

Original citation:

Taylor, Celia A., Gurnell, Mark, Melville, Colin R., Kluth, David C., Johnson, Neil and Wass, Val. (2017) Variation in passing standards for graduation-level knowledge items at UK medical schools. *Medical Education*. doi: 10.1111/medu.13240

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/87020>

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions. Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:

This is the peer reviewed version of the following article: Taylor, C. A., Gurnell, M., Melville, C. R., Kluth, D. C., Johnson, N. and Wass, V. (2017), Variation in passing standards for graduation-level knowledge items at UK medical schools. *Med Educ*. doi:10.1111/medu.13240, which has been published in final form at <http://dx.doi.org/10.1111/medu.13240> . This article may be used for non-commercial purposes in accordance with [Wiley Terms and Conditions for Self-Archiving](#).

A note on versions:

The version presented here may differ from the published version or, version of record, if you wish to cite this item you are advised to consult the publisher's version. Please see the 'permanent WRAP url' above for details on accessing the published version and note that access may require a subscription.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

Variability in passing standards for graduation-level knowledge questions across UK Medical Schools

C A Taylor*, Associate Professor in Quantitative Methods, University of Warwick Medical School, Coventry CV4 7AL.

M Gurnell, Clinical SubDean, University of Cambridge School of Clinical Medicine, Addenbrooke's Hospital, Cambridge, CB2 0QQ.

C R Melville, Head of Lancaster Medical School, Faculty of Health and Medicine Lancaster University, LA1 4YR.

D C Kluth, Reader in Nephrology, MRC Centre for Inflammation Research, University of Edinburgh, 47 Little France Crescent, EH16 4TJ.

N Johnson, Dean, Faculty of Health and Medicine, Lancaster University, LA1 4YW.

V Wass, Emeritus Professor of Medical Education, Faculty of Health, Keele University, Staffordshire, ST5 5BG.

*Corresponding author: celia.taylor@warwick.ac.uk, 02476 524793, University of Warwick Medical School, Coventry CV4 7AL.

ABSTRACT

Objectives: Given the absence of a common passing standard for students at UK medical schools, this paper compares independently-set standards for common “1 from 5” single-best answer (multiple choice) items used in graduation-level applied knowledge examinations and explores potential reasons for any differences.

Methods: A repeated cross-sectional study, with participating schools sent a common set of graduation-level items (55 in 2013/14; 60 in 2014/15). Items were selected against a blueprint and underwent a quality review process. Each school employed their own standard setting process for the common items. The primary outcome was the passing standard for the common items for each medical school using the Angoff or Ebel methods.

Results: 22 (of 31 invited) medical schools participated in 2013/14 (71%) and 30 (97%) in 2014/15. Schools used a mean of 49 and 53 common items in 2013/14 and 2014/5 respectively; around one-third of the items in the examinations in which they were embedded. Data from 19 (61%) and 26 (84%) schools respectively met inclusion criteria for comparison of standards. There were statistically significant differences in the passing standard set by schools in both years (effect size (f^2) 0.041 in 2013/14 and 0.218 in 2014/15, both $p < 0.001$). The inter-quartile range of standards was 5.7 percentage points in 2013/14 and 6.5 percentage points in 2014/15. There was a positive correlation between the relative standards set by schools in the two years (Pearson's $r = 0.57$, $n = 18$, $p = 0.014$). Time allowed per item, method of standard setting and timing of exam in the curriculum did not have a statistically significant impact on standards.

Conclusions: Independently-set standards for common single-best answer items used in graduation-level examinations vary between UK medical schools. Further work to examine standard setting processes in more detail is needed to help to explain this variability and develop methods to help reduce it.

INTRODUCTION

For UK-trained medical students, successful graduation from one of the 31 UK medical schools is a prerequisite for provisional registration to practise from the General Medical Council (GMC) and subsequent eligibility to enter an approved Foundation Programme Year 1 supervised training post. It is each school's responsibility to ensure that only students who meet the GMC's "Outcomes for Graduates" (1) are entitled to graduate. There is no common standard applied across all schools; each medical school designs and implements its own curriculum and multifaceted programme of assessment to comply with GMC requirements, with students being required to pass multiple and varied assessments prior to graduating. Systems of regulation and quality assurance of medical education operate in other countries, although a review of ten international systems undertaken for the GMC found a variety of different approaches to such activity (2).

The GMC's Quality Assurance Framework (3) is designed to determine whether medical schools are meeting its standards for medical education and training, including whether assessments allow a school to robustly "decide whether medical students have achieved the learning outcomes required for graduates" (4). This statement implies that each school is responsible for setting its own passing standards i.e. the minimum level of performance required to pass each assessment. The statement also implies that standards should be "absolute" (setting the level of performance required of any passing student) rather than "relative" (setting the proportion of students who will pass regardless of their performance). Three common methods of setting absolute standards are Angoff, Ebel and Hofstee, as described in Box 1.

