Pass Rate Set by Borderline Regression Method but not by Modified Angoff is Independent on Difficulty of Content in Objective Structured Clinical Exams

CURRENT STATUS: POSTED

Petr Waldauf
Charles University Fac Med 3

Jerome Cockings
Royal Berkshire NHS Foundation Trust

Christian Sitzwohl
Medizinische Universitat Wien

Marco Maggiorini
UniversitatsSpital Zurich

Paul Elbers
VU medisch centrum Amsterdam

Marcus Lindner
Institute for Communication and Assessment, Heidelberg

Konstantin Brass
Institute for Communication and Assessment, Heidelberg

Alexandra Núñez
Institute for Communication and Assessment, Heidelberg

Sven Ballnus
Institute for Communication and Assessment, Heidelberg

Anne Le Roy
Charles University, Fac Med 3

František Duška
Charles University Fac Med 3

frantisek.duska@lf3.cuni.cz*Corresponding Author*
ORCiD: https://orcid.org/0000-0003-1559-4078

DOI:
10.21203/rs.2.19475/v1
SUBJECT AREAS
Educational Philosophy and Theory
KEYWORDS
Objective structured clinical examination, standard setting method, borderline regression, modified Angoff method

Abstract

Background Standard setting is a method of determining the cut-off point on the scoring scale that separates the competent from the non-competent. This is a crucial feature of each exam. Pass rate should ideally be independent on the difficulty of exam content.

Methods We compared the modified Angoff method (MAM) with the borderline regression method (BRM) of standard setting in 185 candidates examined by 137 examiners in the oral part of the European Diploma in Intensive Care exam, June 2018. We then compared the effect of removal of the hardest questions on the performance of the two techniques. The exam comprised 299 items in total across 6 OSCE stations. OSCE stations were of two types; short computer based OSCE stations (3 x 12 minutes), and longer structured discussion stations based on a clinical case (3 x 25 minutes). Our focus was the effect of item difficulty on the performance of the two standard setting techniques in determining the pass mark.

Results MAM and BRM both led to similar pass rates overall for the shorter computer based 12 min OSCE stations. In the longer structured discussion 25 min stations MAM set a pass mark much higher than BRM, failing more of the candidates whose performance during the examination was judged by examiners on their global assessment as above the standard required to pass. Further analysis showed the exam items most affecting this were the more difficult items with lower discrimination; Angoff judges over-estimated the borderline candidates ability for these items. Elimination of these items led to convergence of pass marks achieved by the two methods.

Conclusion Pass mark setting by Modified Angoff Method, but not by Borderline Regression Method, is influenced by the difficulty of exam content. This has practical implication for evaluating the results of OSCE exams.

Practice Points

In OSCE exam, pass rate should be independent on the difficulty of the exam. This was achieved with setting pass mark by Borderline Regression Method, but not by Modified Angoff Method.
The presence of hard items with low discrimination is associated with an increase of the pass mark set by modified Angoff method (and in line a decrease in pass rate) as compared to borderline regression method. This is because Angoff judges overestimate the proportion of candidates giving correct answer for hard items and underestimate the proportion of candidates giving correct answer for easy items.
If Modified Angoff is to be used to set pass mark in OSCEs, we recommend post hoc elimination of items answered correctly by less than 25% of candidates.

Glossary Terms

Facility index: Facility index is the proportion of candidates who answer a test item correctly. We prefer using this term over “item difficulty” as the index ranges from 0 to 1 and the higher the number the easier the question. Indeed it may be considered counterintuitive that easy questions that are answered by more candidates have in fact higher difficulty.

Item discrimination is the degree to which students with high overall exam scores also got the particular item correct. The higher the discrimination, the better the item discriminates between the competent and the non-competent. Easy questions (high facility index) are expected to have lower discrimination than hard questions (low facility index).

Background

The European Diploma in Intensive Care (EDIC)[1] is awarded to candidates who pass a two-stage CESMA-accredited exam [2] process at the end of their intensive care medicine (ICM) training. Part one of the EDIC is a multiple-choice test (MCQ) testing core knowledge of clinical intensive care medicine. Part two is an oral examination, further testing knowledge, but with greater focus on the application and integration of that knowledge into practical decision making at the bedside that comes with experience. Since 2013, the oral part of the exam has been held twice a year simultaneously in 5 to 7 examination centers across Europe. The format has been a modified Objective Structured Clinical Examination (OSCE) [3], with all candidates being asked the same pre-defined questions with pre-specified correct answers.

This Part 2 (oral) examination comprises two sections, each of a different OSCE style; The shorter section uses computer-based questions (3 x 12 minutes) to test radiology, investigations and monitoring, for example. The longer section uses a structured discussion style based around a single clinical case (3 x 25 minutes). A candidate must pass two out of three clinical case scenarios (CCSs) and two of three computer-based stations (CBSs) in order to pass the exam. During the CBSs (12 min each) one examiner tests the candidate’s capability to recognize common patterns of clinical imaging, different curves relevant in critical care (monitor traces, ventilator screens, ECGs etc.) and biochemical abnormalities. During each CCS (25 min each) two examiners test one candidate’s ability

to make routine decisions for a typical intensive care clinical case. The OSCE will start with questions on the initial clinical scenario provided. Questions then further explore the case as it unfolds with the introduction of new information given to the candidate. The clinical cases are derived from ‘real life’ cases intended to reflect real critical care problems as closely as possible (for more details about EDIC exam see www.esicm.org)[1].

After ESICM Examinations Committee decided to stop weighing the exam items according to clinical relevance due to unstable pass mark (See Appendix Part 1 and ref [4]), a new pass mark technique had to be chosen. In turn, for the Spring 2018 examination, the pass mark was derived by two separate and parallel techniques, the modified Angoff technique with unweighted marks (otherwise similar to the previous technique), and the borderline regression method[5–8]. Data were collected and analyzed to compare the performance of these two and the effect of the harder questions. This paper reports on interesting generalizable findings discovered by this comparison.

Methods

The examination process. The oral part is organized twice a year and run over a single day simultaneously in 5-7 exam centers across Europe. Each centre typically accommodates around 36 candidates. Each candidate is exposed to 9 examiners across 3 CCSs and 3 CBSs. The examination begins for the candidate with a 30 min preparation period, during which they are given introductory vignettes to all 3 CCSs; these vignettes resemble an extract from patients’ notes at or near referral to the intensive care physician and contain information pertaining to the history, physical examination and results of laboratory tests and imaging. Following this initial 30 minutes spent on their own reviewing these three vignettes, candidates move on through the 3 CCSs and 3 CBSs.

