Established in 1871

Swiss Medical Weekly

Formerly: Schweizerische Medizinische Wochenschrift

An open access, online journal • www.smw.ch

Original article | Published 08 March 2022 | doi:10.4414/SMW.2022.w30118
Cite this as: Swiss Med Wkly. 2022;152:w30118

Swiss general internal medicine board examination: quantitative effects of publicly available and unavailable questions on question difficulty and test performance

Petra Ferrari Pedrini[a*], Christoph Berendonk[bc*], Anne Ehle Roussy[de], Luca Gabutti[df], Thomas Hugentobler[dg], Lilian Küng[c], Franco Muggli[dh], Florian Neubauer[cd], Simon Ritter[di], Alexandre Ronga[djk], Andreas Rothenbühler[dl], Monique Savopol[dm], Hansueli Späth[dn], Daniel Stricker[c], Daniel Widmer[bko], Ulrich Stoller[pq], Jürg Hans Beer[adr**]**

[a] Department of Internal Medicine, Cantonal Hospital Baden, Switzerland
[b] Former member of the exam board of the SSGIM
[c] Institute for Medical Education, University of Bern, Switzerland
[d] Member of the exam board of the SSGIM
[e] Private Practice General Internal Medicine, Bellevue, Switzerland
[f] Department of Internal Medicine and Nephrology, Bellinzona Regional Hospital, Bellinzona, Switzerland
[g] Private Practice General Internal Medicine, Amriswil, Switzerland
[h] Outpatient's Medical Clinic, Vezia, Switzerland
[i] Department of Internal Medicine, Spitalverbund Appenzell Ausserrhoden, Switzerland
[j] Centre médical de La Source, Lausanne, Switzerland
[k] Département de Médecine de Famille, Unisanté, Lausanne, Switzerland
[l] Private Practice General Internal Medicine, Lyss, Switzerland
[m] Private Practice General Internal Medicine, Fribourg, Switzerland
[n] Practice at the Wolfgraben, Internal Medicine, Langnau am Albis, Switzerland
[o] General Internal Medicine, Lausanne, Switzerland
[p] Medical Centre Thun am Bahnhof, Thun, Switzerland
[q] President of the exam board of the SSGIM
[r] Centre for Molecular Cardiology, University of Zurich, Schlieren, Switzerland
[*] Co-first authors
[**] Co-last authors

Summary

BACKGROUND: Formerly, a substantial number of the 120 multiple-choice questions of the Swiss Society of General Internal Medicine (SSGIM) board examination were derived from publicly available MKSAP questions (Medical Knowledge Self-Assessment Program®). The possibility to memorise publicly available questions may unduly influence the candidates' examination performance. Therefore, the examination board raised concerns that the examination did not meet the objective of evaluating the application of knowledge. The society decided to develop new, "Helvetic" questions to improve the examination. The aim of the present study was to quantitatively assess the degree of difficulty of the Helvetic questions (HQ) compared with publicly available and unavailable MKSAP questions and to investigate whether the degree of difficulty of MKSAP questions changed over time as their status changed from publicly available to unavailable.

METHODS: The November 2019 examination consisted of 40 Helvetic questions, 40 publicly available questions from MKSAP edition 17 (MKSAP-17) and 40 questions from MKSAP-15/16, which were no longer publicly available at the time of the examination. An one factorial univariate analysis of variance (ANOVA) examined question difficulty (lower values mean higher difficulty) between these three question sets. A repeated ANOVA compared the difficulty of MKSAP-15/16 questions in the November 2019 examination with the difficulty of the exact same questions from former examinations, when these questions belonged to the publicly available MKSAP edition. The publicly available MKSAP-17 and the publicly unavailable Helvetic questions served as control.