Box 1: Common methods of absolute standard setting

Each method requires a panel of appropriately qualified individuals (for example, faculty involved with the design and delivery of teaching) to make decisions on individual test items. Decisions can be made independently and the results averaged across panel members, or followed by discussion amongst panel members to agree relevant standards (often known as a 'Modified' approach). For the Angoff and Ebel methods, the panel must also agree the definition of a 'minimally competent' student on the assessment.

Angoff (1): The panel provide the proportion of minimally competent students who would answer each item correctly. The mean standard across all items in the assessment provides the pass mark.

Ebel (2): The panel rate each item in two dimensions: 1. importance (e.g. essential, important, useful to know) and 2. difficulty (e.g. easy, moderate, challenging). This process creates a number of different categories of question (with the examples given there would be nine categories). They then agree what proportion of minimally competent students would answer each category of item correctly and the relevant standard applied to each item. The mean standard across all items in the assessment provides the pass mark.

Hofstee (3): The panel provide four ratings: 1. the minimum acceptable passing score, 2. the maximum acceptable passing score, 3. the minimum acceptable failure rate and 4. the maximum acceptable failure rate. A graph of cumulative student scores on the assessment is plotted, with score on the x-axis and the cumulative percentage of students achieving each score (or lower) on the y-axis. The four points provided by the panel are then plotted to form a rectangle on a graph and a diagonal line drawn between the top left corner of the rectangle (minimum acceptable passing score, maximum acceptable failure rate) and the bottom right corner (maximum acceptable passing score, minimum acceptable failure rate). The pass mark is the point where this diagonal line meets the students' cumulative performance curve.

There are no stipulations in the GMC Framework on *how* standards should be set so, while all schools use methods that are widely accepted as robust (for example, based on Norcini's description of credible standards (5)), there is variation between schools (6). Such variation in standard setting method and panel composition, for example, may therefore result in differences in passing standards across schools.

The External Examiner system, used across the UK Higher Education sector (7), aims to provide reassurance that passing standards are similar across all medical schools. However, evidence from External Examiners is entirely qualitative in nature and may not be sufficient to ensure comparability (6). For example, three studies

comparing the passing standards for Objective Structured Clinical Examination (OSCE) stations across a small number of medical schools identified absolute differences in pass marks at station-level of around 20% between the schools setting the highest and lowest standards (8-10). Similar between-school variation in classification of students' cognitive readiness for internship, as validated by USMLE Step 2 performance, was also found in a study involving 20 US medical schools (11). There is no similar evidence comparing passing standards for written examinations, or studies that seek to include all schools in one country, a gap we aim to address.

It is known that standard setting can influence student outcomes (12-14). If a medical school has a significantly higher passing standard than others, some students who fail at this school could be denied access to the profession if they would have passed elsewhere ('false negatives'), although such students are generally offered the opportunity to remediate and retake the examination. Conversely, if a medical school has a significantly lower passing standard than others, some students who graduate from this school may not have graduated from other schools ('false positives'). It is plausible that such false positive students might not yet be sufficiently competent to obtain provisional registration although no concerns regarding insufficient competence have been raised to date (15) (16). This may, at least in part, be due to the "Transfer of Information" process between UK schools and the Foundation Programme which helps to ensure students considered borderline passes are supported as they enter practice. Nevertheless, despite the existence of mitigating mechanisms, differences in the minimum level of performance required to graduate across schools are important and worthy of empirical study.

This paper seeks to compare the passing standards set for a common set of single-best answer applied knowledge examination items across UK medical schools in two academic years. We therefore test the null hypothesis that there is no difference in standards between schools. The research was supported by the Medical Schools Council Assessment Alliance (the Alliance) which aims to enhance the quality of assessments by sharing best practice and developing a bank of high-quality examination material which can be shared across schools (17).

METHODS

Study design

A cross-sectional study was undertaken in two academic years, July 2013 to June 2014 and July 2014 to June 2015.

The comparison of passing standards was made possible by the inclusion of a set of shared "1 from 5" single-best answer items in participating schools' graduation-level applied knowledge examinations, which were subjected to the schools' usual processes for setting standards. The items used were designed to assess students' application of clinical knowledge, rather than merely factual recall; i.e. they were "two-step" questions (Box 2). "Graduation-level" implies that a student passing the examination is deemed to have sufficient clinical

knowledge on the outcomes assessed in that examination to graduate and hence practise as a provisionally-registered doctor in the UK.

Box 2: Example two-step, 1 from 5, single best answer item used as part of the common content project

A 75 year old man becomes unresponsive in the cardiac catheter laboratory. He was admitted five minutes ago with an acute anterior myocardial infarction. He is unconscious. His pulse rate is 176 beats per minute and BP 70/40 mmHg. His airway is maintained using an oropharyngeal airway and his respiratory rate is four breaths per minute. Cardiac monitoring shows a broad complex tachycardia. Which is the most appropriate immediate treatment?