Each CCS last 25 minutes. Initial questions pertain to the introductory vignette. Further vignettes are then provided as the case develops, containing additional information. There are 2 examiners in the CCS room; one interacts with the candidate, while the other is records the correct answers on a tablet. Once the candidate has left the room at the end, the examiners spend 5 min to agree on which answers were correctly provided, and to agree a global assessment of performance against the defined standard for the examination. The EDIC 2 (final, or part 2) examination is regarded as an exit

examination, testing knowledge and ability to apply that knowledge at a level commensurate with completion of training as an intensive care physician in Europe. The global assessment made jointly by both examiners at the end of each CCS and each CBS is an assessment of where the examiners think the candidate fell along a spectrum for that question and independent of the difficulty of that particular question. The score used comprises a 1 for clearly unsatisfactory, 2 for bare fail, 3 for borderline, 4 for clearly satisfactory, and 5 for superior performance. For any given performance on this scale, a candidate might be expected to achieve a higher number of correct answers for an easier question, or a lower number of correct answers for a harder question.

The CBS is examined by 1 examiner aided by a computer and marked in a similar way with both marks achieved for correct answers, and a global assessment recorded using the same scale as used in the CCS section (above). During 12 min exam time, the candidate is exposed to 8-12 slides containing a brief description of clinical context, an image and one or two questions. These single slide questions aim to tests candidate’s ability to recognize patterns and integrate them with the clinical information. The focus of CBSs is imaging (x-rays and computer tomography), curves (haemodynamics, ventilator and monitor screens) and laboratory or blood test results. The examiner records correct answers provided by the candidate as the exam progresses, again on an electronic tablet. The examiner then has 3 minutes after the candidate leaves the room in which to record their impression of the candidate’s general performance of on the same scale as used in CCSs.

Selection and training of examiners. The ESICM Examinations Committee prepares the exam content with the assistance from a consultant educationalist and supervises the process of selection and training of EDIC examiners in each center. Apart from mandatory training that includes observing one series of the exam, all examiners have a half-day training before every single exam and full explanation of the content of the exam. The measures taken to equalize examiners performance in different centers were described elsewhere [9].

Post-hoc evaluation of exam content. After each exam, the facility index and discrimination of each item is calculated and reviewed by the Examination Committee during a post-exam key validation meeting. Facility index (difficulty) is the proportion of candidates who gave the correct answer (low

being easy and high being hard). The range by convention is between 0 and 1. The measure of discrimination used in this exam is the point-biserial correlation (item-rest correlations[10]). This describes how the answer on the particular item correlated with the total score of whole exam station without that item. Point biserial discrimination ranges between -1.0 and 1.0. Discrimination of an exam item must always be assessed in the context of the facility index of the item as easier items, typically, are poor discriminators. Exam item performance can be assessed by facility index-discrimination plot, where problematic items appear in left lower quadrant (See Figure S4 Left in the Supplementary Appendix). Poorly discriminatory, hard items (low discrimination, low facility index) as well as those that raised concerns of the examiners during the pre-exam workshop are individually discussed and the Examination Committee makes a decision on whether or not to eliminate them; usually, around 5% of items are eliminated this way. The exam results are calculated from the remaining exam items. In order to anchor populations of candidates sitting different series of the exam, 20 to 33% of the content is re-used, whilst the rest of exam content is created *de novo* for each exam series.

Modified Angoff method (MAM) of standard setting. A MAM is an exam-centered standard setting procedure, in which the content of the test is reviewed by the expert judges ahead of the exam. In our setting the exam content was reviewed by 8 members of ESICM Examinations Committee in 2 rounds. In the first round each rater independently reviewed the content item by item. The question was: “How many percent of minimally competent candidates would give the correct answer?”. The minimally competent candidate in this context is one who possesses the minimum level of knowledge and ability to meet the standard of the exam; that is to have completed training as an intensive care physician as considered by the ESICM examinations committee, thus being be able to safely look after patients unsupervised. The experts were instructed to predict each item facility index in range of 10-90% (in accordance with Angoff’s original text, 100% is typically avoided because of the connotation it carries of perfect performance). The original Angoff method [11] was later commonly modified by adding second round of evaluation as described [12]. During the second-round, individual ratings were disclosed to the group. Items in which the raters differed by more than 40% were discussed and

the raters with extreme ratings are allowed but not forced to re-evaluate their probabilities. Predicted probabilities of the minimally competent candidate giving the correct answer for each item was taken as the mean of all the judges scores for that item after the discussion. The pass mark for each station was calculated as the sum of all these means for items in that station.

Borderline regression method (BRM) of standard setting. BRM is a modern examinee-centered standard setting procedure, where an expert's (examiner's) decisions are based on the actual performance of the examinee [6] during the course of the exam [6]. A rater evaluates student's performance at each station by completing both a checklist recording correct answers provided, as well as recording a global rating scale as an assessment of their overall performance as compared with the defined standard for the exam (1 for clearly unsatisfactory, 2 for bare fail, 3 for borderline, 4 for clearly satisfactory, and 5 for superior performance). The scores (total of correct answers for each candidate) from all examinees at each station are subsequently regressed on the global ratings, providing a linear equation. The global score representing borderline performance is then the pass mark (See Figure 1). The pass mark is taken as the score of correct answers corresponding to the borderline global rating on this regression. All candidates achieving that score or above would pass that station, irrespective of their personal global rating given by the individual examiners. In keeping with published descriptions of this method, there was no post-exam modification by the Examinations Committee.

Data processing and statistics.

Examiners entered a record of correct answers provided and their global assessments during the course of the exam via iPads with the tOSCE-App [13–16]. Data for both correct answers given and global rating assessments were relayed in real time at the end of each station to a tOSCE-Server hosted by the Institute for Communication and Assessment Research [17] for further processing. Raw data were exported from the tOSCE-Server using the Examinator [14] evaluation tool [18] and processed by STATA 15.1 StataCorp LLC, Texas USA.

Item facility index and discrimination were calculated as described above. The pass by borderline regression method for each exam station was determined by simple linear regression of individual

points scores plotted against their global assessment ratings as shown in Figures 1 and S3a.

Comparisons of the impact of hard questions on the performance of the two standard setting methods were made two steps; firstly, the two techniques were compared on the full questions set (299 questions) and then again on the 282 questions after the 17 questions removed that had been deemed to be the hardest and least discriminatory by the Examiners Committee. Secondly, the two methods were compared by ranking the questions in terms of hardness (low facility index) and then performing iterative comparisons of the techniques as progressively more and more of the harder questions were removed. These iterative calculations of pass mark and pass rate by MAM and BRM as an increasing proportion of questions from the hard end of the spectrum were removed was automated by script written in Mata language [19] (Figure 2). The comparison between the two standard setting techniques and the effect of removal of the questions at the hard end of the spectrum was explored for both CCS type questions and CBSs (Figure 3). Relative reliability G coefficients (See Appendix, Part 6) were calculated using EduG 6.1 software with Candidates/Items measurement design [20].