RESULTS: The analysis of the November 2019 exam showed a significant difference in average item difficulty between Helvetic and MKSAP-17 questions (71% vs 86%, p <0.001) and between MKSAP-15/16 and MKSAP-17 questions (70% vs 86%, p <0.001). There was no significant difference in item difficulty between Helvetic and MKSAP-15/16 questions (71% vs 70%, p = 0.993). The repeated measures ANOVA on question use and the three question categories showed a significant interaction (p <0.001, partial eta-squared = 0.422). The change in the

Correspondence:
Prof. Jürg Hans Beer, MD, FACP.
Senior consultant
Leiter Gerinnungssprechstunde
Cantonal Hospital Baden
Im Ergel 1
CH-5404 Baden
hansjuerg.beer[at]ksb.ch

availability of MKSAP-15/16 questions had a strong effect on difficulty. Questions became on average 21.9% more difficult when they were no longer publicly available. In contrast, the difficulty of the MKSAP-17 and Helvetic questions did not change significantly across administrations.

DISCUSSION: This study provides the quantitative evidence that the public availability of questions has a decisive influence on question difficulty and thus on SSGIM board examination performance. Reducing the number of publicly available questions in the examination by introducing confidential, high-quality Helvetic questions contributes to the validity of the board examination by addressing higher order cognitive skills and making rote-learning strategies less effective.

Background

In most European countries, some form of certifying examination is required to complete the skills evaluation and obtain the title of specialist in internal medicine [1]. Like the postgraduate education itself, these examinations vary considerably in format and content. Table 1 gives an overview of the examination formats used in some selected countries. There is an unmet need for a standardised, fair and reproducible examination which assures a predefined competency level. In spite of the influence these examinations have on physician trainees, there are relatively few reports on validation of assessment for postgraduate medical certification. In their systematic review on this topic, Hutchinson et al. concluded that rigour and transparency in the postgraduate assessment process should be reflected in publications [2]. Our paper exemplifies how the analysis of item difficulty can contribute to accuracy in assessment and how this was used to improve the quality of the Swiss board examination in general internal medicine.

The specialist degree in general internal medicine has existed in Switzerland only since 2013, when the Societies of General Medicine and Internal Medicine merged. At that time, it was decided that the board examination had to be a written high-stake assessment that should primarily test applied knowledge, clinical reasoning skills, the identification of key problems, the generation of a diagnosis and decision upon a management plan. After several attempts with combinations of multiple-choice, short answer [4] and script concordance test questions [5, 6], an format consisting of just multiple-choice questions was introduced in 2017. Questions are in English to unify the examination and avoid the translation into three different languages (German, French and Italian). In agreement with the American College of Physicians, the Medical Knowledge Self-Assessment Program (MKSAP) served as a primary question pool. MKSAP is a comprehensive learning management system that has a long-standing tradition as an excellent resource for American board examination preparation [7]. An MKSAP edition includes a theoretical part and a collection of more than one thousand multiple-choice questions for self-assessment. A new edition is released every three years.

In 2017, the Swiss general internal medicine board examination consisted of 120 multiple-choice questions based on MKSAP. To some extent, the questions were part of the current MKSAP-17 edition, which was available to candidates for examination preparation. Another portion derived from earlier editions, which were no longer publicly available to candidates. It was obvious that the use of publicly available questions in the examination would encourage memorising approaches as part of the preparation, but the extent of this effect on performance was controversial. In addition to the issue of memorising, there was the need to supplement the examination with questions that reflect the specific circumstances of the Swiss healthcare system. To address this issue the SSGIM decided to develop specific Helvetic questions (HQ) and introduce them into the board examination with the goal to improve its validity.

The goal of this analysis was to measure the effect of memorising publicly available MKSAP questions and to demonstrate the validity of HQ in the SSGIM board examination. The specific aims of the present study were (1) to quantitatively assess the difficulty of the Helvetic questions compared with publicly available and unavailable MKSAP questions and their performance over time and (2) to investigate the difficulty of MKSAP questions as they changed from publicly available to unavailable.

Methods

The November 2019 examination, mandatory for gaining the title Specialist in General Internal Medicine, consisted of 40 Helvetic questions, 40 publicly available MKSAP-17 questions, and 40 MKSAP-15/16 questions, which were not publicly available at the time of the examination. In a first step, item difficulty (primary outcome) in the Novem-

Table 1:
Comparison of the board examinations for general internal medicine in different countries.

