Lastly, the content of each group was summarised in generalised descriptions and illustrated by selected quotes. All comments were read and themes discussed by three authors (SSS, VG, BL) to widen the analytic space. Quotes were translated, and edited only to improve readability.

3. Results

3.1. Review decisions

Of the 1353 items that were externally reviewed, 282 (20.8%) were not considered approved. Of these, 229 (16.9%) were judged as needing revision and 53 (3.9%) were rejected by reviewers (Table 2). Item writers and examination committees made changes to 115 (40.8%) of disapproved items. Of these, 96 (34.0%) were revised and 19 (6.7%) were deleted. In total, 8.5% of all the items reviewed were changed following external review.

3.2. Reviewer comments

Reviewers' comments fell within three main themes, each with three subthemes: *content relevance* (level of difficulty, importance of content, and cognitive level), *content accuracy* (missing information, content errors, and uncertainty about professional content) and *technical flaws* (spelling and language, structure, and lack of explanation of correct option).

3.2.1. Content relevance

The relevance of item content for medical students was a frequent reason for disapproving items. This included the level of difficulty, importance of content and cognitive level tested by the item.

Level of difficulty: Many reviewers commented that the item content was too difficult for undergraduates. Some remarked that the knowledge asked for was too in-depth, whereas others wrote that the content would be better suited in graduate medical education. One reviewer wrote: "This topic has far more relevance in specialist training than in final undergraduate examinations." Only three items were disapproved by reviewers on the basis of being too easy. One reviewer stated that the content should be presumed knowledge, and therefore unnecessary to ask about in an examination.

Importance of content: Some reviewers commented that the topic covered by items was peripheral as opposed to core areas of the curriculum, and others remarked that the item covered rare symptoms and diseases, and therefore unlikely to be encountered by junior doctors. One such comment was: "The item is irrelevant, and this type of detailed knowledge cannot be deemed essential for clinical practice." Irrelevance for later clinical practice was a frequent reason for not approving items.

Table 2. Review decision. Review decision and subsequent changes made to items by item writer/examination committee

	Review decision			Total
	Approved	Revision needed	Rejected	
Unchanged, *n*	n/a	125	42	1238
Revised, *n*	n/a	90	6	96
Deleted, *n*	n/a	14	5	19
Total, *n*	1071	229	53	1353

Cognitive level: Many items were disapproved because they only tested recall of knowledge. Reviewers commented that such facts would either not be relevant in clinical practice or when needed can be looked up in the literature, as illustrated by the following two quotes: "A very narrow question that only tests students' ability to recall knowledge," and "the question should be more comprehensive, enabling students to use their reasoning skills to a greater extent." This applied especially to items asking for numbers or percentages, for example prevalence of disease.

3.2.2. Content accuracy

The accuracy of the item content was also commonly remarked by reviewers. These comments related to items that were missing key bits of information, had errors in the content or items where reviewers were unsure about the accuracy of the content.

Missing information: Most comments on content accuracy related to missing information in the stem or options, as exemplified by the quote: "There is not enough information in the stem to provide a good and unambiguous answer." In a few cases, reviewers specified that certain details were missing, thereby making more than one option correct. In one such item, the reviewer wrote: "It should be specified in the stem that this applies for children with a birth weight above 2.5 kg. For children with low birth weight, option C is the correct answer." Many reviewers commented that the stem or options did not provide information that would normally be present in a real clinical situation, making the item hypothetical rather than realistic: "The item should include more information such as temperature, heart rate and blood pressure, which you would have access to in a real clinical situation."

Content errors: Reviewers also came across content errors, some of which related to improbable symptoms or findings. Other items were based on outdated guidelines or classification systems. The following quotes illustrate typical examples of errors:

> "The question asks for a probable diagnosis in a patient with a broad complex tachycardia with a ventricular rate that fits well with an atrial flutter with 2:1 conduction. Most patients with a broad complex tachycardia and previous history of MI will have ventricular tachycardia, but not at this ventricular rate … If the ventricular rate is changed to a higher rate, the answer will be correct."

> "The classification used for endometrial hyperplasia is outdated based on WHO guidelines … "

Uncertainty about professional content: Several reviewers expressed uncertainty about the accuracy of the content. This included uncertainty about whether the content was in line with updated guidelines or current literature, or whether the stem and option were realistic or had missing information, such as: "[I am] unsure whether 15% is right. I have found 20–30% in the literature," and "I am unsure whether the correct option complies with national guidelines … "

3.2.3. Technical flaws

Item writing flaws that related to language or structure of the items, here termed technical flaws, were often commented by reviewers.