- A. Amiodarone
- B. Intubation and ventilation
- C. Lidocaine
- D. Proceed to primary coronary intervention
- E. Synchronised DC cardioversion

Answer key: E

Participants – medical schools

All UK medical schools with graduation-level examinations (N=31) were invited to participate before the beginning of each academic year and those agreeing to participate were sent, via the Alliance's secure item banking software, the same 55 (2013/14) and/or 60 (2014/15) common content items. The common content items were selected by the Alliance's *Final Clinical Review Group*, which comprises a number of clinicians from different specialties and representing different medical schools/geographical areas. The Review Group aimed to select items covering the spectrum of body systems/specialties and learning outcomes (e.g. diagnosis, investigations and management) that would be considered core for a medical graduate from any UK medical school. Prior to selection, all items underwent a two-stage process of quality review, with individual questions subjected to scrutiny and revision by an expert panel, before final revision and approval by the Final Clinical Review Group (18). Schools were not obliged to include all items in their examinations, but were given a target of 50 items. Schools were asked not to change any items unless strictly necessary. The selected common content items were then included in schools' usual graduation-level examinations.

Examinations and standard setting within medical schools

To avoid bias and maximise participation, no attempt was made to influence the processes for standard setting or the conduct of the examination at any medical school. As a result, the examinations in which the common content items were used were held at varying times during students' last two years at medical school and were not all identical in format – for example they included other item types, had

different total numbers of items, time allowed per item/mark and different marking schemas (e.g. whether negative marking was used).

Data collection

Relevant data were provided by each medical school, including the passing standard set for each common content item, as a percentage score. To check item quality and reliability, data on whether each student answered every item in the examination correctly or not were collected. To investigate potential reasons for any differences in standards, data on when in the curriculum the examination was sat, the marking schema, the method of standard setting and time allowed per item were obtained. Finally, schools were asked to report what, if any, changes had been made to each common content item, whether any items had been excluded from scoring or scored but with two answers allowed. To minimize the burden of data collection, we did not ask schools to report precise details of their standard setting processes, such as the number or composition of their panels.

Data analysis

The data from each medical school were anonymised prior to analysis. Data from examinations where negative marking was used were excluded, as were data related to items with changes to wording (beyond very minor changes such as changing “Emergency Department” to “Accident and Emergency”), those not scored or where two answers were allowed. These exclusions were made as standards set under such circumstances would not be directly comparable across schools. Aggregated datasets of standards set and common content item performance were constructed for each academic year.

The primary outcome was the overall passing standard set for the set of common content items at each school. To compare standards across schools, a general linear mixed model with repeated measures was undertaken for each year using maximum likelihood estimating in Stata v11. This method is similar to a repeated measures Analysis of Variance, which could not be used in this study because not every school used every common content item: there were some “missing” data. The model estimates the passing standard that would have been set for all of the common content items at each school. The reference school for the analysis was chosen as the school with the lowest passing standard for the common content items used. The fixed factor in the model was the schools and the random repeated factor the items. The overall effect of schools on standards was summarized using Cohen’s f^2 measure of effect size, which shows the proportion of the total variation in standards accounted for by schools. A p-value less than 0.05 for this overall comparison was considered statistically significant.

For the schools participating in the study in both years, the relationship between the estimated model coefficients (i.e. the estimated absolute difference between the standard at an individual school and the standard at the school with the lowest standard) in both years was explored using a Pearson’s correlation coefficient.

Schools that did not use item-level standards (e.g. used the Hofstee method of standard setting (19)) were excluded from the analysis because composite standards across the common content items used by an individual school could not be adjusted for differences in item usage between schools. In addition, one medical school was excluded on the basis that the standards for its examinations were being set at the time of the study by a partner school which was also participating.

To check item quality, item facility (the percentage of students answering correctly) and discrimination (the “item-rest” correlation using the Pearson’s point-biserial correlation between students’ scores on each item and their total score on all other items combined) were calculated. Our internally-agreed targets for these measures were 50-90% for facility and >0.2 for discrimination. The internal consistency of the common content items was estimated using Cronbach’s alpha, using the aggregated dataset across all schools. The Spearman Brown formula (20) was used to estimate the number of similar items required to achieve an alpha of 0.90, as a benchmark for a very high-stakes test (such as one determining entry to medical practice) (21).

We examined three potential reasons for differences in standards between schools: time allowed per item (for schools using only single-best answer/multiple choice items rather than those also using questions requiring written short answers or essays), method of standard setting (comparing the two most common broad approaches, Angoff/Modified Angoff and Ebel/Modified Ebel) and timing of exam in the curriculum. Because time per item (in seconds) and time from start of the Foundation programme (in months) were skewed we calculated the Kendall’s tau-b correlation coefficient between time allowed and the estimated mixed model coefficient for each variable. The variance in the school-level coefficients was greater amongst schools using Ebel than for those using Angoff, so mean coefficients in each group were compared using a t-test assuming unequal variances. Each analysis was undertaken separately for each year.

