Improving assessment practice through cross-institutional collaboration: an exercise on the use of OSCEs

Bunmi Malau-Aduli, Peta-Ann Teague, Richard Turner, Benjamin Holman, Karen D'Souza, David Garne, Clare Heal, Paula Heggarty, Judith Nicky Hudson, Ian Wilson, Cees van der Vleuten

College of Medicine and Dentistry, James Cook University, Townsville, Australia
School of Medicine, University of Tasmania, Hobart, Australia
School of Medicine, Deakin University, Geelong, Australia
Graduate School of Medicine, University of Wollongong, New South Wales, Australia

Corresponding Author: Bunmi Malau-Aduli, College of Medicine and Dentistry, Division of Tropical Health and Medicine, James Cook University, Townsville, Australia

Tel: +61747814418; Fax: +6174781 5870

E-mail: bunmi.malauaduli@jcu.edu.au

Abstract

Background: This study was undertaken to improve assessment practice on OSCEs through collaboration across geographically dispersed medical schools in Australia.

Methods: A total of eleven OSCE stations were co-developed by four medical schools and used in summative 2011 and 2012 examinations for the assessment of clinical performance in the early clinical and exit OSCEs in each school's medical course. Partial Credit Rasch Model was used to evaluate the psychometric properties of the shared OSCE data. Evaluation of the quality assurance reports was used to determine the beneficial impact of the collaborative benchmarking exercise on learning and teaching outcomes.

Results: The data for each examination demonstrated sufficient fit to the Rasch model with infit mean square values ranging from 0.88 to 0.99. Person separation (1.25 to 1.63) indices indicated good reliability. Evaluation of perceived benefits showed that the benchmarking process was successful as it highlighted common curriculum areas requiring specific focus and provided comparable data on the quality of teaching at the participating medical schools.

Conclusion: This research demonstrates the validity of the psychometric data and benefits of evaluating clinical competence across medical schools without the enforcement of a prescriptive national curriculum or assessment.

Introduction

Concerns about junior doctor competencies (Goldacre et al, 2003; Boursicot et al, 2006), has led to a call for greater integration of medical education (McGrath et al, 2006). In addition, there has been a move towards outcomes-based medical education with identifiable core competencies (Harden et al, 1999; Wilkinson, 2010; Medical Deans Australia and New Zealand, 2011) and this has coincided with a shift in societal expectations regarding the accountability of doctors. Medical scandals have shaped public perception and have increased pressure for medical training providers to be more confident in certifying the ability of their graduates (Norcini et al, 2008).

Two major proposals have emerged worldwide in response to these pressures. One proposal suggests a national medical education curriculum with a mandatory national exit exam to allow medical school benchmarking (Wilkinson, 2010; Chapius et al, 2010). At the same time, there is recognition that having a national exit examination or a highly prescriptive academic standard framework would drive counter-productive standardisation and uniformity (Australian Medical Students' Association, 2010). A rigid national assessment may not be readily applied across the diversity of medical schools.

The second proposal promotes voluntary assessment collaboration between medical schools (Harden, 2009). The argument for this approach is that it would allow greater flexibility for medical courses to maintain relevant curricula with benefits such as improved cost-effectiveness, transparency, accountability, and standardisation of process and relevant content among participating medical schools (Wilkinson, 2010; Wilkinson et al, 2014; Muijtjens et al, 2008). The applicability of either proposal merits research.

International and national consortia such as the International Database for Enhancement of Assessment and Learning (IDEAL), Universities Medical Assessment

Partnership (UMAP) and Australian Medical Assessment Collaboration (AMAC)) have developed databases and processes to provide their member schools with access to quality assessment items. However, in many countries, there is a lack of data on the attainment of clinical competencies by medical graduates. OSCEs are widely used at important clinical assessment checkpoints (Brailovsky et al, 1992; Whelan, 1999; Medical Council of Canada, 2002; Turner & Dankoski, 2008). Benchmarking medical school assessment standards, with an evaluation of the psychometric impact, may prove valuable in establishing the attainment of clinical competencies (McGrath et al, 2006; Roberts et al, 2006), as it has done for written assessment (Wilkinson et al, 2014). Given this context, there is the need for a benchmarking process that helps evaluate assessment standards and provides individual medical schools with feedback about the strengths and weaknesses of their clinical curriculum, teaching and assessment.

