The Australian Medical Schools Assessment Collaboration: benchmarking the preclinical performance of medical students

Medical school characteristics account for only a small part of overall variation in student performance

Deborah A O'Mara BA(Hons), DipEd, PhD[1]

Ben J Canny BMedSc(Hons), MB BS, PhD[2]

Imogene P Rothnie BA BSc, PostGradDipPsychol, MEd[1]

Ian G Wilson MAssess&Eval, PhD, FRACGP[3]

John Barnard MSc, MEd, PhD[4]

Llewelyn Davies MB BS, MD, FRACP[1,5]

1 Sydney Medical School, Sydney, NSW.

2 Monash University, Melbourne, VIC.

3 University of Wollongong, Wollongong, NSW.

4 EPEC (Excel Psychological and Educational Consultancy), Melbourne, VIC.

5 Royal Prince Alfred Hospital, Sydney, NSW.

deborah.omara@sydney.edu.au

doi: 10.5694/mja14.00772

Abstract

Objectives: To report the level of participation of medical schools in the Australian Medical Schools Assessment Collaboration (AMSAC); and to measure differences in student performance related to medical school characteristics and implementation methods.

Design: Retrospective analysis of data using the Rasch statistical model to correct for missing data and variability in item difficulty. Linear model analysis of variance was used to assess differences in student performance.

Setting and participants: 6401 preclinical students from 13 medical schools that participated in AMSAC from 2011 to 2013.

Main outcome measures: Rasch estimates of preclinical basic and clinical science knowledge.

Results: Representation of Australian medical schools and students in AMSAC more than doubled between 2009 and 2013. In 2013 it included 12 of 19 medical schools and 68% of medical students. Graduate-entry students scored higher than students entering straight from school. Students at large schools scored higher than students at small schools. Although the significance level was high ($P < 0.001$), the main effect sizes were small (4.5% and 2.3%, respectively). The time allowed per multiple choice question was not significantly associated with student performance. The effect on performance of multiple assessments compared with the test items as part of a single end-of-year examination was negligible. The variables investigated explain only 12% of the total variation in student performance.

Conclusions: An increasing number of medical schools are participating in AMSAC to monitor student performance in preclinical sciences against an external benchmark. Medical school characteristics account for only a small part of overall variation in student performance. Student performance was not affected by the different methods of administering test items.

Since 2000, the number of medical schools in Australia has expanded rapidly. Seven new schools have been established in the past 15 years, so there are now 19 medical schools. The schools differ with regard to entry point, curricula, course duration and teaching methods. Schools may identify themselves as aiming to achieve specific graduate attributes, for instance, producing future specialists, rural health practitioners or medical researchers, while others are more general in their aims.

Medical schools need a process that allows them to measure changes in their students' and graduates' performance relative to other medical schools, so they can evaluate changes to their selection criteria, curricula and teaching methods.

The Australian Medical Schools Assessment Collaboration (AMSAC) was established by a group of seven medical schools in 2008. It aims to allow Australian medical schools to share assessment items and performance statistics within an anonymous framework. While other assessment collaborations exist for item sharing, there has been no previous sharing of Australian student performance data. The formation principles guard against the construction of league tables. In 2013, the AMSAC collaboration comprised 12 Australian medical schools (Appendix 1).

The collaboration assesses preclinical medical education, in the sense that schools generally administer the items at the point where students make the transition from predominantly campus teaching to mainly clinical settings. This division is more clear cut in some schools than others. Almost all AMSAC collaborators have embedded the items in second-year examinations, irrespective of the total length of their programs.

AMSAC assessment generation

The project creates an agreed set of 50 items for participating schools to include in summative assessments. A multiple choice format with one correct answer is used. This is the most widely used written item type in assessment of basic and clinical sciences[1,2] and, if well designed, assesses reasoning as well as factual recall.[3,4] We have used the five-option format standardised by the United States National Board of Medical Examiners,[4] although we acknowledge that varying the option number has little effect on item performance.[5,6] The items are mapped to a blueprint (Appendix 2) that covers two broad basic science domains — function and structure — and are managed through the Sydney Medical School's assessment database.[7] All participating schools contribute items, which are reviewed at an annual collaborators' meeting. A short list of 60 items is circulated and items nominated by multiple schools as problematic are eliminated to produce the final set of 50. About half the items have been used previously with good performance and serve as "anchors" for interyear comparison.