Results

The oral part of EDIC (the Part 2) was sat on 14th July 2018 by 185 candidates, examined by 137 examiners and in 6 centers. Most candidates were from Europe (49%), India (37%) or Middle East (14%). The exam as sat comprised 299 items. 17 (6%) were eliminated *post hoc* by the examination committee as per the process described above (15 items from CCSs and 2 items from CBSs). The details of facility index/discrimination characteristics of the deleted items are can be found in Supplementary Appendix Part 2.

A	Max points	BRM pass mark			MAM pass mark		
		Absolute	Relative	Pass rate	Absolute	Relative	Pass
CBS 1 imaging	34		20	71%	20	60%	
CBS 2 curves	26		15	53%	14	55%	
CBS 3 labs	26		15	67%	15	59%	
All CBSs				*67%*			
CCS 1	75		41	63%	45	60%	
CCS 2	77		36	56%	45	59%	
CCS 3	61		36	56%	38	62%	
All CCSs				*61%*			
Exam				50%			

B	Max points	BRM			MAM		
		Absolute	Pass mark	Pass rate	Absolute	Relative	Pass
CBS 1 imaging	33		20	71%		20	7
CBS 2 curves	25		15	53%		14	6
CBS 3 labs	26		15	67%		15	6
All CBSs				*67%*			*6*
CCS 1	69		41	63%		43	5
CCS 2	69		35	59%		40	2
CCS 3	60		36	56%		37	5
All CCSs				*62%*			*4*
Exam				51%			4

Table 1. Results of EDIC Spring 2018 series before (A – top) and after (B – bottom) elimination of 17 items. Note: BRM = borderline regression method, MAM = modified Angoff method

Comparison was made on the performance of the two standard setting techniques both for the original set of 299 questions and again for the reduced set of 282. Removal of the 17 questions made little difference to the performance of BRM (overall pass rate 50% for all 299 questions and 51% for the 282 question set), but made a substantial difference when MAM was used (30% pass rate for the 299 questions and 41% for the 282 question set (Table 1). The use of MAM without removal of the 17 questions would have resulted in a pass rate well below the lowest pass rate in EDIC history and also below the expectation of the committee and the examiners. Even after removal of the 17 questions, the pass rate remained lower by MAM than by BRM (41% vs 51%).

The CCS question types was the main factor in the different effects seen between the two techniques when the 17 questions were removed. Removing these 17 questions made little difference

to the pass mark for CBS questions with either MAM or BRM. MAM suggested an exceptionally low pass rate for CCS questions when these 17 were included, only moving towards a more usual or expected pass rate when they were excluded. The pass rate when judged by BRM varied little with or without these 17 questions and either way fell within the expected range.

The borderline regression method (BRM) requires, as described above, a contemporaneous judgement by the examiners along a global assessment scale as to where the candidate's performance fell along a pass-fail spectrum. Using this, we employed a sub-group analysis on the candidates deemed to be borderline. In this sub-group, the pass rate was close to 50% if BRM was used in both the full 299 questions and the 282 after the 17 were removed, but much lower if MAM was used to set the standard (See Table 2). Of note, 22%, 73% and 23% candidates rated as "clear pass" or "superior performance" by examiners during CCSs 1-3, respectively would have failed that station had MAM been used for standard setting.

Pass rate of borderline candidates				
	Borderline Regression Method		Modified Angoff Method	
	All 299 items	Cleaned (282 items)	All 299 items	Cleaned (28
CBS 1 (n= 34, eliminated 1)	51.9%	51.9%	51.9%	51.9%
CBS 2 (n= 26, eliminated 1)	42.3%	42.3%	51.9%	51.9%
CBS 3 (n= 26, eliminated 0)	52.9%	52.9%	52.9%	52.9%
CCS 1 (n= 75, eliminated 6)	35.3%	35.3%	11.8%	29.4%
CCS 2 (n= 77, eliminated 8)	47.4%	59.6%	3.5%	15.8%
CCS 3 (n= 61, eliminated 1)	52.0%	52.0%	36.0%	44.0%

Table 2.: Pass rates of subgroup of candidates judged as "borderline" by the examiner(s) in all exam stations before and after elimination of 17 poorly performing items.

In summary, our primary results showed that BRM produced believable or acceptable results with or without the 17 questions removed, and in both CCS and CBS type questions. In contrast, MAM produced an unacceptable and improbable pass mark, particularly on the full 299 question dataset and particularly when used on CCS type questions. We then looked more closely at CCS type questions for reasons for this.

We noticed that CCSs had much higher proportion of harder (low facility index) and poorly discriminatory items (27%) as compared to CBSs (3%) – See Supplementary Appendix Part 4 for details. We therefore hypothesized that elimination of poorly performing exam items would be

affecting standard setting by MAM, but not by BRM. In order to further evaluate this, we looked at the effect on the two methods that progressive removal of an increasing proportion of the harder items would have. We iteratively removed items in a stepwise manner starting with those with item that were harder (lowest facility index) and with lowest discrimination (See Fig.2) and calculated the impact of such operation on the hypothetical pass rates achieved by application of MAM and BRM to set the standard. As shown in Fig. 2, the pass rate set by BRM remains almost unaffected, whilst pass rate set by MAM increases and converges with the pass rate obtained by BRM. Of note, the difference between pass rates by MAM and BRM at baseline is proportional to the proportion of low-facility-low-discrimination items.

In order to explore why MAM is influenced by the presence of hard items, we looked more closely at the relation between the observed and MAM-predicted facility index of individual exam items in the subgroup of candidates marked as borderline by the examiners (on average n = 55 [30%] candidates, range 51 [28%] to 68 [37%] depending on exam station). Ideally, the intercept of such a relation should be 0 as items with facility index = 0 (nobody gave the correct answer) should have the predicted facility index of 0. However, in all exam stations the intercept was in fact positive (0.25 to 0.46). The slope, which should ideally equal 1.0 was in fact lower (0.26 to 0.59). Indeed, as shown in Figure 3, CBSs had significantly lower intercept and higher slope than CCSs, and thus were significantly ($p=0.012$) closer to the ideal line. It can be inferred that MAM leads to the same results as BRM at around the real facility index of 65%. For harder questions (lower difficulties), Angoff rates underestimate the hardness of the question (i.e. leads to overestimation the item facility index). Or, using other words, MAM overestimates the percentage of candidates who will give the correct answer for harder questions.