This research project was designed and conducted with an overarching aim of improving assessment practice on OSCEs through collaboration across geographically dispersed medical schools in Australia. The improvement was evaluated in two ways, firstly through the study of the psychometric properties of the student performance data across the different participating medical schools. For scores to be meaningfully interpreted, content-related evidence of the adequacy of the content tested, statistical evidence of score reproducibility and the assessment item's statistical quality are required (Downing, 2004). The second evaluation procedure was to explore how useful the data from the exercise was in providing the participating schools with feedback on the learning outcomes of their students. Based on the overarching aim, this research was therefore designed to answer the following questions:

1. To what extent do the OSCE performance data form a unidimensional and locally independent construct according to the Rasch measurement model?
2. What are the benefits of the exercise for the participating schools?

Methods

Participating Medical Schools

Four regional and geographically dispersed Australian medical schools (A, B, C and D – letters randomly assigned) participated in this study by sharing OSCE stations which were co-developed by an expert committee. The selected schools originated from 4 different states in Australia. Two of the schools run 4-year graduate-entry programmes, while the other two schools run undergraduate-entry medical programmes. All schools have similar horizontally and vertically integrated outcomes-based curricula. The selected year groups (early clinical and exit level) were chosen because of their comparable levels of intended learning outcomes. This collaborative venture is known as the Australian Collaboration for Clinical Assessment in Medicine (ACCLAiM).

Shared OSCE Stations

There were two phases of the collaboration in which a total of eleven OSCE stations were collaboratively developed by a committee comprising clinical and educator colleagues from each participating school. Competencies were chosen from prospectively reviewed clinical blueprints which represent a fair and reasonable assessment and which mapped to the Medical Deans Australia and New Zealand (MDANZ) medical competencies project (2011). The assessed competencies included history taking, physical examination, communication, diagnostic reasoning and knowledge of basic sciences and they were similarly weighted at each school. After achieving consensus on content and marking criteria the stations were

incorporated into the summative OSCEs at each school. Figure 1 summarises the process that was followed for the development and implementation of the shared OSCE stations.

Appendix 1 depicts the eleven stations used over the three distinct OSCE cycles, their descriptors and the competencies that each of them assessed. The first phase of this study was conducted in 2011, when four of the stations were embedded in the end of early clinical (EC) phase OSCEs. The second phase of the study was conducted in 2012 involving the same four schools and following the same procedure, but involved embedding four new stations into the early clinical (EC) exam, as well as three new stations used in the exit level exams (EE). The scoring sheets consisted of a checklist and an overall global rating scale.

Examination procedure

Each collaborative set of OSCE stations were embedded into the OSCEs (comprising either 10 or 12 stations) in each school. The collaborating schools inserted these stations into their blueprint, and designed the other OSCE assessment items around the shared stations. This approach permitted locally relevant content to be examined alongside the benchmarked competencies, without the need to fully align the entire curriculum sequence of the medical schools. The participating schools arranged the shared station 'paperwork' to fit with their local practice, to ensure that the shared OSCE stations appeared identical to the local medical school stations. Due to large numbers of students, concurrent multiple circuits of each station were used at each school. All schools had one internal local examiner per station who were experienced clinicians involved in student teaching.

To standardise marking at the four schools, a secure on-line examiner training/calibration program was developed and made available to all the assessors of the shared OSCE stations one week prior to the examination (Malau-Aduli et al, 2012). Each

school retained their pre-existing local practice in relation to examiner and role player training for the other OSCE stations.

As a means of quality assurance (QA), the consistency of assessment processes at each school was evaluated by the ACCLAiM clinical co-ordinators from the other three participating schools. To ensure QA validity, the QA examiners were selected from staff possessing an expert level of experience of OSCE design, implementation and analysis. For each examination, they were required to serve as QA examiners on the shared ACCLAiM stations as well as the internal local OSCE stations, and provided a combined QA report to the visited school based on a predetermined template. This report constructively critiqued the administration of the OSCE, and provided feedback on the academic content and student/examiner views on the OSCE. Where OSCEs were run at more than one clinical site per medical school, the QA examiners were split to cover each site.