	Austria	France	Germany	Italy	Switzerland	United Kingdom	USA
Responsibility for organising the exams	«Akademie der Ärzte»	«Coordination Nationale des Collèges d'Einseg-nants en Médecine»	«Landesärzte-kammern»	University (without national board)	SSGAIM	Royal College of Physicians	American Board of Internal Medicine
Format	Written	No national final exam	Oral	No national final exam	Written	Written and oral	Written
	Part 1: 120 MC.				120 MC	Part 1: 200 MC.	Up to 240 MC
	Part 2: 150 MC.					Part 2: 200 MC .	
						Part 3: oral	
Time	Part 1: 4h	-	30-45 min	-	5h	Part 1: 6h	Up to 10 h
	Part 2: 5h					Part 2: 6h	
Evaluation	Pass or fail	-	Pass or fail	-	Pass or fail	Pass or fail	Pass or fail

Please note that besides Switzerland, the US, the UK and Austria have introduced multiple-choice (MC) Board examinations. Importantly, the questions of their examinations are not publicly available.

ber 2019 examination was compared for these three item categories. For this purpose individual answers were evaluated in a dichotomised manner as correct or wrong, and the aggregated proportion of correct answers was used in a completely anonymised file for further analyses. The inverse proportion of correct answers to an item in the examination defines its difficulty. The less an item is answered correctly, the more difficult it is.

The second step examined the influence of public availability on the difficulty of a question. Twenty-nine MKSAP-15/16, 25 MKSAP-17 and 16 Helvetic questions had already been used in earlier examinations. We compared the difficulty of the MKSAP-15/16 questions in the examination of November 2019 (when these questions were no longer publicly available) with the difficulty of the identical questions in the years 2014–16, when they were still publicly available to candidates. The publicly available (MKSAP-17) and unavailable (Helvetic) questions served as controls. For this retrospective observational analysis without intervention and based on properly anonymised data, informed consent or ethical approval are not necessary.

Statistical considerations

Internal consistency of the November 2019 exam was measured with Cronbach alpha. Values above 0.8 are required for summative assessments. A one-factor analysis of variance followed by Tukey test for post-hoc pairwise comparison was used to compare item difficulty of Helvetic, MKSAP-17 and the non-publicly available MKSAP-15/16 questions. A repeated analysis of variance (ANOVA) was conducted to compare item difficulties of questions of the three categories that had been used in earlier examinations. For this analysis, *question use* (first versus second examination) was defined as repeated measures factor and *question category* (MKSAP-15/16, MKSAP-17 and HQ) was defined as grouping factor. Question difficulty served as dependent variable in the analysis. It was expected that question difficulty would vary only between the first and second use for the MKSAP-15/16 questions, since their status of public availability changed between the two administrations, which was not the case for MKSAP-17 and Helvetic questions.

The sample size is given by the (large) number of approximatively 500 candidates per examination and this is a strength of the study, because this stabilised the items' difficulty (primary outcome) across administrations and thus minimised irrelevant variation with regard to the items' difficulty.

All analyses were performed using the Statistical Package for Social Sciences (SPSS), version 27 (IBM Corp., Armonk, NY). For all analyses, significant effects are reported with p-values <0.05 and effect size is reported in terms of partial eta-squared. Partial eta-squared measures the proportion of variance explained by a given variable of the total variance in which the effects of other independent variables and interactions are partialled out. As a rule of thumb, one can assume that partial eta-squared values lower than 0.07 denote a small, above 0.14 denote a large and in between a medium-sized effect [8].

Results

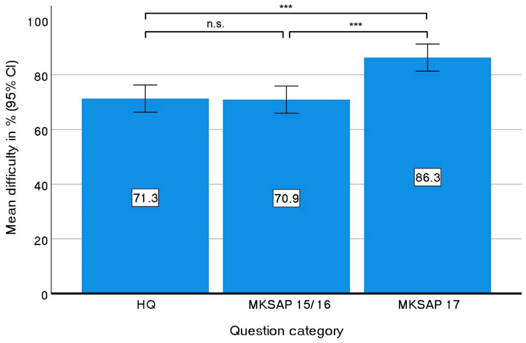
Comparison of the three question sets in the November 2019 examination