Sample size

A sample size calculation was undertaken in G*Power 3.1.6 (22) using a repeated measures within factors Analysis of Variance model due to difficulties in estimating sample sizes for general linear mixed models. Data from the pilot study undertaken with 19 schools in 2012/13 were used, together with an estimate of the minimum important “educationally significant” difference between passing standards at the schools setting the highest and lowest passing standards of 4 percentage points (from a survey at a meeting of the Alliance’s Reference Group, formed of representatives with responsibilities for assessment from each UK medical school). With a mean standard deviation of item-level standards within each school of 12%, this implied a small effect size (f^2) of 0.054. The correlation between repeated measures (the items) and non-sphericity correction were both estimated at 0.2. For an alpha of 0.05 and 95% power, the total sample size required was 684, or 36 items x 19 schools. Although we anticipated increased participation, some missing data were expected and it was therefore decided to use 55 items in 2013/14 and 60 items in 2014/15.

Ethical approval

Ethical approval for this study was obtained from the Science, Technology, Engineering and Mathematics Ethical Review Committee at The University of Birmingham (ERN_13-0598). All schools were requested to include a standard opt-out wording in their information for both students and standard-setters. Opt-in consent was not required by the Ethical Review Committee as all data were anonymized (i.e. the names of students and standard setters were not provided).

RESULTS

Participation, item usage, item quality and reliability

Table 1 summarises participation in the project, item usage and quality and internal consistency for each year, and details the number of schools included in the comparison of standards and reasons for exclusion. The common content items accounted for around one-third of the examinations in which they were embedded. The fall in the proportion of items meeting the target for facility (50-90%) is primarily due to the use of more difficult/challenging items in 2014/15: five items (9%) had facility below 50% in 2013/14 compared with 13 items (22%) in 2014/15. Taking the lower of the two annual values for Cronbach's alpha of 0.73 and assuming this reflects an examination of 50 items (similar to the average number used), the Spearman-Brown formula suggests that 166 similar items would be required to achieve an alpha of 0.90.

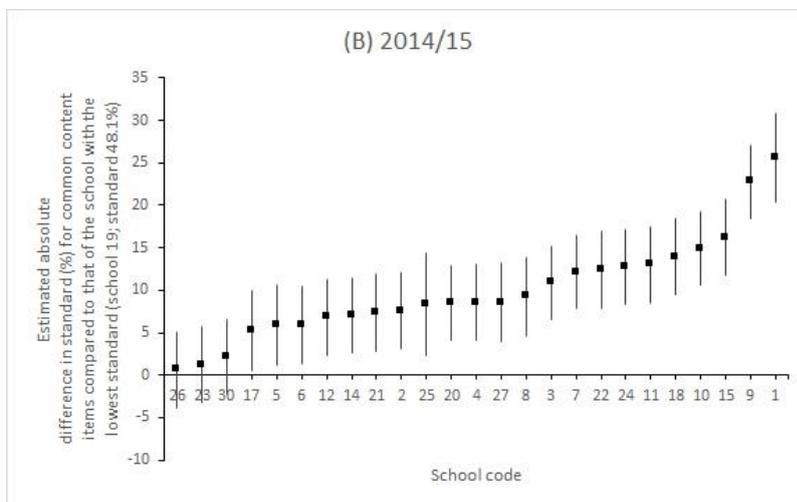
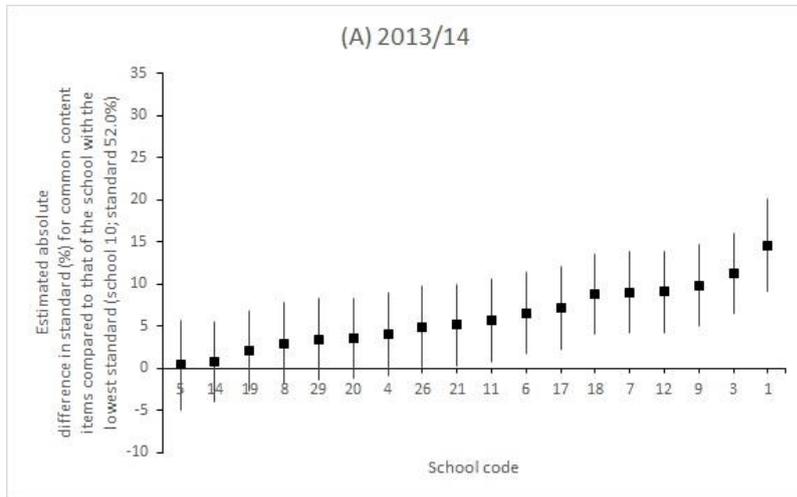
Table 1: Participation, item usage and item performance, by year