Including the items in assessments

The chosen items are delivered to a single student cohort in the collaborating schools as part of one or more summative assessments over a calendar year. Schools vary in terms of the number of items they include in their assessments (Appendix 3), as assessment and curriculum timing

1 Participation in the Australian Medical Schools Assessment Collaboration

	Year of administration				
	2009	2010	2011	2012	2013
No. of medical schools (n = 19)	6	6	8	11	12
No. of students assessed	1035	1293	1666	2358	2377
Proportion of all medical students in equivalent year*	33%	39%	51%	65%	68%

*Data were obtained from Medical Deans of Australia and New Zealand 2013 medical student statistics (http://www.medicaldeans.org.au/wp-content/uploads/Website-Stats-2013-Table-2.pdf); Table 2(a): Total student enrolments 2013 by year of course (Australia). ◆

affect item relevance. The time allowed for each item varies between collaboration members; from a low of 60 seconds to a high of 120 seconds.[8]

The performance of individual students on the AMSAC items is collated and analysed by an independent consultancy, so schools can preserve their anonymity through the use of confidential identifiers. The data are analysed using the Rasch model which, unlike classical test theory, accounts for missing data in estimates of item difficulty and student performance, enabling valid comparisons to be made across the cohort, irrespective of which of the 50 items are administered. Rasch analysis has been applied widely in medical assessment.[9,10]

Study objectives

Institutional variation in the quality of assessments in other countries has been noted,[11,12] but there have been few reports comparing medical school assessment outcomes in Australia.[13] Our aim in this study was to report participation by medical schools and their students in AMSAC and to determine whether there were differences in student performance related to medical school characteristics and test administration methods.

Methods

We used Medical Deans of Australia and New Zealand (MDANZ) data on year two enrolments at schools outside the collaboration to calculate the equivalent-year population base for 2009–2013. (Box 1).

The Rasch measure (Winsteps, version 3.80.1; Winstep Software Technologies) for each student was used to investigate differences related to medical schools and variations in implementation. Although Rasch estimates were derived for both domains (structure and function), some schools implemented too few questions to provide a reliable basis for analysis of individual domains.

We used a general linear model analysis of variance (ANOVA) in SPSS Version 21 (IBM). Five independent variables (year of administration, entry requirement, school size, number of assessments and time per item) were applied to the models. Significance was defined at $P < 0.01$ to account for Type I errors. Effect size was assessed using the F test and partial η^2, multiple interactions using the Scheffé test and reliability using the Kuder-Richardson Formula 20 index (KR20).

Results

Participation in AMSAC

The representation of Australian medical schools and students in AMSAC more than doubled between 2009 and 2013. In 2013 it included 12 of 19 medical schools (Box 1) and 68% of medical students.

Although the initial collaboration was formed by seven medical schools, one was unable to field questions in 2009, and another was unable to in 2010. Two medical schools participated for 2 years and withdrew owing to difficulty in matching the agreed blueprint within a single cohort; one recently rejoined.

The proportion of graduate-entry schools grew from half of the participating schools in 2009 to two-thirds in 2013. The proportion of students who attended a graduate-entry school and participated in AMSAC was relatively stable; 69% in 2009, 62% in 2011 and 67% in 2013. The proportion of AMSAC schools that could be classified as large (intake of 150 or more students) remained at about half, with the corresponding student representation being 61% in 2009, 74% in 2011 and 73% in 2013.

In 2013, AMSAC comprised six of the 10 small schools and six of the nine large schools. In terms of entry requirement, it comprised eight of the 11 schools with graduate entry and four of the eight schools admitting school leavers.