Collectively 540 candidates took the general internal medicine board examination in November 2019. Overall test reliability was 0.91 (Cronbach alpha). The average of correctly answered questions over all candidates was 76.4% ± standard deviation ± 11.4% for all questions in the examination, 71.3% ± 18.4% for Helvetic questions, 70.9% ± 19.0% for MKSAP-15/16 questions and 86.3% ± 7.5% for MKSAP-17 questions (fig. 1). Univariate ANOVA revealed significant differences between the difficulty of questions of the three different question pools ($F_{(2,117)}$ = 12.2, p <0.001, partial eta-squared = 0.173). A Tukey HSD test for post-hoc comparison of the mean values revealed a significantly lower difficulty for the MKSAP-17 questions compared with MKSAP-15/16 (p <0.001) and Helvetic questions (p <0.001), but no difference in difficulty between the publicly unavailable MKSAP-15/16 and the Helvetic questions (p = 0.993).

Quantification of difficulty of MKSAP 15/16 and MKSAP 17 questions and Helvetic questions over time

A total of 29 questions of MKSAP-15/16 were used in the board examination of November 2019, when these questions were no longer publicly available, as well as in the different examinations in 2014–16, when these questions belonged to the then most recent MKSAP edition and were publicly available to candidates. Average difficulty of these 29 questions in the November 2019 examination was 65.8%, whereas average difficulty of the identical questions in the examination sessions 2014–16 was 87.7%. The difficulty of 16 reused Helvetic questions as well as the difficulty of 25 MKSAP-17 questions did not change over time between first and second administration. A repeated ANOVA showed that this interaction question use * question category is significant ($F_{(2,67)}$ = 24.5, p <0.001, partial eta-squared = 0.422). Figure 2 depicts this interaction together with the 95% confidence intervals for the means. The difficulty changed only for the MKSAP-15/16 questions between the first and second use and not for either the

Figure 1: Average item difficulty of Helvetic, MKSAP-15/16 and MKSAP-17 questions. Comparison of question difficulty (lower levels indicate higher degree of difficulty) of questions from different sources in the November 2019 board examination. HQ and MKSAP-15/16 questions were much more difficult to answer than MKSAP-17 questions. *** p <0.001. Error bars denote the 95% confidence intervals of the means.

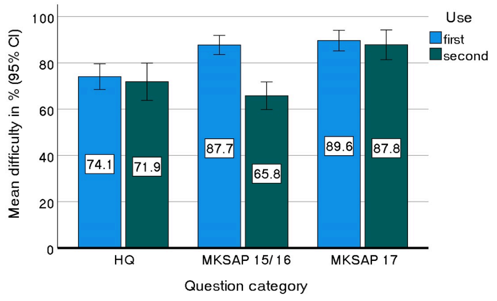


Helvetic questions or the MKSAP-17 questions. Although the main effects question use ($F_{(1,67)}$ = 35.636, p <0.001, partial eta-squared = 0.347) and question category ($F_{(2,67)}$ = 9.538, p <0.001, partial eta-squared = 0.222) also showed significance, the observed effect size was biggest for the interaction. Detailed results of the repeated ANOVA are presented in table 2.

Discussion

The fact that the availability status may influence the degree of difficulty of questions is not surprising as such. However, the extent of this change in difficulty, as documented and quantified in the present study, is striking; our results suggest that applicants intensely rehearse available questions when preparing for their examinations. The purpose of a board examination is to ensure that candidates not only have factual knowledge, but also possess adequate clinical reasoning and decision-making skills and are able to perform critical analyses. These higher order cognitive abilities are associated with deep-learning strategies [9, 10]. Surface learning, in contrast, relates to instrumental motivation for learning, reproducing content and memorising in order to pass the tests. Solving high-quality multiple-choice questions fostering application, evaluation, and analysis as self-assessment is an excellent exam preparation and has a positive impact on the exam outcome [11, 12]. Thus, our findings not only suggest that the public availability of questions is actively used by the candidates but also quantifies the impressive and unexpected size of this effect. These findings are of particular relevance because the effect of using publicly available questions in summative examinations on test performance has been scarcely studied to date in medical or other relevant educational literature. Other countries such as the US and the UK have dedicated questions writers [13, 14], who create new questions and evaluate their performance after each use, and do not use any publicly available questions. If an examination is intended to stimulate deep learning strategies, we recommend that no publicly accessible questions are used or that their number is reduced as much as possible, which is a central mission and message of the current examination board.