The hardness of an exam overall should not affect the pass rate for any given cohort. The ideal standard setting technique should take this into account, setting a higher pass mark for an easier question set and vice versa. By progressively removing an increasing percentage of the questions from the harder end of the spectrum, we were able to test the effect of the two standard setting methods on the pass mark and pass rate with progressively an easier on average question set. BRM provided a more constant pass rate as the average question set difficulty was varied, which was reassuring. MAM, however, progressively resulted in a higher pass rate as the average question set became easier; conversely, MAM made the exam harder to pass if the question set was harder on average. MAM predicted the harder questions to be easier than they really were, so a greater failure rate on these harder questions pulled the average pass rate down when they were included.

In our data, when plotting pass rate we found that if the question set hardness and discrimination was such that the pass mark was assessed at around 65%, both techniques concurred. At the point of convergence of the pass rate by both methods, the pass mark was in range 63-66%

(See Table S1 in Part 5 of Supplementary appendix for detailed results).

The intercept of the line in Fig. 3 is the predicted facility index of a hypothetical item with a real difficulty index of 0 (no candidate getting it right). For example, an extreme question with an answer that nobody could possibly guess or know. Yet, MAM predicts a percentage of correct answers (albeit low).

Discussion

The main finding of our analysis is that in the oral OSCE-type exam, pass rate set by modified Angoff method (MAM), but not by Borderline regression method (BRM), is dependent on exam content. The presence of larger proportion of harder items with low discrimination leads to arbitrary increase in the pass mark set by MAM, leading to failing candidates rated as clear passers or superior performance by the examiners. BRM is not affected by these harder nondiscriminatory items, suggesting that examiners on site are not adversely influenced by them. Of particular note and importance, the inclusion of these items distort the relationship between MAM-predicted and real difficulty of items in the subgroup of borderline candidates; this is the sub-group of candidates who will be most affected by the incorrect setting of the pass mark. These data show that harder lower discriminatory items represent a challenge to Angoff raters and therefore, in turn, a challenge to candidates.

This finding is important as MAM is still widely used for standard setting in many high-stake stakes oral exams. As opposed to written exams with options laid in front of candidates' eyes (e.g. MCQs), it is much more difficult to give the pre-specified correct answers when an open question is asked. Our analysis clearly demonstrates that Angoff judges systematically fail to identify items that are next to impossible to answer. It would require a detailed analysis of the discussion during the second round of the MAM to find out whether there was a "emperor's new clothes" phenomenon, where some of the judges are guarding their feeling of competence [21] and consciously or subconsciously deny to admit that they did not know the answer either. It has already been found that Angoff judges tend to reach this consensus at a higher cut-off score when group discussion is allowed (Wood 2006), a phenomenon that has been called group polarization [22, 23].

Harder, poorly discriminatory items were mostly found in longer (25min) CCSs as opposed to shorter (12 min) CBSs. CCSs are designed to resemble a real-life situation when a clinician spends time with

the patient, evaluates incoming information and makes decisions. It can be hypothesized that CCS type questions with realistic clinical scenarios require the presence of harder question items to be representative of common clinical cases. In which case, a standard setting method is required that can correctly handle these questions. CBS type questions on the other hand can very focused and test very discrete points of knowledge or understanding. The optimal length of OSCE station is between 5 and 15 min [3] and our data confirm the agreement of MAM with BRM for the 12 min CBSs.

Evaluation of something ideally requires a gold standard against which to compare. No such standard exists in standard setting. However, we examine the qualities required of the hypothetical gold standard, namely the ability of the method to result in pass rate that is unaffected by the difficulty of the question set for any given cohort of candidates. BRM comes out in our data as superior in this regard. MAM shows clear limitations that likely unfairly compromises candidates when the proportion of harder and poorly discriminatory questions is varied.

BRM is an extension of the borderline group method [24, 25], where the mean score of borderline group is used as the station pass mark. In contrast, BRM uses the regression line based on actual performance of all examinees [5]. BRM cannot be used in exams sat by fewer candidates (e.g. below 50), which is often the case for high-stake specialist exams. Given the relatively high number of both examinees and examiners (>100), we consider this method optimal for standard setting for EDIC exam.

One weakness of this study is that we used only 8 Angoff raters, a limited number, and who were not independent from exam content authors; 1 out of the 8 raters was always the author of the content of one respective exam station. This could have contributed to the judges' failure to identify "impossible" answers during Angoff. Indeed, generalizability of our results remains unknown and different exam setting may lead to different results. Nonetheless, whenever the Angoff method is used to establish the pass mark in oral OSCE exams, we recommend a careful analysis of the facility index and discrimination of exam items and awareness that hard questions, particularly those with limited discrimination, may skew the pass mark to the candidates disadvantage. If the proportion of hard (low-facility index), low-discrimination items is not negligible (e.g. >10%), we recommend either

recalculating pass rates after the elimination of these items, or the use of alternative pass mark setting method such as BRM, if appropriate. MAM performance can also be assessed by the analysis of the relation between predicted and real difficulty of exam items. Ideally, the regression line of this relation should have intercept 0 and slope 1. We have demonstrated (See Appendix Part 8) that there is no need to use the subgroup of borderline candidates, as the results for this subgroup and all candidates were almost identical ($R^2 > 0.95$ for all exam stations).

Conclusion

We demonstrate that in OSCE exams with open questions and pre-specified expected answers, the modified Angoff method of standard setting only leads to similar results as borderline regression method in the absence of exam hard (low facility index), low discrimination items. If these are present, the modified Angoff method sets the pass mark artificially higher. Too easy questions have the opposite effect on the pass mark setting by modified Angoff. On the contrary, borderline regression method seem to be very robust and independent on the difficulty of exam content.

List Of Abbreviations

BRM = Borderline Regression Method

BGM = Borderline Group Method

CCS = Clinical Case Scenario

CBS = Computer Based Station

EDIC = European Diploma in Intensive Care Medicine

MAM = Modified Angoff Method

MCC = Minimally Competent Candidate

OSCE = Objective structured clinical examination

Declarations

Ethics approval and consent to participate. The study was conducted in accordance with the EU Regulation 2016/679 (“GDPR”, https://gdpr-info.eu/), International Ethical Guidelines for Biomedical Research Involving Human Subjects (https://cioms.ch/revising–2002-cioms-ethical-guidelines-biomedical-research-involving-human-subjects/), and Ethics and Data Protection Guidelines issued for E. U. Horizon 2020 projects

(https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-data-protection_en.pdf). Accordingly, ethical concerns arising from the use of personal data were mitigated by anonymising them so that they no longer relate to identifiable persons. Data that no longer relate to identifiable persons, such as aggregate and statistical data, or data that have otherwise been rendered anonymous so that the data subject cannot be re-identified, are not personal data and are therefore outside the scope of data protection law (as per Article 4.1 GDPR). In turn, informed consent and Research Ethics Board opinion were deemed unnecessary.