Statistical analysis

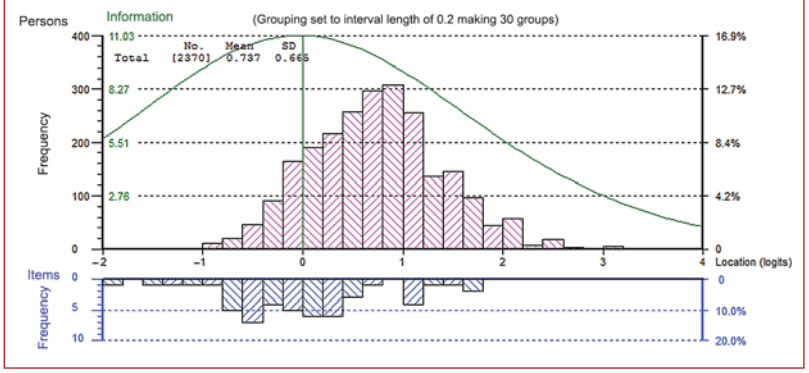
Rasch analysis found the item set to be psychometrically sound with a good fit of items to the model (Box 2), and the ability estimates for student performance are robust. The KR20 score for the 2013 AMSAC question cohort was 0.85. The average student performance varied slightly across the five implementations.

Our statistical analysis was performed only on the 2011–2013 data owing to variations in item difficulty, fewer participating schools in the early years of the collaboration and increased availability of reliable marker questions. As the number of reliable marker questions has increased over the years, the performance of the question set has stabilised.

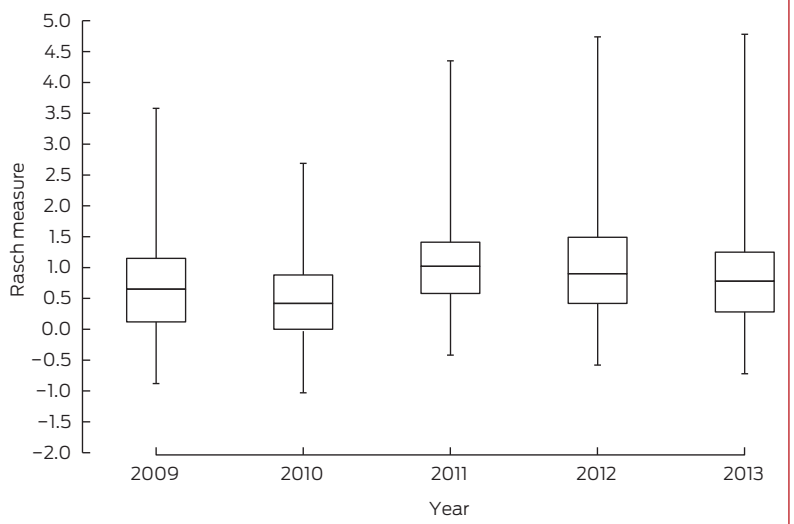
Effect of independent variables

The ANOVA showed significant differences ($P < 0.001$) for four of the five independent variables. Entry type was the most significant main effect (F test and partial η^2), followed by school size, year of implementation and number of assessments (Appendix 4). The effect of time allowed per multiple choice question was not significant ($P = 0.681$). The interaction of year and each of the four independent variables was also significant on ANOVA. The mean and, to a lesser extent, the standard deviation of student performance for each interaction clarifies the strength

2 Australian Medical Schools Assessment Collaboration student and item Rasch distributions, 2013



3 Rasch measure* for each year



*Each box displays the median score (horizontal line within the box) and the interquartile range. ◆

and direction of the significant differences (Box 4).

The set of test items for the 2013 implementation was significantly harder (Scheffé test) than the 2011 and 2012 sets (2011 mean difference (MD) = – 0.22; $P < 0.001$; 2012 MD = – 0.18; $P < 0.001$). The mean scores for 2011 and 2012 were not significantly different (MD = 0.04; $P = 0.29$). Although the difference in the means was statistically significant, the magnitude was less than 0.5 logit, the definition of a substantive difference[14] in the Rasch model. There was a large overlap between the interquartile ranges of student performance in each year (Box 3).

Students from large schools performed better than students from small schools. The difference in performance increased over time and was significant in 2012 and 2013 (Appendix 4); however, the difference attributable to school size explained only 2.3% of variation. The interaction with year of implementation explained a further 0.5%. Even at its greatest (2013; – 0.38), the difference due to school size was still less than 0.5 logit.

Graduate-entry students performed better than those entering medicine directly from school (ANOVA $P < 0.001$; Appendix 4). The proportion of variance attributable to entry type was 4.5%. The interaction of entry type with year of implementation was not as strong ($\eta^2 = 2.3$) due to the variation over the 3 years for the two options (Box 4).