Figure 2: Change of question difficulty of three different question categories over time. Comparison of question difficulty of MKSAP-15/16 questions in the November 2019 board examination (when these questions were not publicly available any more) with the difficulty of the exact same questions in previous examinations (when these questions were publicly available). The two other question categories did not change availability status between first and second use: on both administrations, MKSAP-17 questions were publicly available and HQ were publicly unavailable. Error bars denote the 95% confidence intervals of the means.



The results of the present study further allow appraisal of the newly developed Helvetic questions in comparison with the other question categories used in the board examination. The importance of contextualised vignettes, which feature clinical scenarios that are relevant to daily practice, was consistently taken into account. Whether it is possible to test competencies such as clinical reasoning with multiple-choice questions depends on the stimulus format of the questions [15–17]. The stimulus specifies what the candidate is expected to answer with the question. If measuring clinical reasoning skills is in fact the desired purpose of assessment, a vignette rich in information and context is an essential feature of the questions. If questions are contextualised and ask for decisions, the thought processes invoked are vastly different from those triggered by context-free questions [18, 19]. Furthermore, quality assurance encompassed iterations of rigorous review in the item development process by content and methodology experts, professional translation from the original languages German or French into English, as well as post-hoc analysis for flawed items. This elaborate process ensured that Helvetic questions were of a quality highly comparable to MKSAP questions, which are considered a reference for quality by international experts [20]. Item difficulty of Helvetic questions was very much akin to those of publicly unavailable MKSAP; furthermore, and most importantly for the quality assessment of the examination, difficulty values remained stable over time.

The use of publicly available question also affects standard setting procedures (cut score). For licencing examinations such as the SSGIM board examination, a criterion-based passing grade is desirable [21]. Criterion-referenced assessment measures examinees' performance against a pre-

Table 2:
Detailed results of the variance analysis with within effect *question use* (first versus second examination) and between effect *question category* (MKSAP-15/16, MKSAP-17 and Helvetic questions). Within effects are the main effect *question use* and the interaction question use * *question category*. Their variances (mean sum of squares) are both tested at the error variance for the within factor mean sum of squares of Error (Use), whereas the between group effect's variance is tested at the error variance. Displayed are the resulting F-values together with their corresponding degrees of freedom, the significance of the effects and their partial eta-squared values as an estimate of the observed effect size. With a partial eta-squared of 0.422 the interaction is the largest observed effect which manifests itself in the fact that the difficulty between the first and second exam changes only for the MKSAP 15/16 questions.

	Source of variance	Typ III sum of squares	df	Mean sum of squares	F	p-value	Partial eta-squared
Within effects	Use	2457.7	1	2457.7	35.636	<0.001	0.347
	Use * Category	3380.0	2	1690.0	24.504	<0.001	0.422
	Error (use)	4620.8	67	69.0			
Between effects	Category	6004.0	2	3002.0	9.538	<0.001	0.222
	Error	21088.4	67	314.8			

defined competency level that each examinee is expected to achieve. In a criterion-based method, such as Angoff, the examination committee discusses and estimates the proportion of the minimally competent trainees who would respond correctly to each question [22]. For the committee, however, it is largely impossible to assess the "true" difficulty level of a publicly available question, since the difficulty depends less on the inherent difficulty level of the question itself and more on its availability status. This implies that publicly available questions are ill-suited for calculating a meaningful passing grade in a licensing examination, despite their accepted and well-respected quality.

The strengths of our study were (a) a detailed and sound analysis of the effects of publicly available questions on the national general internal medicine board examination, (b) the comparison of the outcome with the same source of questions that are not publicly available, (c) the comparison with the newly generated and validated Helvetic questions, and (d) the evidence of their stable performance over time. Given the large number of candidates each year (almost 1000), the data and the analysis of the high quality of the questions are of utmost interest to the candidates, the examination boards and clinical teachers. Physician trainees spend a substantial amount of their time in preparation efforts and would appreciate useful, didactic and state of the art questions, which ideally will stimulate and contribute to their clinical education.