Consent for publication: Not applicable.

Availability of data and material. All data generated or analysed during this study are included in this published article and its supplementary information files.

Competing interests: The authors declare that they have no competing interests

Funding: The work presented in this study is the result of voluntary contribution of all the authors that was not funded.

Authors' contributions. PW, CS, KB, AN and ML collected the data, PW processed the data. PW, JC, MM, PE, SB, ALR and FD contributed to the idea and concept of the paper. FD and JC wrote the first draft, to which all authors contributed. All authors have read and approved the final version of the manuscript.

Acknowledgements. We thank the ESICM office staff, namely Dominique DeBoom, Estelle Pasquier and Nikolas Stylianides for help with data acquisition and processing and Dr Reinhardt Westkämpfer for advice.

References

1. European Diploma in Intensive Care Medicine—ESICM. https://www.esicm.org/education/edic2–2/[last accessed 31–07–2019]
2. UEMS-CESMA. https://www.uems.eu/areas-of-expertise/postgraduate-training/cesma[last accessed 31–07–2019]
3. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: An historical and theoretical

perspective. Med Teach. 2013;35.

4. Sandilands D, Gotzmann A, Roy M, Zumbo BD, De Champlain A. Weighting checklist items and station components on a large-scale OSCE: Is it worth the effort. Med Teach. 2014;36:585–90.
5. Kramer A, Muijtjens A, Jansen K, Düsman H, Tan L, Van Der Vleuten C. Comparison of a rational and an empirical standard setting procedure for an OSCE. Med Educ. 2003;37:132–9.
6. Hejri SM, Jalili M, Muijtjens AMM, Van Der Vleuten CPM. Assessing the reliability of the borderline regression method as a standard setting procedure for objective structured clinical examination. J Res Med Sci. 2013;18:887–91.
7. Pell G, Fuller R, Homer M, Roberts T. How to measure the quality of the OSCE: A review of metrics AMEE guide no. 49. Med Teach. 2010;32:802–11.
8. Dwyer T, Wright S, Kulasegaram KM, Theodoropoulos J, Chahal J, Wasserstein D, et al. How to set the bar in competency-based medical education: standard setting after an Objective Structured Clinical Examination (OSCE). BMC Med Educ. 2016;16:1.
9. Waldauf P, Rubulotta F, Sitzwohl C, Elbers P, Girbes A, Saha R, et al. Factors associated with success in the oral part of the European Diploma in Intensive Care. J Intensive Care Soc. 2017;18.
10. Nunnally JC, Bernstein IH. Psychometric theory. 3rd edition. New York: McGraw-Hill; 1994.
11. Angoff W. Scales, norms, and equivalence scores. Educational Measurement. 1984;:1–139.
12. Cizek GJ. Variations on a Theme The Modified Angoff, Extended Angoff, and Yes/No Standard Setting Methods. In: Setting Performance Standards Foundations, Methods, and Innovations. 2nd edition. Routledge; 2012. p. 181–200.

13. Tablet-based OSCE: UCAN, iTunes.

14. Tablet-based OSCE: UCAN.

15. Hochlehnert A, Schultz J-H, Möltner A, Tımbıl S, Brass K, Jünger J. Electronic acquisition of OSCE performance using tablets. GMS Z Med Ausbild. 2015;32:Doc41.

16. Monteiro S, Sibbald D, Coetzee K. i-Assess: Evaluating the impact of electronic data capture for OSCE. Perspect Med Educ. 2018;:110–9.

17. Institute for Communication and Assessment Research, Heidelberg, Germany.

18. Examinator evaluation tool.

19. Gould WW. The Mata book : a book for serious programmers and those who want to be.

20. Cardinet J, Johnson S, Pini G. Applying generalizability theory using EduG. Routledge; 2011.

21. Morrison T. Emotional intelligence, emotion and social work: Context, characteristics, complications and contribution. Br J Soc Work. 2007;37:245–63.

22. Lamm H, Myers DG. Group-Induced Polarization of Attitudes and Behavior. Adv Exp Soc Psychol. 1978;11:145–95.

23. Fitzpatrick AR. Social Influences in Standard Setting: The Effects of Social Interaction on Group Judgments. Rev Educ Res. 1989;59:315–28.

24. Livingston S, Zieky M. Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. Educational Testing Service; 1982.

25. Cizek GJ, Bunch MB. Standard setting: a guide to establishing and evaluating performance standards on tests. Sage Publications; 2007.