	2013/14	2014/15
Schools participating/31	22	30
Schools included in comparison of standards	19	26
<i>Exclusions:</i>		
<i>Schools not using item-level standards</i>	3	3
<i>Schools not using independently set standards</i>	0	1
Items allocated	55	60
Items used: mean, N/% (range)	49/89% (35 to 55)	53/88% (22 to 60)
Students sitting examinations using common content items	6,093	7,706
Students included in item performance analysis	6,093	7,504
<i>Exclusions:</i>		
<i>Students from 1 school using negative marking</i>	N/A	202
Item facility		
Mean (range)	0.72 (0.39 to 0.97)	0.67 (0.10 to 0.98)
N/% items with facility 50-90%	44/80%	35/58%
Item discrimination across common content items in all schools		
Mean (range)	0.21 (0.03 to 0.45)	0.17 (-0.08 to 0.35)
N/% items with discrimination >0.2	29/40.0%	20/33%
Cronbach's alpha for common content items (data from all schools combined)	0.79	0.73

Comparison of passing standards

The way in which standard setting was undertaken varied across schools, even within common frameworks such as 'Angoff' (23) or 'Ebel' (24). Figure 1 shows the estimated difference in the absolute standard set by each school, compared to that by the school with the lowest standard, for the full set of 55 or 60 common content items for (A) 2013/14 and (B) 2014/15. The lowest passing standards were 52.0% in 2013/14 and 48.1% in 2014/15. The graphs for each year are ordered by coefficient (smallest to largest) and thus the ordering of schools is different in the two graphs. Only the schools included in the analysis for each individual year are included on the graph for that year. Overall, both models were statistically significant at $p < 0.001$, with f^2 values of 0.041 and 0.218 for 2013/14 and 2014/15 respectively. The overall range of the estimated standards required to pass the full set of common content items across schools was 14.5 percentage points in 2013/14 and 25.0 percentage points in 2014/15; the considerably smaller interquartile ranges of 5.7 percentage points and 6.5 percentage points imply the presence of outliers, particularly in 2014/15.

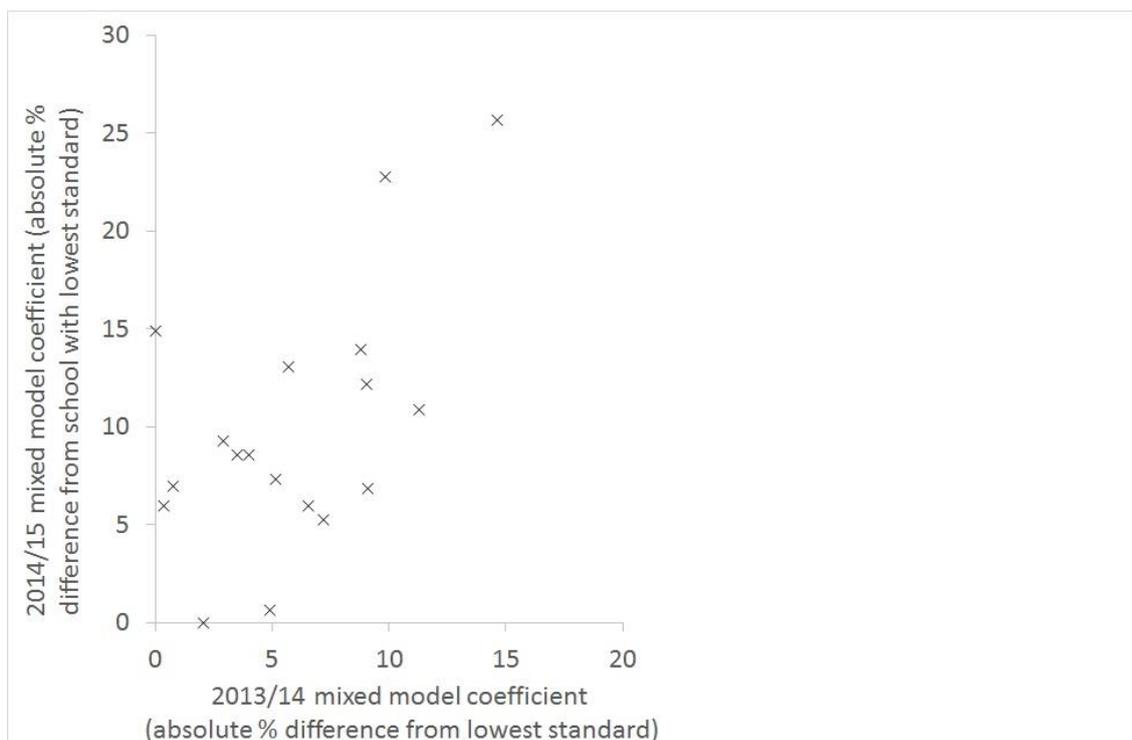
Figure 1: Estimated absolute difference in passing standards, compared to the school with the lowest standard, for each participating medical school for the full set of 55/60 common content items, for (A) 2013/14 (compared to school 10) and (B) 2014/15 (compared to school 19), with 95% confidence intervals.



