*Journal of Advances in Medicine and Medical Research*

***26(11): 1-6, 2018; Article no.JAMMR.42576***
***ISSN: 2456-8899***
*(Past name: British Journal of Medicine and Medical Research, Past ISSN: 2231-0614, NLM ID: 101570965)*

# National Medical Exams in Australia – is it Possible?

**John J. Barnard[1,2*]**

*[1]EPEC Pty Ltd., Australia.*
*[2]University of Sydney, Australia.*

***Author's contribution***

*The sole author designed, analyzed and interpreted and prepared the manuscript.*

***Article Information***

DOI: 10.9734/JAMMR/2018/42576
Editor(s):
(1) Muhammad Torequl Islam, Nuclear of Pharmaceutical Technology (NTF), Federal University of Piaui, Brazil.
Reviewers:
(1) Berna Musal, Dokuz Eylul University, Turkey.
(2) Nazan Karaoglu, Necmettin Erbakan University, Turkey.
Complete Peer review History: http://www.sciencedomain.org/review-history/25226

***Original Research Article***

***Received 8th April 2018***
***Accepted 18th June 2018***
***Published 22nd June 2018***

## ABSTRACT

The question of national medical exams in Australia has sporadically been debated and both advantages and disadvantages have been discussed. The value and importance of monitoring minimum standards over schools is generally accepted but arguments that national exams may impede flexibility of schools' curricula and instructional methods are used to oppose such exams. Different methods which can be used to administer national exams without limiting schools are discussed in this paper. By including questions agreed on by different schools in their assessments or by using different subsets of questions, valid comparisons of performance can be achieved. If enough questions are included, robust results can be obtained.

*Keywords: Medical; national; exams; standards; Rasch.*

## 1. INTRODUCTION

Whilst some countries such as the USA have national medical exams (the United States Medical Licensing Examination (USMLE)) others do not, e.g. the United Kingdom which relies on external examiners and General Medical Council (GMC) inspections. Two of the main reasons for considering national licensing exams are the increase of medical schools and a lack of consistency in assessment methods. It is argued that national exams can help ensure a common standard so that minimal standards are met, can improve public trust and confidence and such

**Corresponding author: E-mail: JohnBarnard@bigpond.com, John@EPECat.com;*

exams can give a reliable, transparent and valid measure of candidates' qualifications. Some arguments against such exams are efforts and costs to maintain the process, possible exposure of weaknesses in medical schools and a fear of impeding flexibility within medical schools' curricula. Proponents against national exams claim that national exams will cause curriculum and instructional methods to be standardized, eliminating creativity and innovation, will stratify medical schools into explicit league tables and that a national exam is not as comprehensive as ongoing assessments [1].

Research has shown that standards over medical schools do differ [2,3,4,5,6]. In a review of the impact of large scale licensing examinations in highly developed countries, Archer, Lynn, Coombes, Roberts, Gale and Price [7] concluded from their review of databases from 2005 to 2015 that the debate on licensure examinations is characterized by strong opinions but is weak in terms of validity evidence.

There are currently 19 medical schools in Australia. Although there are significant similarities amongst the schools, there are also important differences such as graduate versus undergraduate entry, course duration and teaching methods. Koczwara, Tattersall, Barton, Coventry, Dewar, Millar, Olver, Schwarz, Starmer, Turner and Stockler [8] point out that although it is commonly assumed that the quality of medical school education in Australia is uniformly high, there is no national process for assessing its outcomes. They conclude that there is substantial variability in the content of medical school curricula and that a national exit examination could provide a uniform standard of assessment for all medical school graduates in Australia, as well as foreign graduates applying to work in Australia.

In this article three different studies to determine if national medical exams in Australia are practically possible to administer with the aim of setting minimum standards without impeding on medical schools' teaching programs are discussed.

## 2. METHODS

Benchmarking data collected by the Australian Medical Schools Assessment Collaboration (AMSAC) and the Medical Deans of Australia and New Zealand (MDANZ) in 2017 was used to explore different methodologies to ensure that minimum standards are met in medical schools without limiting medical schools in their curriculum and instructional methods.

### 2.1 AMSAC

The Australian Medical Schools Assessment Collaboration (AMSAC) was formed by a group of seven medical schools in 2008 [9] with the aim of schools to compare confidentially the performance of their own students with students at other Australian medical schools at a similar stage of their respective programs. The number of schools and students have gradually increased and in 2017 AMSAC included 18 of the 19 medical schools with over 4000 participating students.