The time allowed per item was not significantly associated with performance ($P = 0.681$). While the number of assessments had a significant effect on performance ($P < 0.001$), the effect size was too low to be meaningful (partial $\eta^2 = 0.003$) (Appendix 4). The frequency of summative assessments increased each year of the collaboration, so some medical schools were reclassified over time (Box 4).

There was no interdependence between the independent variables explored in this study. The strongest association was between size of medical school and use of AMSAC questions in more than one assessment ($r = 0.44$). The multiple regression with all five independent variables explained only 12% of total variance ($R^2 = 0.12$). The standardised beta weights were 0.28 for entry type, 0.26 for size of school, 0.14 for time per item, – 0.10 for year and – 0.07 for number of assessments. Thus, 88% of variation in medical student performance was due to factors outside of our model.

Discussion

The AMSAC project demonstrates the viability of linking assessments across medical schools. As few as 25 common items enable reliable interschool comparisons. Individual medical schools can use AMSAC data to assess the effect of changes in their curriculum or entry requirements and to evaluate the need for change; however, fitting all questions into all school curricula for a single cohort is challenging.

AMSAC is a broad collaboration of medical schools that vary in size, selection criteria, course duration and syllabus content. The overall sampling of Australian medical school students has grown from one-third of the preclinical cohort in 2009 to over two-thirds in 2013 and includes 14 schools in 2015. Two schools have left the collaboration due to difficulty with curricula mapping, though one has recently rejoined.

A key outcome of this process has been the collegiate interaction of the medical school representatives. The schools involved have acquired a better understanding of the structure

4 Mean Rasch scores by school type and implementation method

		Year of implementation			
		2011	2012	2013	Total
School type					
Small school	Mean (SD)	0.94 (0.82)	0.67 (0.71)	0.51 (0.63)	0.68 (0.73)
	No. of students	431	420	638	1489
Large school	Mean (SD)	1.03 (0.64)	1.03 (0.86)	0.89 (0.70)	0.98 (0.75)
	No. of students	1235	1938	1739	4912
School leaver entry	Mean (SD)	0.82 (0.70)	0.61 (0.68)	0.73 (0.71)	0.71 (0.70)
	No. of students	635	985	775	2395
Graduate entry	Mean (SD)	1.12 (0.66)	1.23 (0.85)	0.82 (0.69)	1.03 (0.77)
	No. of students	1031	1373	1602	4006
Implementation method					
< 90 seconds per item	Mean (SD)	0.92 (0.59)	1.03 (0.95)	0.79 (0.73)	0.91 (0.81)
	No. of students	556	1219	1205	2980
⩾ 90 seconds per item	Mean (SD)	1.05 (0.74)	0.91 (0.70)	0.79 (0.67)	0.91 (0.71)
	No. of students	1110	1139	1172	3421
Single assessment	Mean (SD)	1.24 (0.76)	1.00 (0.72)	0.68 (0.68)	0.98 (0.75)
	No. of students	575	791	588	1954
Multiple assessments	Mean (SD)	0.88 (0.62)	0.95 (0.90)	0.82 (0.71)	0.88 (0.76)
	No. of students	1091	1567	1789	4447
Total					
	Mean (SD)	1.01 (0.69)	0.97 (0.84)	0.79 (0.70)	0.91 (0.76)
	No. of students	1666	2358	2377	6401

and content of other schools' syllabuses and assessment strategies. The increase in the quality and stability of the items during the project is a reflection of the broad engagement of course leaders in the question collection and review process, a phenomenon which has also been found in overseas collaborations.[15]

Rasch analysis allows valid comparisons between schools using less than the full set of test items and has enabled valid comparisons to be made by medical school and implementation methods. The slightly better performance of graduate-entry students in the early years of medical school may reflect their increased maturity and previous success but is also likely to reflect the substantial proportion of candidates with a medical science degree (eg, 27% of entrants to Sydney Medical School, 2011–2013). It would be surprising if an additional 3 years of study in the medical sciences did not confer any advantage in a preclinical medical science examination. A previous study from Melbourne Medical School found a similar small advantage for graduate-entry students.[16]

The small but significant difference in performance by size of school probably reflects the small size of all but one of the new Australian medical schools. The new schools are unlikely to have the depth of resources in the preclinical sciences of the established schools and are also in a stage of course stabilisation as their early cohorts graduate. Most variation in student performance on AMSAC items is due to other factors, most likely individual student ability.