Spelling and language: In some items, spelling mistakes and typographical errors were pointed out. A few reviewers commented that abbreviations, eponyms and dialect words should be avoided for clarity, as in this case: "Eponyms such as Conn's syndrome should be avoided." Other comments related to imprecise wording in the stem and question, long and information dense options, response options with lists of shuffled words, and negative wording. These comments can be summarised in the following quote: "Some students might answer this incorrectly because they are confused by the question."

Structure: Reviewers commented that some items had superfluous stems or did not have stems at all. Other comments pertained to clues as to which option was correct, for example grammatical clues or longest option being correct.

Explanation of correct option: Many items lacked an explanation of the correct option. Where an explanation was provided, reviewers often requested that explanations be more in-depth or to a larger degree explain certain concepts of the item. Some reviewers commented that distractors should also have an explanation of what did not make this the best option, thereby increasing the learning potential of the item.

4. Discussion

In this study, we implemented external double-blinded peer review of MCQs for in-house examinations. Results showed that of the 1353 items reviewed, 20% of items were either rejected or judged as needing revision by reviewers. Subsequently, changes were made to 40% of disapproved items, which constitutes almost 10% of the total number of MCQs that were reviewed. Relevance and accuracy of item content, as well as technical item writing flaws, were the three main reasons for disapproving items for use.

The double-blinded peer review system ensures that review is not biased by gender, affiliation or seniority, and that reviews can be honest and without fear of retaliation (Shaw, 2015). In higher education, a limitation of internal review can be a reluctance to criticise colleagues, especially when the individual writing that item is considered an expert on the topic (Jozefowicz et al., 2002). In this study, we chose to use junior doctors and general practitioners as reviewers. We asked that they review items on the basis of being clinicians, thereby providing a practitioner's perspective which draws on experience and tacit knowledge. There may be advantages of using junior doctors and general practitioners rather than content experts. Their generalist perspective may contrast that of experienced academic staff responsible for teaching and developing test items, who may overestimate the importance of learning the details in their field (Mcleod & Steinert, 2015). Indeed, standard setting studies have demonstrated that expert judges tend to set unrealistically high passing scores, which could indicate that they expect too much of novice learners (Kane, Crooks, & Cohen, 1999). However, by allowing item writers to decide whether to change the item following review, experts remained responsible for item content. In this way, reviews provided input on item content, rather than a final say. Assessment authenticity and validity may increase when content is informed both by experienced academic staff and front line clinicians' perspective on what is important to know (American Board of Internal Medicine, 2016).

Content relevance emerged as one of the main themes in reviewers' comments on disapproved items. Reviewers commented that many items were too difficult to be appropriate in an undergraduate setting, demanding knowledge that was too in-depth or that concerned rare conditions. Another aspect was the importance of the content tested, with reviewers commenting that items were irrelevant for clinical practice or that items only tested recall of knowledge, as opposed to application and reasoning. This finding is in line with Koens, Rademakers, and Ten Cate (2005) who found that although test items were designed by item writers to assess core medical knowledge, many were judged as testing non-core knowledge by clinicians. The occurrence of test items of low relevance may reflect differing views on what constitutes relevance (Janssen-Brandt, Muijtjens, & Sluijsmans, 2017; Koens, Custers, & Ten Cate, 2006; Koens et al., 2005). In order to reach a more consistent and accurate interpretation of the relevance, Janssen-Brandt et al. (2017) suggest using a rubric of five criteria: 1) medical knowledge (requires study and understanding of medicine), 2) ready knowledge (cannot be looked up quickly), 3) incidence in practice (how often knowledge is needed in practice) 4) prevalence or high-risk (needed for high-prevalence or high-risk situations), and 5) foundation in the medical curriculum. The link between test content and professional practice is especially important in professional higher education in order to make sound inferences about licensing and certification, and irrelevant content is therefore a major threat to test validity (Downing, 2002; Norcini & Grosso, 1998).

Another main theme from reviewers' comments was content accuracy. While most comments related to lack of sufficient information, leaving items too imprecise to identify one best option, others related to errors in the content. These ranged from uncertainties about the accuracy of the professional content to content errors, such as items that were based on outdated guidelines or classification systems. The rapid growth of medical knowledge poses a challenge to deciding and updating curriculum and assessment content. Additionally, if items contain information that is

medically inaccurate, the testing effect may increase the likelihood of students remembering erroneous information (Rohrer & Pashler, 2010). This may be especially relevant when storing items in an item bank for reuse on later examinations, running the risk of items becoming outdated in a short period of time (Sadaf, Khan, & Ali, 2012).