Figures

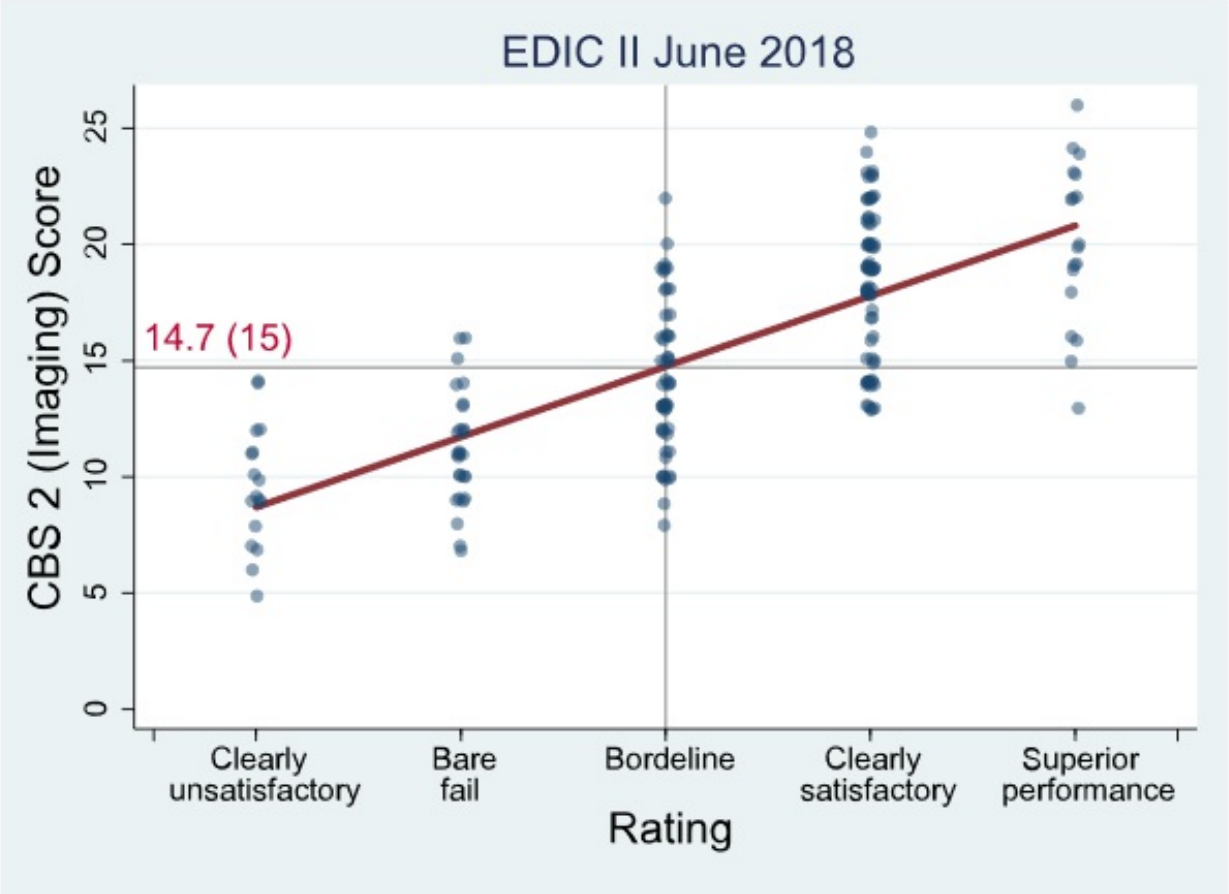

Figure 1

Example of setting the pass mark by borderline regression method. Real data from Spring 2018 Exam series. Note: CBS = Computer-Based Station. EDIC = European Diploma in Intensive Care

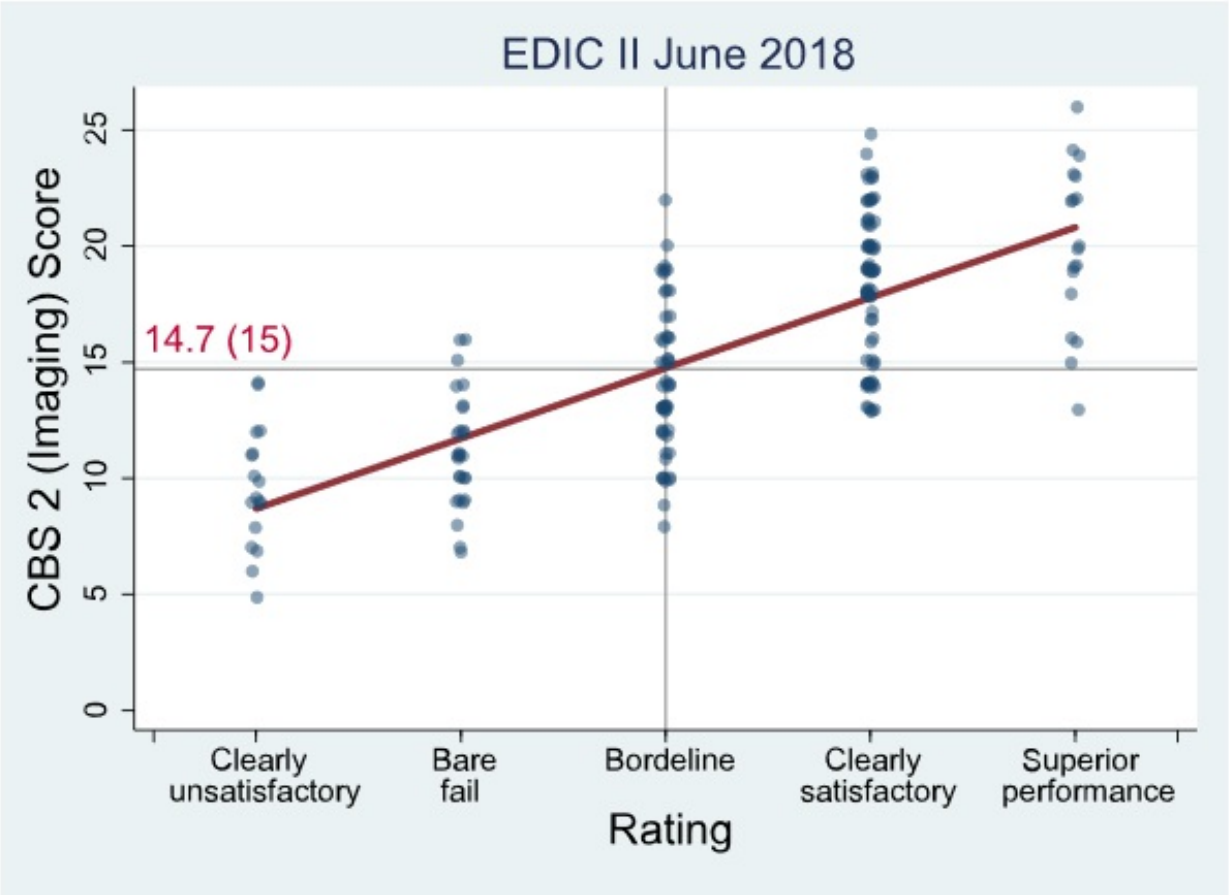

Figure 1

Example of setting the pass mark by borderline regression method. Real data from Spring 2018 Exam series. Note: CBS = Computer-Based Station. EDIC = European Diploma in Intensive Care