AMSAC provides an opportunity for comparison of assessment strategies across schools. Over time there has been an increase in schools using multiple assessments as opposed to a single end-of-year exam. This pattern is particularly strong among the larger schools, perhaps reflecting better resourcing. The use of multiple small assessments did not improve student performance.

The project allows participating schools some ability to compare their students' knowledge base and reasoning skills with those of other schools and with a national average. This project does not assess other graduate attributes and only looks at mid-course performance, but other national collaborations are in process to examine clinical skills and knowledge and reasoning in the pregraduation phase.

The project has demonstrated that medical schools can collaborate on a benchmarking process without the need for external regulation. Early fears about data being misused to create league tables or unique syllabus content being damaged have not been realised. The AMSAC project is a model for national collaboration between medical schools to meet government and community demands for accountability without loss of school autonomy.

Acknowledgements: We thank all participating AMSAC medical schools and academic and professional assessment staff.

Competing interests: No relevant disclosures.

Received 28 May 2014, accepted 17 Sep 2014. ■

1 Schuwirth LW, van der Vleuten CP. General overview of the theories used in assessment: AMEE Guide No. 57. *Med Teach* 2011; 33: 783-797.

2 Norcini JJ, Swanson DB, Grosso LJ, Webster GD. Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. *Med Educ* 1985; 19: 238-247.

3 Berk RA. A consumer's guide to multiple-choice item formats that measure complex cognitive outcomes. Pearson, 1996. http://images.pearsonassessments.com/images/NES_Publications/1996_12Berk_368_1.pdf (accessed Sep 2014).

4 Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. Philadelphia, Pa: National Board of Medical Examiners, 1998.

5 Zoanetti N, Beaves M, Griffin P, Wallace EM. Fixed or mixed: a comparison of three, four and mixed-option multiple-choice tests in a Fetal Surveillance Education Program. *BMC Med Educ* 2013; 13: 35.

6 Taylor AK. Violating conventional wisdom in multiple choice test construction. *College Student Journal* 2005; 39: 141-148.

7 O'Mara D, Quinnell R, Rothnie I, et al. ExamBank: a pedagogic and administrative system to provide effective student feedback and stable assessment across disciplines. *International Journal of Innovation in Science and Mathematics Education* 2014; 22: 62-73.

8 Wilson I, O'Mara D, Davies L. Australian Medical Schools Assessment Collaboration: what do the differences mean? Paper presented at the Australian and New Zealand Association for Health Professional Educators; 2011 Jun 27-30; Alice Springs, Australia.

9 De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. *Med Educ* 2010; 44: 109-117.

10 Wright BD, Stone MH. Best test design: Rasch measurement. Chicago: MESA Press, 1979.

11 Reichert TG. Assessing the use of high quality multiple choice exam questions in undergraduate nursing education: are educators making the grade? [Masters of Arts in Nursing Thesis]. St Paul, Minn: St Catherine University, 2011. http://sophia.stkate.edu/cgi/viewcontent.cgi?article=1014&context=ma_nursing (accessed Sep 2014).

12 Jozefowicz RF, Koeppen BM, Case S, et al. The quality of in-house medical school examinations. *Acad Med* 2002; 77: 156-161.

13 Edwards D, Wilkinson D, Canny BJ, et al. Developing outcomes assessments for collaborative, cross-institutional benchmarking: progress of the Australian Medical Assessment Collaboration. *Med Teach* 2014; 36: 139-147.

14 Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. 2nd ed. London: Psychology Press, 2013.

15 Naeem N, van der Vleuten C, Alfaris EA. Faculty development on item writing substantially improves item quality. *Adv Health Sci Educ Theory Pract* 2012; 17: 369-376.

16 Dodds AE, Reid KJ, Conn JJ, et al. Comparing the academic performance of graduate- and undergraduate-entry medical students. *Med Educ* 2010; 44: 197-204. ■