Schools supply single best answer questions annually, blueprinted to an agreed pre-clinical curriculum with a focus on selected topics in Structure (Anatomy) and Function (Physiology) covered in the first two years of medicine. From this initial question set items are removed if they are repetitive, technically flawed or do not map to the blueprint. A reviewing body with representatives from the participating schools then sits to review the questions and approximately 60 questions are retained. This set of questions is circulated to the schools for their assessment committees to review and they are asked to nominate any questions they would reject in their assessment process. The question set is then culled of any questions rejected by multiple schools and the final set of 50 items is drawn from those remaining so as to match the blueprint.

To ensure anonymity, an independent organization, EPEC Pty Ltd, assigns unique identification codes to each school only known to the school. The codes are changed each year and the results are presented in terms of percentiles to preserve anonymity of the size of schools. Any school can see where it stands against other schools without knowing the identity of the other schools.

Although the time of delivery of the test is at the discretion of the participating schools, most schools don't include all 50 items mainly due to differences in the timing of the delivery of some curricula and not all schools use a single end of the year summative assessment. Schools are requested to administer the test soon after the completion of the teaching of the content assessed to ensure that results are as

comparable as possible. In 2017 schools included between 15 and 50 items.

Schools send their data to EPEC for analysis and reporting. A Rasch calibration [10] yields question difficulties and student abilities on a single interval-level logit scale. Since the measures on such scales range between -3 and 3 and include decimals, EPEC reports the measures for each school as a scaled score on a scale with a question difficulty mean of 500 and standard deviation of 100. Questions not answered by all students in a school is treated as missing data and if one or more students in a school didn't respond to one or more questions, such responses are treated as incorrect. All responses are scored as either correct or incorrect. Fig. 1 shows the distributions of the scaled scores as box-and-whisker plots.

Number correct scores cannot be used for comparisons since some schools may include more difficult question sets than others. Table 1 shows the mean scaled score in rank order and the number-correct mean obtained from the questions responded as a percentage for each school.

The difference in rank between the scaled scores and the number-correct score means is due to the difference in difficulty between the sets of questions responded to, not accounted for in the number-correct scores.

## 2.2 MDANZ

Two different benchmarking projects of the Medical Deans of Australia and New Zealand (MDANZ) to develop benchmarks for medical schools are briefly discussed. The first study reports on annual benchmarking of final year students in one major discipline area assessing the knowledge base of graduating students whilst the second study reports on a project in which all disciplines are included.

### 2.2.1 Study 1

In this study the same protocol was followed as for AMSAC except that items from a single discipline was included in the final 60 single best answer question test. In 2014 internal medicine questions were included, in 2015 paediatric and child health questions, in 2016 surgery questions and in 2017 psychiatry questions.