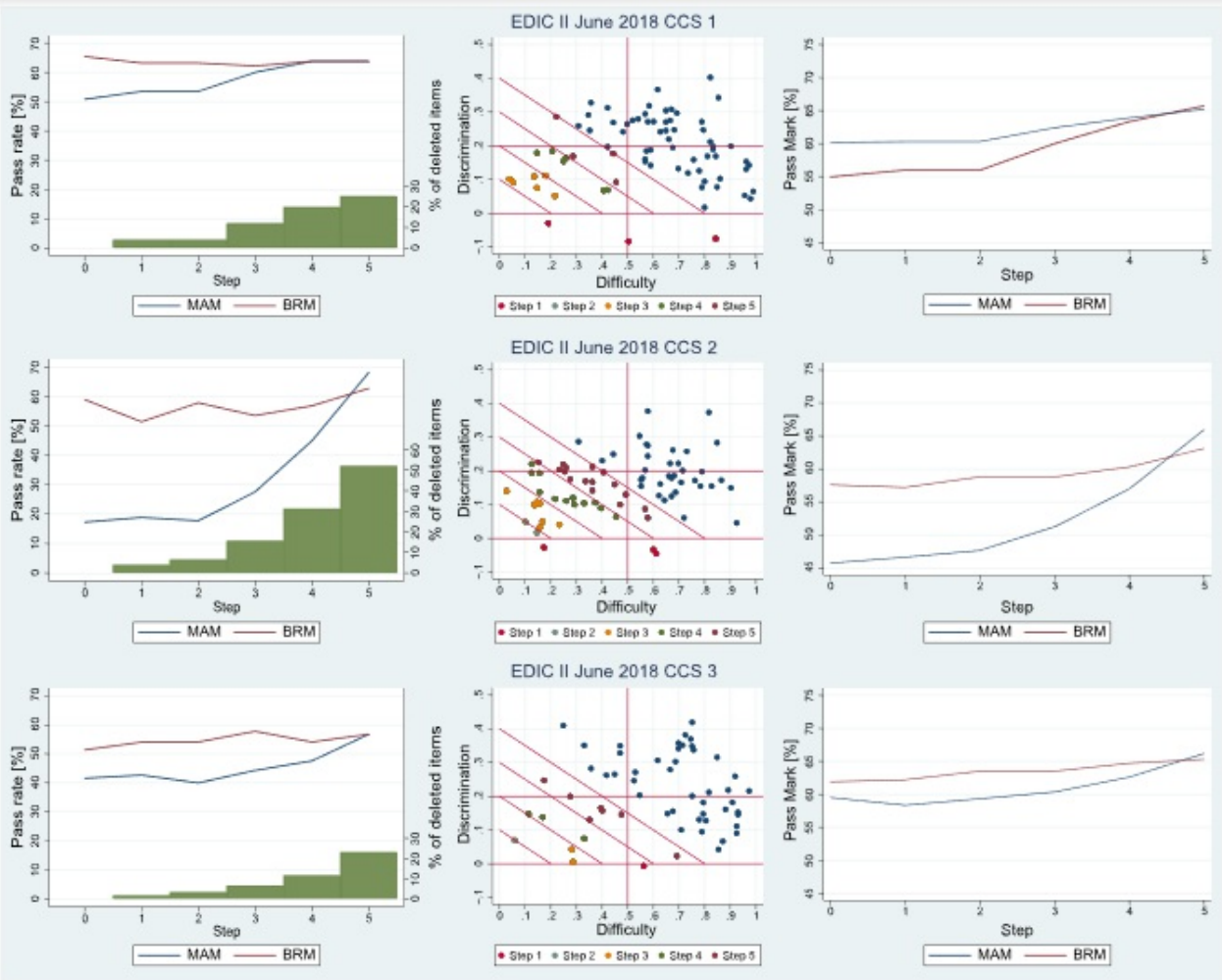

Figure 2

The effect of eliminating items with low facility indexdifficulty and low discrimination on pass rate derived from standard setting by Modified Angoff Method (MAM, blue line) or Borderline Regression Method (BRM- brown line)

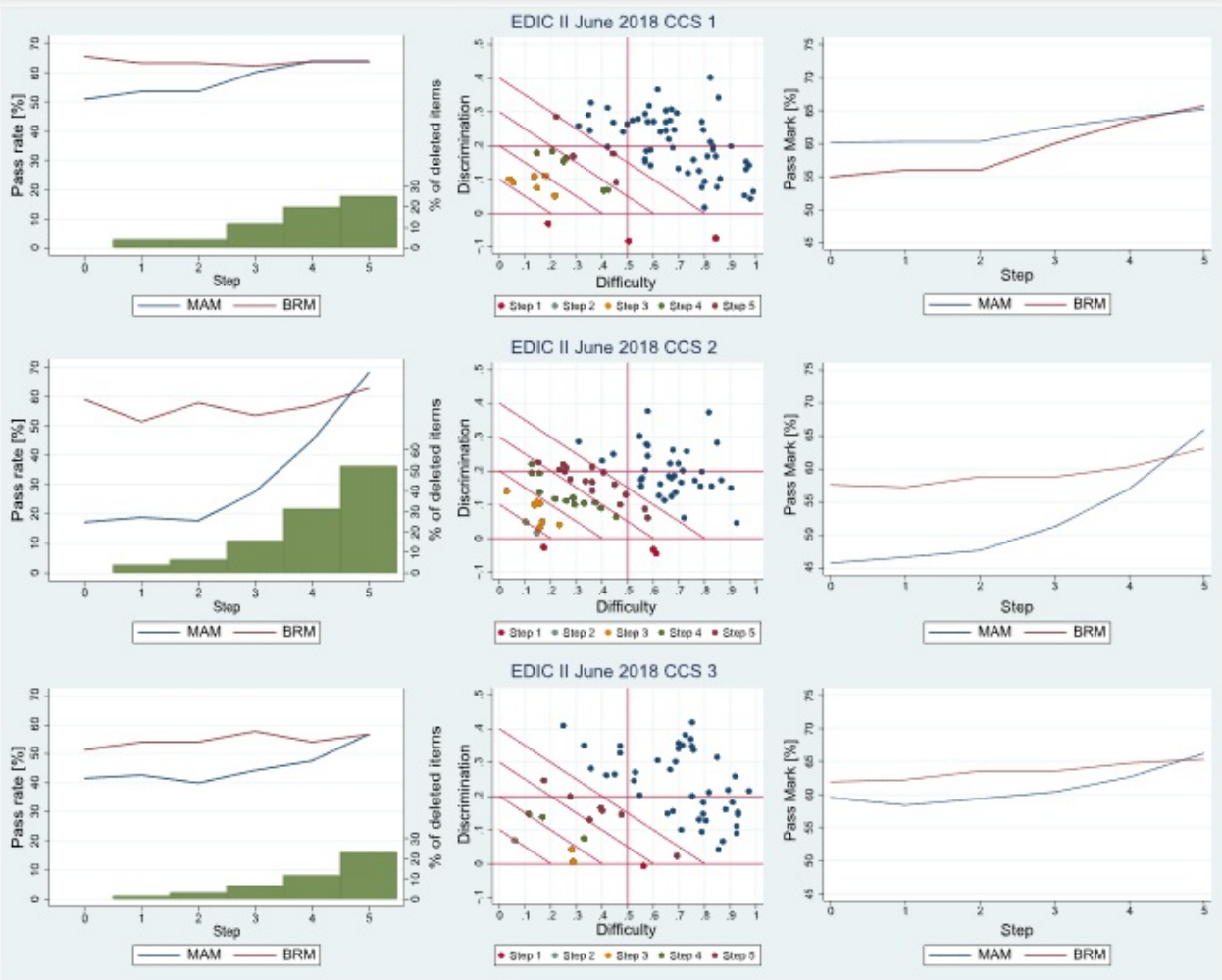

Figure 2

The effect of eliminating items with low facility indexdifficulty and low discrimination on pass rate derived from standard setting by Modified Angoff Method (MAM, blue line) or Borderline Regression Method (BRM- brown line)