Comparison of results in 2013/14 and 2014/15

Figure 2 shows a scatter diagram of the estimated mixed model coefficients for the 18 schools who participated in both years and for whom item-level standards were available. There was a reasonably strong positive correlation between the mixed model coefficients between years (Pearson's $r=0.57$, $p=0.014$), suggesting that schools setting relatively high standards in 2013/14 tended to do so in 2014/15.

Figure 2: Comparison of mixed model coefficients in 2013/14 and 2014/15.



Impact of potential mediating effects

None of the potential mediating effects, time allowed per item, method of standard setting and time from the start of the Foundation Programme when the examination was sat were explored had a statistically significant effect on the standards set by schools (Appendix Table 1).

DISCUSSION

Main findings

This study suggests that there are differences in the passing standards set for common single-best answer items used as part of graduation-level applied knowledge examinations across UK medical schools. The effect size for the differences would be considered small to medium for 2013/14 and medium to large for 2014/15 (25). The effect size for the correlation between schools' standards set in 2013/14 and 2014/15 ($r^2 = 0.32$) would be considered large (25). The time allowed per item or timing of the examination in the curriculum did not appear to influence the

standards set for the common content items, nor did the method of standard setting used when comparing all “Angoff” approaches with all “Ebel” approaches.

Interpretation of results and practical implications

Our findings regarding differences in standards build on and concur with those of studies comparing standards set for clinical examinations across a small number of UK medical schools (8-10) and that from the US (11). Identifying any variability in standards is important because UK schools are responsible for making decisions regarding their own students’ readiness to graduate and begin Foundation training. Our results are therefore likely to be of interest to a wide constituency including UK and international medical schools and their students, those responsible for postgraduate training, regulators and employers.

It is plausible that some schools are making false positive and some false negative decisions on students’ performance on the common content items, although the number of students affected would be small. As an illustration, had 3% of students passed the examination with a score within 5 percentage points of the passing standard (i.e. were 'borderline passes') at the school with the lowest standard, and a further 3% of students had scores within each subsequent 5 percentage points up to the passing standard at the school with the highest standard (i.e. 15 percentage points higher in 2013/14 and 25 percentage points higher in 2014/15), then in a school with 200 students, 18 students who passed at the school with the lowest passing standard would have failed at the school with the highest standard in 2013/14 and 30 in 2014/15 and vice-versa. Furthermore, it is also important to recognise that the common content items comprise around only one-third of longer examinations which are themselves part of multifaceted programmes of assessment which must be passed (with opportunities for remediation) prior to graduation. Hence it is important to recognise that our results do not necessarily imply that some students lack the competence to begin Foundation training.

Strengths and weaknesses of the study

This study aimed to include all UK medical school. While the standards of 84% of schools with graduation-level examinations were included in the primary analysis in 2014/15, our results for 2013/14 in particular could be affected by response bias with only 61% included. In addition, our work only includes one type of assessment included in one examination. The repeated nature of this study has ensured that conclusions are not drawn on the basis of data from only one year. The study team were also careful not to influence the standard setting process used at any school although, in line with good practice guidance provided by the GMC (26), the use of a criterion-based approach (e.g. Angoff or Ebel rather than Hofstee) was being encouraged independently of this project to enhance practice across schools.

Within the two broad groups of Angoff and Ebel, there are likely to be differences in how standard setting was undertaken and the composition of the standard setting panel. Existing evidence suggests that differences in processes “within” one method might actually lead to a greater variation in standards than the use of different

methods (27, 28). We may therefore have benefited from seeking more detail on the composition of standard setting panels in terms of exploring potential reasons for differences in standards. Furthermore, standard setting involves expert judgment and, as such, the passing standard set for the same assessment by two independent standard setting panels using the same standard setting process may differ. The repeated nature of this study which identified a positive correlation between relative standards between the two years, does however imply a level of consistency within each school.

While studies have been undertaken to examine medical school factors that might influence examination performance (for example the work of the Australian Medical Schools Council Assessment Collaboration (29), who found an effect of level of entry (graduate vs. undergraduate) and school size on student performance) this study did not seek to assess the performance of students or compare performance across medical schools. This was for two reasons: first, the Cronbach alpha levels achieved suggest that around 170 items would be required to allow reliable comparisons (with an alpha of 0.9) and second, some schools scheduled their examinations early in students' penultimate years while others were much closer to graduation, rendering meaningful comparisons of performance potentially unfair.