A limitation of our study was that we did not have candidate demographics. This is an issue because item difficulty is not only a property of the item itself, but a result of the interaction between the item and the candidates' abilities [23]. Accordingly, it would have been important to determine whether the 2019 cohort differed from previous cohorts in terms of clinical experience, number of years of training, or study site. If the candidate population had changed relevantly over the years, this should also have affected the difficulty levels of the reused Helvetic questions and MKSAP-17 questions. However, this was clearly not the case. Accordingly, the changed difficulty values of the MKSAP 15/16 questions are quite obviously due to the change in availability and not to a systematically different composition of the candidate population. Nevertheless, these aspects should be re-examined in future (prospective) studies. There is no doubt that a written multiple-choice examination– despite the validated inherent clinical reasoning – will not and cannot replace bedside teaching, guided bedside examinations and professional guidance by the experienced clinician/tutor. The SSGIM educational programme requires and provides for this reason at least 5 years of postgraduate training, the completion of a logbook, an annual evaluation by the senior supervisor,four workplace-based assessments (mini clinical examination exercise [MiniCEX] or direct observation of procedural skills [DOPS]) and the publication of an article in a peer-reviewed journal. A comparable, objective, fair evaluation and a reproducible rating for such a large number of candidates in a bedside examination every year would represent a huge task and a personal and financial hurdle, which could hardly be achieved. In fact, exactly these reasons have led to the development of the multiple-choice examinations to master the complex strategy of evaluating clinical competences and skills in most countries including Switzerland (see table 1).

In conclusion, public availability of examination questions has a decisive influence on question difficulty and on test performance of the general internal medicine board examination. Reducing the number of publicly available questions is likely to reduce rote learning strategies. If confidential high-quality vignette questions that ask for decisions replace publicly available questions, it is conceivable that deep-learning strategies are promoted and allow for setting criterion-based cut scores. Our thoroughly modified approach holds the promise of actually testing the application of knowledge and of clinical reasoning skills; our results will encourage the candidates to prepare for the examination accordingly

Financial disclosures

Dr Ch Berendonk reports grants from the Swiss National Foundation of Science (100019_200831), Prof JH Beer reports grants from the Swiss National Foundation of Science (324730_182328), the Swiss Heart Foundation, the Kardio foundation, grant support and consultation fees to the institution from Bayer, Sanofi and Daichii Sankyo.

Conflicts of interest

All authors have completed and submitted the International Committee of Medical Journal Editors form for disclosure of potential conflicts of interest. There was no conflict of interest concerning this paper except that all authors are current or former members of the examination board of the Swiss Society of Internal Medicine.

Dr D. Widmer represents UEMO to the European Commission – Health Technology Assessment Regulation as Health care professional stakeholder.

References

1. Cranston M, Slee-Valentijn M, Davidson C, Lindgren S, Semple C, Palsson R; European Board of Internal Medicine Competencies Working Group. Postgraduate education in internal medicine in Europe. Eur J Intern Med. 2013 Oct;24(7):633–8. http://dx.doi.org/10.1016/j.ejim.2013.08.006. PubMed. 1879-0828
2. Hutchinson L, Aitken P, Hayes T. Are medical postgraduate certification processes valid? A systematic review of the published evidence. Med Educ. 2002 Jan;36(1):73–91. http://dx.doi.org/10.1046/j.1365-2923.2002.01120.x. PubMed. 0308-0110
3. Torre DM, Hemmer PA, Durning SJ, Dong T, Swygert K, Schreiber-Gregory D, et al. Gathering Validity Evidence on an Internal Medicine Clerkship Multistep Ex-am to Assess Medical Student Analytic Ability. Teach Learn Med. 2020 Apr;•••:1–8.1040-1334
4. Sam AH, Field SM, Collares CF, van der Vleuten CP, Wass VJ, Melville C, et al. Very-short-answer questions: reliability, discrimination and acceptability. Med Educ. 2018 Apr;52(4):447–55. http://dx.doi.org/10.1111/medu.13504. PubMed. 1365-2923
5. See KC, Tan KL, Lim TK. The script concordance test for clinical reasoning: re-examining its utility and potential weakness. Med Educ. 2014 Nov;48(11):1069–77. http://dx.doi.org/10.1111/medu.12514. PubMed. 1365-2923
6. Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. Med Educ. 2013 Dec;47(12):1175–83. http://dx.doi.org/10.1111/medu.12283. PubMed. 1365-2923
7. https://acponline.org/
8. Cohen J. Statistical Power Analysis for the Social Sciences (2nd. Edition). Hillsda-le, New Jersey: Lawrence Erlbaum Associates; 1988.
9. May W, Chung EK, Elliott D, Fisher D. The relationship between medical students’ learning approaches and performance on a summative high-stakes clinical performance examination. Med Teach. 2012;34(4):e236–41. http://dx.doi.org/10.3109/0142159X.2012.652995. PubMed. 1466-187X
10. Feeley AM, Biggerstaff DL. Exam Success at Undergraduate and Graduate-Entry Medical Schools: Is Learning Style or Learning Approach More Important? A Critical Review Exploring Links Between Academic Success, Learning Styles, and Learning Approaches Among School-Leaver Entry (“Traditional”) and Graduate-Entry (“Nontraditional”)