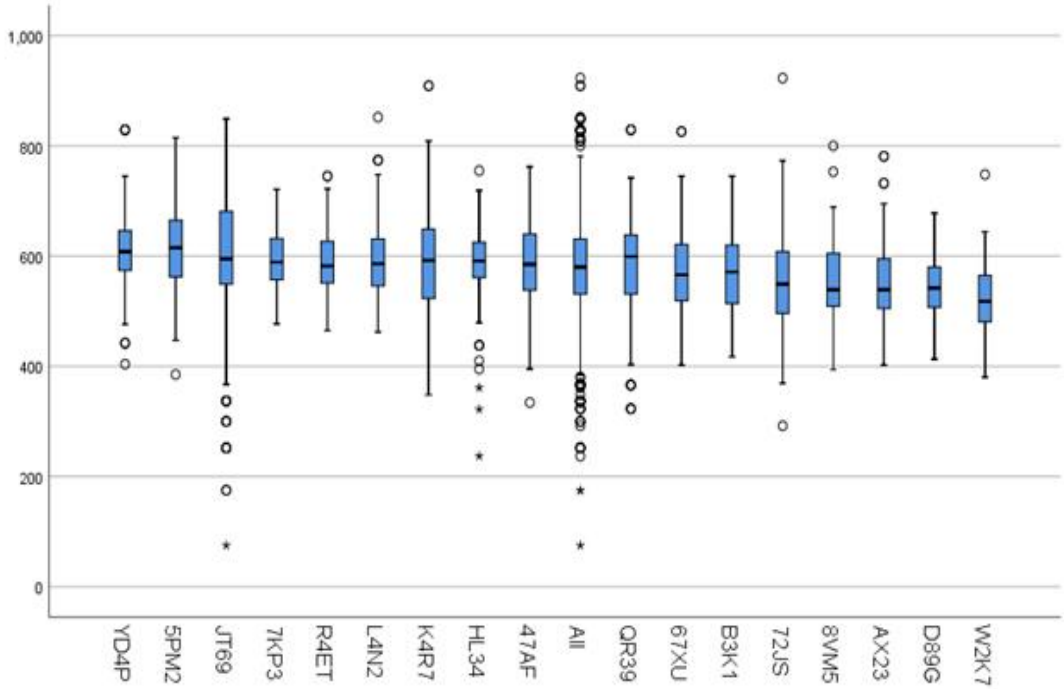


**Fig. 1. Distribution of scaled scores of 2017**

**Table 1. Mean scaled score and mean number-correct score as a percentage**

| School | Scaled | Mean |
|---|---|---|
| YD4P | 611.01 | 66.0% |
| 5PM2 | 609.28 | 73.0% |
| JT69 | 597.01 | 66.7% |
| 7KP3 | 595.28 | 68.4% |
| R4ET | 594.76 | 68.6% |
| L4N2 | 593.32 | 68.4% |
| K4R7 | 592.59 | 67.6% |
| HL34 | 590.07 | 69.5% |
| 47AF | 581.06 | 65.0% |
| ALL | 580.82 | 67.0% |
| QR39 | 576.54 | 63.6% |
| 67XU | 572.92 | 64.8% |
| B3K1 | 571.33 | 66.5% |
| 72JS | 554.49 | 62.7% |
| 8VM5 | 552.48 | 58.4% |
| AX23 | 549.14 | 61.9% |
| D89G | 548.67 | 62.2% |
| W2K7 | 520.73 | 54.4% |

In 2017 more than 3 500 students from 16 schools participated. There were 21 different data sets as some schools reported on two different cohorts. Cohorts included between 9 and 60 items. After investigating data fit to the model, Rasch calibration yielded item difficulties that could be anchored so that scaled scores for individual cohorts could be derived. As in the case of AMSAC the rank order of the schools differed when based on the Rasch scaled scores and the number-correct percentage scores.

Table 2 shows the mean scaled score in rank order and the mean number-correct score as a percentage for each cohort of the 2017 study.

As in the case of AMSAC, the rank order of the scaled score means and the number-correct means differed mainly due to cohorts responding to subsets of items which differed in difficulty.

### 2.2.2 Study 2

Instead of using a fixed test as in Study 1 and AMSAC, Study 2 asked medical schools to select any number of questions from any discipline from a pool of questions to match their curricula and other specifications. The Australian Medical Council (AMC) made 1 200 questions from their calibrated item bank used to screen International Medical Graduates (IMGs) available for selection. To ensure that there will be some common items over all participating universities, schools were asked to include ten questions nominated by the MDANZ in their selections.

**Table 2. Mean scaled score and number-correct mean as a percentage**

| School | Scaled | Mean |
|---|---|---|
| ZLC-48 | 662 | 80.1% |
| HNJ-17 | 658 | 83.0% |
| DKZ-89 | 657 | 75.8% |
| TCX-64 | 656 | 76.2% |
| ESG-67 | 655 | 76.2% |
| FZL-46 | 642 | 72.7% |
| GMT-36 | 637 | 71.1% |
| YAM-37 | 637 | 78.5% |
| ALL | 629 | 71.9% |
| AXP-38 | 629 | 72.4% |
| SQB-51 | 625 | 73.4% |
| PNC-74 | 623 | 79.3% |
| BRG-23 | 621 | 71.2% |
| BYR-79 | 620 | 70.3% |
| NFG-28 | 618 | 71.3% |
| KAD-57 | 616 | 73.6% |
| MKP-42 | 614 | 67.0% |
| XTR-45 | 613 | 70.2% |
| CQN-52 | 608 | 70.2% |
| REQ-69 | 600 | 71.2% |
| DLT-81 | 571 | 63.7% |
| URF-31 | 562 | 61.1% |

A similar process was followed as for AMSAC, except that the difficulties of all questions were known beforehand from calibration of the AMC bank. These difficulties could be used to derive measures for the schools as they were already linked to a common interval scale. The 15 cohorts of schools that participated in the study in 2017 included between 7 and 170 questions from all disciplines in their assessments.

Part 1 of the analysis considered the ten nominated items only. Four schools didn't include any of the ten items and the remaining schools included between six and ten items. The reliability of the 10-item test was 0.43, and although scaled scores could be calculated, the results were not robust enough to draw sound conclusions - more common items are required. Using the Spearman-Brown prophecy formula, it was determined that at least 53 such items need to be included to achieve a reliability of 0.80.

Whereas Part 1 of the study was similar to AMSAC and MZANZ Study 1, but with much less common items, Part 2 of the study considered all the items included by the schools in their assessments. Seven cohorts responded to more than 50 items and thus had robust results [11]. The mean scaled scores of these seven cohorts ranged from 548 to 609.