Conclusion

This study found differences in passing standards set by different UK medical schools for a common set of single-best answer applied knowledge examination items. There is a lack of similar work from other countries without a national examination against which performance can be 'benchmarked', so we do not know if our finding is unique to the UK. The results of our study raise questions about the use of local standard setting for high stakes assessment of readiness to undertake initial post-graduate training. Reducing the variability in standards is important as students should be required to meet the same minimum standard regardless of where they trained. However, as we did not find any statistically significant mediators of standards, we are currently unable to offer suggestions as to how the differences identified may be reduced.

One potential explanation for the differences in standards worthy of further study is that local standard setters are influenced by the ability of the students they see on a day to day basis when making standard setting decisions, such that schools with students with higher than average ability would have higher passing standards and vice versa. No evidence has previously been published suggesting that this is the case for medical school examinations, but a similar finding was reported by Livingston and Zieky when standards were set for tests of basic skills in reading and mathematics (30). Additional work to understand why differences in standards occur could explore schools' standard setting processes in detail. Such a study might identify a need for an agreed definition of the "minimally competent candidate" that could be shared across schools, or recommendations for good practice regarding the composition of standards setting panels including the range of background experiences that should be sought. These aspects could then be included in published guidance to help reduce the subjectivity inherent in standard

setting and thus facilitate the application of an applicable and acceptable methodology for standard setting across all schools. With freedom of movement for doctors across many international borders without the need for further tests of clinical competence, such methodology could even be fruitfully applied across countries as well as within them.

ACKNOWLEDGEMENTS

Professors John Cookson and Tim Lancaster were instrumental in setting up this project and their support is gratefully acknowledged. We would also like to thank school staff providing data and responding to queries, Gareth Booth and Emma Horan (both Medical Schools Council Assessment Alliance, MSCAA) for aggregating data and Veronica Davids (MSCAA) for corresponding with schools, and the standard setters and students for allowing their data to be analysed anonymously.

The authors thank the Medical Schools Council, MSCAA Board and Reference Group members for support and helpful comments, although the views expressed do not necessarily reflect those of these organisations and groups.

CAT is supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care West Midlands. MG is supported by the NIHR Cambridge Biomedical Research Centre. This paper presents independent research and the views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

DETAILS OF CONTRIBUTORS

CAT designed the study, managed the data collection process, undertook the data analysis and drafted the manuscript. MG, CM and DK led the item review and selection processes and advised on data analysis and interpretation. NJ advised on data analysis and interpretation. VW led the implementation of the study and advised on data analysis and interpretation. MG, CM, DK, NJ and VW provided comments on drafts of this manuscript. All authors had access to the data collected for this study and all approved the final submitted version.

FUNDING AND ROLE OF THE FUNDER

This study was funded by the Medical Schools Council Assessment Alliance (MSCAA). The MSCAA Board members advised on study design and the paper itself and the MSC Executive provided feedback on a draft of this paper and agreed on submission of the results for publication. (See competing interests section.)

COMPETING INTERESTS

(1) CAT has financial support from MSCAA for the submitted work; (2) all authors are, or were Board members of the MSCAA (the funder), who has an interest in the submitted work, but unremunerated for this purpose; (3) MG participated in the standard setting process at their medical school during either 2013/14 and/or 2014/15; and (4) VW Chairs, and CAT and NJ are members of, the GMC Assessment Advisory Board.

Appendix Table 1: Effect of factors potentially affecting relative passing standards, by year

Factor potentially effecting relative passing standard	2013/14 Total N=22 N in primary analysis = 19	2014/15 Total N=30 N in primary analysis = 26
Time per item (schools only using single best answer/multiple choice items)		
Number of schools	17	20
Median (range) time per item, seconds	72 (60 to 120)	72 (55 to 108)
Kendall's tau-b correlation coefficient (p-value)	-0.13 (0.520)	-0.26 (0.134)
Method of standard setting (schools using Angoff/Modified Angoff vs. Ebel/Modified Ebel; the exact method of implementation within these broad groups varied across schools)		
Number of schools	19	26
Angoff: Mean coefficient (SD), N	-0.07 (3.57), 12	-0.22 (4.50), 19
Ebel: Mean coefficient (SD), N	1.76 (4.61), 7	4.25 (8.82), 7
T statistic (p-value)*	-0.91 (0.386)	-1.28 (0.240)
Time from start of Foundation Programme (durations >13 months indicate examinations held in the penultimate year of study)		
Number of <i>examinations</i> **	20	27
Median (range) time in months	5 (2 to 16)	5 (2 to 16)
Kendall's tau-b correlation coefficient (p-value)	-0.18 (0.293)	-0.12 (0.423)

* unequal variances assumed; ** one school used the CC items across graduation-level examinations held in both penultimate and final years (i.e. with different students) and is included twice in this analysis.