The last main theme that emerged from the peer review encompass technical aspects of MCQs, such as errors relating to structure, clues, language and spelling, and items missing an explanation of the correct option. Poorly written MCQs may be falsely more difficult or easy, and be differentially confusing for different subgroups of students, thereby decreasing the fairness of the assessment (Downing, 2002; McCoubrie, 2004; Tarrant & Ware, 2008). Although important, technical aspects could probably be equally or better reviewed by strengthening in-house review. By reducing the frequency of IWFs, in-house peer review has been shown to improve psychometric properties of examinations (Abozaid et al., 2017; Malau-Aduli & Zimitat, 2012; Wallach et al., 2006).

Feasibility of the peer review process was important for implementation in an in-house setting, with fewer financial and staff resources available. The number of items reviewed in one year is approximately the number of MCQs needed yearly for examinations and reassessment in our programme. A small annual work load per reviewer and an IT solution that enabled reviewers to work from home were essential for recruiting reviewers as they received no financial compensation. Furthermore, the IT solution (our web-based item bank) supported the entire review process, including distribution of items to reviewers, review, distributing reviewer comments to item writers and editing items, thereby minimising administrative costs.

The novelty of this study is the implementation of quality assurance of MCQs that is new to an in-house setting. External review could be suitable for other professional higher education programmes, where front-line practitioners can provide useful input on assessment content. In this study, external reviewers received limited training in item writing guidelines and were asked to assess items on the basis of being clinicians. The qualitative data give insight into why junior doctors and general practitioners thought many items should be revised or not be used in examinations. In order to see whether external peer review can affect measures such as reliability and item discrimination, the authors suggest future studies should look into psychometric effects, in addition to its long-term effects on item quality.

5. Conclusions

This study showed that external, double-blinded peer review of MCQs can be implemented for in-house examinations. Approximately one in five items were rejected or judged as needing revision and of these, two in five items were later changed by the item writer. There were three main reasons for not approving items for use: (i) Relevance of item content, (ii) accuracy of item content, and (iii) technical item writing flaws. Using front-line practitioners to review examination content may lead to more relevant and accurate items, increasing the validity of summative assessments.

Acknowledgements
The authors would like to thank all peer reviewers in hospitals and GP surgeries across Norway.

Funding
This work was supported by NTNU Teaching Excellence.

Author details
Susanne Skjervold Smeby[1]
E-mail: susanne.s.smeby@ntnu.no
Børge Lillebo[2,3]
Vidar Gynnild[4]
ORCID ID: http://orcid.org/0000-0001-7589-6057
Eivind Samstad[1,5]
Rune Standal[6]
Heidi Knobel[1,7]
Anne Vik[8,9]
Tobias S. Slørdahl[1,10]
ORCID ID: http://orcid.org/0000-0001-7488-4863
[1] Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.
[2] Department of Circulation and Medical Imaging, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.
[3] Clinic of Medicine and Rehabilitation, Levanger Hospital, Nord-Trøndelag Hospital Trust, Levanger, Norway.
[4] Department of Education and Lifelong Learning, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.

[5] Clinic of Medicine and Rehabilitation, Ålesund Hospital, Møre og Romsdal Hospital Trust, Ålesund, Norway.
[6] Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.
[7] Department of Oncology, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway..
[8] Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.
[9] Department of Neurosurgery, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway.
[10]Department of Haematology, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway.

Declaration of interest statement
The authors declare that they have no competing interests.

Citation information
Cite this article as: Improving assessment quality in professional higher education: Could external peer review of items be the answer?, Susanne Skjervold Smeby, Børge Lillebo, Vidar Gynnild, Eivind Samstad, Rune Standal, Heidi Knobel, Anne Vik & Tobias S. Slørdahl, *Cogent Medicine* (2019), 6: 1659746.

References

Abdulghani, H. M., Ahmad, F., Irshad, M., Khalil, M. S., Al-Shaikh, G. K., Syed, S., … Haque, S. (2015). Faculty development programs improve the quality of multiple choice questions items' writing. *Scientific Reports*, *5*. doi:10.1038/srep09556

Abozaid, H., Park, Y. S., & Tekian, A. (2017). Peer review improves psychometric characteristics of multiple choice questions. *Medical Teacher*, *39*(sup1), S50–S54. doi:10.1080/0142159X.2016.1254743

American Board of Internal Medicine. (2016). More physicians invited to rate exam topics by relevance in practice and to help set exam standard. Retrieved from https://www.abim.org/news/abim-engages-physicians-on-updates-potential-changes-to-moc-assessments.aspx.

Baartman, L. K., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programs. *Studies in educational evaluation*, *32*(2), 153–170.