## 3. DISCUSSION

Whist the sensitivities around national medical exams in Australia need to be addressed, these studies demonstrate that such exams is technically and practically possible. The studies which included a fixed test yielded sound and similar results when enough questions were included. The AMSAC 50-item test had a (KR-20) reliability index of 0.89, and using Spearman-Brown's prophecy formula [12] it was found that at least 25 items were required for a reliability of 0.80. Only 11 schools in the study included 25 or more items in their assessments. A similar result was obtained for the 60-item MDANZ study whilst the 10-item MDANZ study confirmed that ten items were not enough to obtain robust results. The second MDANZ study showed that if enough items are included in the assessments, none of these need to be common over schools as long as they come from a pool of items calibrated on a common scale. In fact, whilst the other studies showed that the difference between the lowest scaled mean and the highest scaled mean was around 100, this range was less for the integrated assessments.

## 4. CONCLUSION

The studies showed that valid comparisons cannot be made if number-correct scores are used since such scores don't account for differences in difficulty of item subsets. A measurement framework such as Rasch or Item Response Theory which places performance on an interval level scale is required. Such a scale can be constructed from the available data or if the items come from a pre-calibrated pool of items, the external scale can be used. Rasch and IRT measurement frameworks can also elegantly deal with missing data so that all schools don't have to include exactly the same questions or the same number of questions in their assessments. The second MDANZ study demonstrated that questions can also come from different disciplines.

These methods can be used to ensure that minimum standards are met without limiting medical schools in their curriculum and instructional methods.

## CONSENT AND ETHICAL APPROVAL

It is not applicable.

## ACKNOWLEDGEMENTS

The Australian Medical Schools Assessment Collaboration (AMSAC), the Medical Deans of Australia and New Zealand (MDANZ) and the Australian Medical Council (AMC) for permission to use the data.

## COMPETING INTERESTS

Author has declared that no competing interests exist.

## REFERENCES

1. Bajammal S, Zaini R, Abuznadah W, Al-Rukban M, Aly SM, Boker A, Al-Zalabani A, Al-Omran M, Al-Habib A, Al-Sheikh M, Al-Sultan M, Fida N, Alzahrani K, Hamad B, Shehri MA, Abdulrahman KB, Al-Damegh S, Al-Nozha MM, Donnon T. The need for national medical licensing examination in Saudi Arabia. BioMed Central Medical Education 2008;8:53. DOI: 10.1186/1472-6920-8-53.
2. Boursicot KA, Roberts TE, 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;11:173–183. DOI: 10.1007/s10459-005-5291-8.
3. Boursicot KA, Roberts TE, Pell G. Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. Med Educ. 2007; 41:1024–1031. DOI: 10.1111/j.1365-2923.2007.02857.x.
4. Fowell SL, Maudsley G, Maguire P, Leinster SJ, Bligh J. Student assessment in undergraduate medical education in the United Kingdom, 1998. Med Educ. 2000; 34:1–49. DOI:10.1046/j.1365-2923.2000.0340s1001.x.
5. McManus IC, Elder AT, de Champlain A, Dacre, JE, Mollon J, Chis L. Graduates of different UK medical schools show substantial differences in performance on MRCP(UK) Part 1, Part 2 and PACES examinations. BMC Med. 2008;6:5. DOI: 10.1186/1741-7015-6-5
6. Wakeford R, Foulkes J, McManus C, Southgate L. MRCGP pass rate by medical school and region of postgraduate training.

Royal College of General Practitioners. BMJ. 1993;307:542–543.
7. Archer J, Lynn N, Coombes L, Roberts M, Gale T, Price T, Regan de Bere S. The impact of large scale licensing examinations in highly developed countries: A systematic review. BioMed Central Medical Education 2016;16:212. DOI: 10.1186/s12909-016-0729-7
8. Koczwara B, Tattersall MH, Barton MB, Coventry BJ, Dewar JM, Millar JL, Olver IN, Schwarx MA, Starmer DL, Turner DR, Stockler MR. Achieving equal standards in medical student education: Is a national exit examination the answer? Med J Aust. 2005;182(5):228-30.
9. 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. The Medical Journal of Australia. 2015;202(2):95-98.
10. Bond TG, Fox CM. Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed). New Jersey: Lawrence Erlbaum Associates; 2013.
11. Wright BD, Stone MH. Best Test Design Chicago: MESA Press; 1979.
12. Crocker LM, Algina J. Introduction to classical and modern test theory. New York, NY: Holt, Rinehart and Winston Inc.; 1986.



*Peer-review history:*
*The peer review history for this paper can be accessed here:*
*http://www.sciencedomain.org/review-history/25226*