Medical Students. Teach Learn Med. 2015;27(3):237–44. http://dx.doi.org/10.1080/10401334.2015.1046734. PubMed. 1532-8015

11. Riggs CD, Kang S, Rennie O. Positive Impact of Multiple-Choice Question Authoring and Regular Quiz Participation on Student Learning. CBE Life Sci Educ. 2020 Jun;19(2):ar16. http://dx.doi.org/10.1187/cbe.19-09-0189. PubMed. 1931-7913
12. Jensen JL, McDaniel MA, Woodard SM, Kummer TA. Teaching to the Test…or Testing to Teach: Exams Requiring Higher Order Thinking Skills Encourage Greater Conceptual Understanding. Educ Psychol Rev. 2014;26(2):307–29. http://dx.doi.org/10.1007/s10648-013-9248-9. 1040-726X
13. www.mrcpuk.org/get-involved-examiners/question-writers
14. www.abim.org/about/exam-information/exam-development
15. Norman G, Swanson D, Case S. Conceptual and methodology issues in studies compar-ing assessment formats, issues in comparing item formats. Teach Learn Med. 1996;•••:8.1040-1334
16. Scully D. Constructing Multiple-Choice Items to Measure Higher-Order Think-ing. Pract Assess, Res Eval. 2017;22:4.1531-7714
17. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, et al. Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance. Acad Med. 2019 Jun;94(6):902–12. http://dx.doi.org/10.1097/ACM.0000000000002618. PubMed. 1938-808X
18. Schuwirth LW, Verheggen MM, van der Vleuten CP, Boshuizen HP, Dinant GJ. Vali-dation of short case-based testing using a cognitive psychological methodology. Med Educ. 2000;35:348–56. http://dx.doi.org/10.1046/j.1365-2923.2001.00771.x. PubMed. 0308-0110
19. Schuwirth LW, van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ. 2004 Sep;38(9):974–9. http://dx.doi.org/10.1111/j.1365-2929.2004.01916.x. PubMed. 0308-0110
20. McBee E, Pitkin NEB, Durning SJ, Burke MJ. Commentary: a View from the Inside-A Per-spective on How ABIM is Innovating in Response to Feedback. Eval Health Prof. 2019 Dec 23:163278719895080. doi: http://dx.doi.org/10.1177/0163278719895080. . Online ahead of print.Eval Health Prof. 2019. PMID: 31868003.
21. De Champlain AF, Book Editor(s):Tim Swanwick, Kirsty Forrest, Bridget C. O'Brien, First published: 05 October 2018, https://doi.org/http://dx.doi.org/10.1002/9781119373780.ch24.
22. Ricker K. Setting cut-scores: a critical review of the Angoff and modified Angoff methods. Alberta J Educ Res. 2006;52(1):53–6.0002-4805
23. Steven J. Durning SJ, Dong T, Artino AR, Van der Vleuten C, Holmboe E, Schuwirth L. Dual processing theory and experts' reasoning: exploring thinking on national multiple-choice questions. Perspectives on medical Education 4.2015; 168-175.