Case, S. M., & Swanson, D. B. (1998). *Constructing written test questions for the basic and clinical sciences*. Philadelphia: National Board of Medical Examiners.

Downing, S. M. (2002). Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education*, *7*(3), 235–241.

Downing, S. M. (2005). 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. *Advances in Health Sciences Education*, *10*(2), 133–143. doi:10.1007/s10459-004-4019-5

Downing, S. M., & Yudkowsky, R. (2009). *Assessment in health professions education*. New York: Routledge.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309–333. doi:10.1207/S15324818AME1503_5

Janssen-Brandt, X. M., Muijtjens, A. M., & Sluijsmans, D. M. (2017). Toward a better judgment of item relevance in progress testing. *BMC medical education*, *17*(1), 151.

Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in-house médical school examinations. *Academic Medicine*, *77*(2), 156–161. doi:10.1097/00001888-200202000-00016

Kane, M., Crooks, T., & Cohen, A. (1999). Designing and evaluating standard-setting procedures for licensure and certification tests. *Advances in Health Sciences Education*, *4*(3), 195–207. doi:10.1023/A:1009849528247

Koens, F., Custers, E. J., & Ten Cate, O. T. (2006). Clinical and basic science teachers' opinions about the required depth of biomedical knowledge for medical students. *28*(3), 234–238. doi:10.1080/01421590500271183

Koens, F., Rademakers, J. J., & Ten Cate, O. T. (2005). Validation of core medical knowledge by postgraduates and specialists. *Medical teacher*, *39*(9), 911–917. doi:10.1111/j.1365-2929.2005.02246.x

Malau-Aduli, B. S., & Zimitat, C. (2012). Peer review improves the quality of MCQ examinations. *Assessment & Evaluation in Higher Education*, *37*(8), 919–931. doi:10.1080/02602938.2011.586991

Malterud, K. (2012). Systematic text condensation: A strategy for qualitative analysis. *Scandinavian Journal of Public Health*, *40*(8), 795–805. doi:10.1177/1403494812465030

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: A literature review. *Medical Teacher*, *26*(8), 709–712. doi:10.1080/01421590400013495

Mcleod, P., & Steinert, Y. (2015). Twelve tips for curriculum renewal. *Medical Teacher*, *37*(3), 232–238. doi:10.3109/0142159X.2014.932898

Melnick, D. E. (2009). Licensing examinations in North America: Is external audit valuable? *Medical Teacher*, *31*(3), 212–214.

Norcini, J., & Grosso, L. J. (1998). The generalizability of ratings of item relevance. *Applied Measurement in Education*, *11*(4), 301–309.

Rohrer, D., & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher*, *39*(5), 406–412. doi:10.3102/0013189X10374770

Sadaf, S., Khan, S., & Ali, S. K. (2012). Tips for developing a valid and reliable bank of multiple choice questions (MCQs). *Education for Health*, *25*(3), 195. doi:10.4103/1357-6283.109786

Schuwirth, L. W., & Van Der Vleuten, C. P. (2004). Different written assessment methods: What can be said about their strengths and weaknesses? *Medical Education*, *38*(9), 974–979. doi:10.1111/j.1365-2929.2004.01916.x

Shaw, D. M. (2015). Blinded by the light. *EMBO Reports*, *16*(8), 894–897. doi:10.15252/embr.201540943

Swanson, D. B., & Roberts, T. E. (2016). Trends in national licensing examinations in medicine. *Medical Education*, *50*(1), 101–114. doi:10.1111/medu.12810

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, *42*(2), 198–206. doi:10.1111/j.1365-2923.2007.02957.x

Van der Vleuten, C., Schuwirth, L., Scheele, F., Driessen, E., & Hodges, B. (2010). The assessment of professional competence: Building blocks for theory development. *Best Practice & Research Clinical Obstetrics & Gynaecology*, *24*(6), 703–719. doi:10.1016/j.bpobgyn.2010.04.001

Verhoeven, B., Verwijnen, G., Scherpbier, A., & Schuwirth, L. (1999). Quality assurance in test construction: The approach of a multidisciplinary central test committee/ Commentary. *Education for Health*, *12*(1), 49.

Wallach, P. M., Crespo, L., Holtzman, K., Galbraith, R., & Swanson, D. (2006). Use of a committee review process to improve the quality of course examinations. *Advances in Health Sciences Education*, *11*(1), 61–68. doi:10.1007/s10459-004-7515-8

Ware, J., & Vik, T. (2009). Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations. *Medical Teacher*, *31*(3), 238–243. doi:10.1080/01421590802155597

Cogent Medicine **(ISSN: 2331-205X) is published by Cogent OA, part of Taylor & Francis Group.**