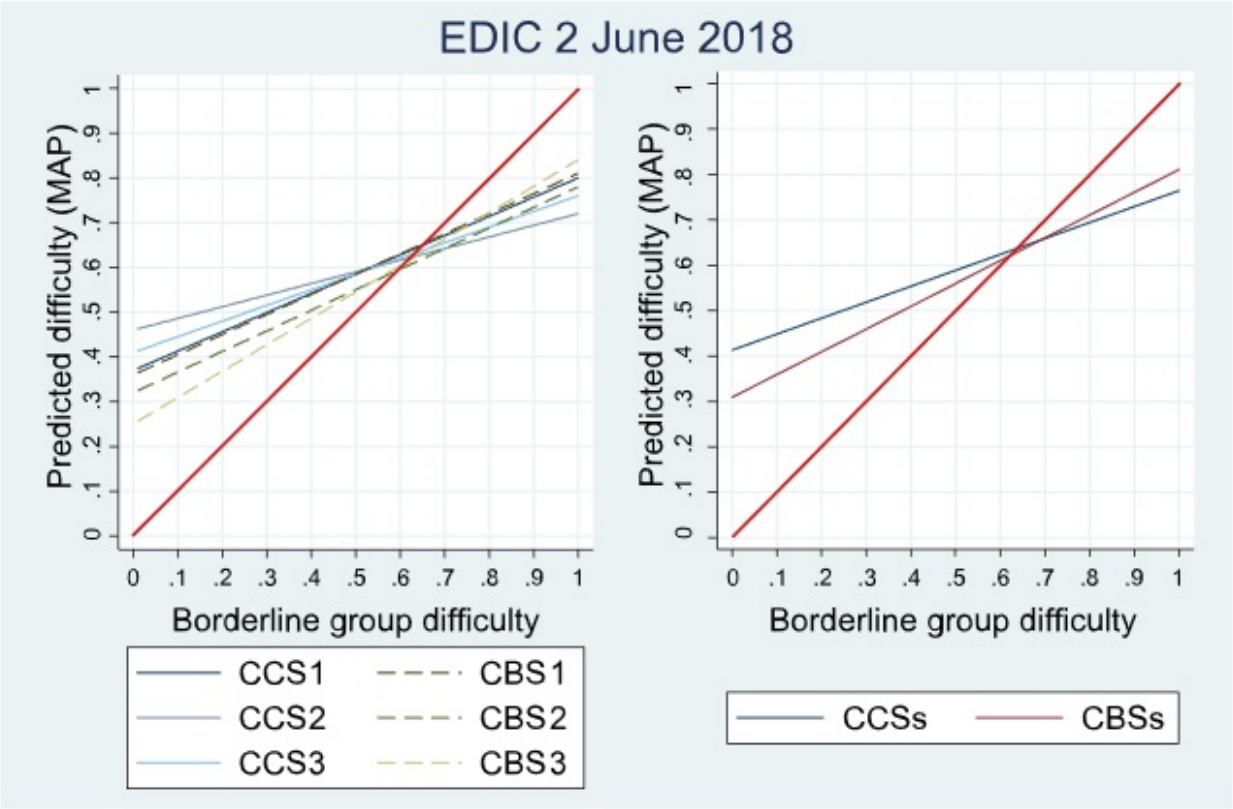

Figure 3

The relationship between MAM-predicted and real difficulties for individual (left) and grouped (right) exam domains. Red line is the ideal line with intercept 0 and slope 1.0.

Note: CBS = computer-based station; CCS = clinical case scenario.

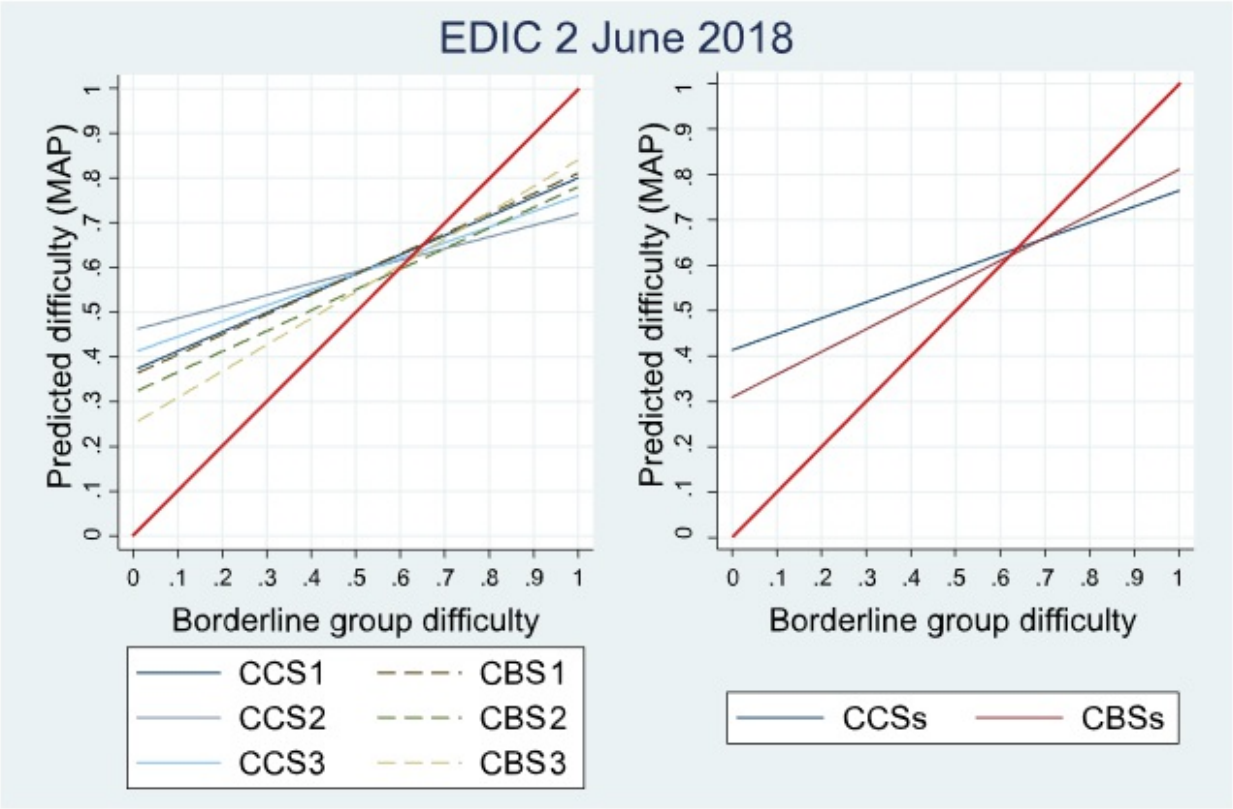

Figure 3

The relationship between MAM-predicted and real difficulties for individual (left) and grouped (right) exam domains. Red line is the ideal line with intercept 0 and slope 1.0.

Note: CBS = computer-based station; CCS = clinical case scenario.