REFERENCES

1. General Medical Council. Promoting Excellence: Standards for medical education and training. London: GMC, 2016.
2. de Vries H, Sanderson P, Janta B, Rabinovich L, Archontakis F, Ismail S, et al. International comparison of ten medical regulatory systems. Cambridge: RAND Europe, 2009.
3. General Medical Council. How we monitor the quality of education and training - the Quality Assurance Framework 2016 [25/04/2016]. Available from: <http://www.gmc-uk.org/education/27080.asp>.
4. General Medical Council. Theme 5: Developing and implementing curricula and assessments 2016 [25/04/2016]. Available from: <http://www.gmc-uk.org/education/27394.asp>.
5. Norcini J. Setting standards on educational tests. Medical Education. 2003;37(5):464-9.
6. McCrorie P, Boursicot KAM. Variations in medical school graduating examinations in the United Kingdom: Are clinical competence standards comparable? Med Teach. 2009;31(3):223-9.
7. QAA. UK Quality Code for Higher Education - Chapter B7: External examining. Gloucester: 2011.
8. Boursicot K, Roberts T, Pell G. Standard setting for clinical competence at graduation from medical school: a comparison of passing scores across five medical schools. Adv Health Sci Educ Theory Pract. 2006;11(2):173-83.
9. Boursicot KAM, Roberts TE, Pell G. Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. Med Educ. 2007;41(11):1024-31.
10. Chesser A, Cameron H, Evans P, Cleland J, Boursicot K, Mires G. Sources of variation in performance on a shared OSCE station across four UK medical schools. Med Ed. 2009;43(6):526-32.
11. Case SM, Ripkey DR, Swanson D. The relationship between clinical science performance in 20 medical schools and performance on Step 2 of the USMLE licensing examination. 1994-95 Validity Study Group for USMLE Step 1 and 2 Pass/Fail Standards. Academic Medicine. 1996;71(1):S31-3.
12. Homer M, Darling JC. Setting standards in knowledge assessments: Comparing Ebel and Cohen via Rasch. Medical Teacher. 2016:1-11.
13. Cohen-Schotanus J, van der Vleuten C. A standard setting method with the best performing students as point of reference: Practical and affordable. Medical Teacher. 2010;32(2):154-60.
14. George S, Haque M, Oyebode F. Standard setting: Comparison of two methods. BMC Medical Education. 2006;6(1):46.
15. Greenaway D. Securing the future of excellent patient care. London: General Medical Council, 2013.
16. General Medical Council. Progress of doctors in training split by medical school. London: GMC, 2014.
17. Medical Schools Council. Medical Schools Council Assessment Alliance 2015 [08/12/2015]. Available from: <http://www.medschools.ac.uk/MSCAA/Pages/default.aspx>.
18. Melville C, Gurnell M, Wass V. Quality assurance of SBA questions: Developing a bank of high quality questions for undergraduate finals: Poster presented at AMEE, 2015 Glasgow2015 [22/10/2015]. Available from: <http://www.medschools.ac.uk/SiteCollectionDocuments/AMEE-2015-Poster.pdf>.
19. Cizek G, Bunch M. Standard Setting: A guide to establishing and evaluating performance standards on tests. Thousand Oaks, CA: Sage Publications; 2007.

20. Rust J, Golombok J. *Modern Psychometrics: The Science of Psychological Assessment*. London: Routledge; 1999.
21. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med*. 2006;119(2):166.e7-.e16.
22. Faul F, Erdfelder E, Lang A-G, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods*. 2007;39:175-91.
23. Angoff W. Scales, norms and equivalent scores. In: Thorndike R, editor. *Educational Measurement*. Washington, DC: American Council on Education; 1971. p. 508-600.
24. Ebel RL. Procedures for the analysis of classroom tests. *Educ Psychol Meas*. 1954;14:352-64.
25. Cohen J. *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum; 1988.
26. General Medical Council. *Assessment in undergraduate medical education*. London: General Medical Council, 2011.
27. Cusimano MD. Standard setting in medical education. *Academic Medicine*. 1996;71(10):S112.
28. Hurtz GM, Auerbach MA. A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement*. 2003;63(4):584-601.
29. O'Mara DA, Canny BJ, Rothnie IP, Wilson IG, Barnard J, Davies L. The Australian Medical Schools Assessment Collaboration: benchmarking the preclinical performance of medical students. *Med J Aust*. 2015;202(2):95-8.
30. Livingston SA, Zieky MJ. A comparative study of standard-setting methods. *Appl Meas Educ*. 1989;2(2):121-41.

References for Box 1

1. Angoff W. Scales, norms and equivalent scores. In: Thorndike R, editor. *Educational Measurement*. Washington, DC: American Council on Education; 1971. p. 508-600.
2. Ebel RL. Procedures for the analysis of classroom tests. *Educ Psychol Meas*. 1954;14:352-64.
3. Cizek G, Bunch M. *Standard Setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications; 2007.