Communication between participating schools

The ACCLAiM committee met several times per year during the study, with additional communication by teleconference and email. Administrative and academic staff members at each participating school were free to contact the other member schools at any stage during the study period to seek answers to any OSCE related queries.

Analysis

Research question 1: Partial Credit Rasch Model (PCRM) was used to evaluate the unidimensionality (i.e. the common underlying construct across the stations) and local independence (i.e. the probability of a person correctly responding to an item does not depend on the other items in the test) of the shared OSCE data. PCRM is a powerful method for interrogating clinical assessment data as it estimates students' true measures of clinical

competence by portioning the variance in raw scores into variance due to item difficulty and student ability (Marais & Andrich, 2008). The Rasch model serves as a quality assurance framework for measurement, in that it uses a unidimensionality measurement scale to determine the probability of an item score (Bond & Fox, 2012). Total students' percent scores on each shared station were converted to standardised scores and collapsed into 10 categories – zero to nine, to be fitted into the PCRM, using the Winsteps software (Linacre, 2009). This allowed for the aggregation of scores across multiple sites. Unidimensionality and local independence are assessed using fit statistics, person-item distribution, reliability and differential item functioning (DIF) measures.

Fit statistics gives an indication of the consistency of the hierarchy of station difficulty across the various students' clinical competence on the scale. It estimates the extent to which responses show adherence to the modelled expectations. Overall fit of the items to the model was examined by assessing the mean item log residual test of fit statistics. Good-fit and misfit items were identified using infit and outfit mean-square values. Expected value is 1.0 and the ideal range that is deemed productive for measurement is 0.8-1.2 (Gustafson, 1980). Lower values indicate observations are too predictable (i.e. data overfit the model) and higher values indicate unpredictability (i.e. data underfit the model).

Reliability refers to the replicability of the observed responses and it is estimated for both persons and items. The person measure reliability (PRI) indicates how well the scale can distinguish amongst persons in terms of their latent trait locations (Bond & Fox, 2012) e.g. clinical competence. A measure of person separation (PSI) is calculated to indicate the efficiency of the items in separating the persons measured (Bond & Fox, 2012). The item measure reliability (IRI) indicates how well the scale can distinguish between items, on the basis of their difficulty (Bond & Fox, 2012), and the item separation index (ISI) indicates the efficiency of the sample of persons in separating the items used. Reliability ranges from 0 to

1.0, with a score of 1.0 denoting that less of the measurement variability can be attributed to measurement error. For the separation indices, values less than 1.0 are unsatisfactory.

Item-person map visually represents the order of difficulty of items relative to each other and can easily ascertain where any individual person is located in relation to all items (Wright & Masters, 1981). Person and item locations are logarithmically transformed and plotted on the same continuum using a common unit of measurement termed logit; thereby converting ordinal data to equal-interval data, implying equal difference in ability or latent trait possession (Masters, 1982).

Differential Item functioning (DIF) tests measurement invariance by detecting test items biased towards different subgroups of test takers according to construct irrelevant factors (Hagquist & Andrich, 2004). DIF was used in this study to examine whether the OSCE stations functioned differently by entry program (graduate & undergraduate entry); gender (males & females) and origin (domestic & international students). A value of <0.43 is not significant; ≥0.43 indicates slight to moderate difference and ≥0.64 indicates moderate to large difference (Bond & Fox, 2012).

Research question 2: The impact of the benchmarking process was examined by the assessors from the participating schools. Based on an evaluation of the quality assurance (QA) reports provided to each school after the examinations, the assessors deliberated on and documented the benefit(s) and impact of the benchmarking exercise on teaching and learning in their respective schools. Over the 2-years study period, fifteen clinicians were involved in the QA examination across all participating schools and their experiences and QA reports were collated and coded for emerging themes.

Results

This research used data collected from 4470 student records, from four Australian medical schools.

Research question 1: *To what extent do the OSCE performance data form a unidimensional and locally independent construct according to the Rasch measurement model?*

The mean item log residual test of fit statistics, measuring the overall fit of the data to the Rasch model, showed that all items fitted the model with infit and out-fit values for person and item ranging from 0.88-0.90 and 0.95-0.99 respectively.

Table 1 depicts reliability measures of the three examinations. In all examinations, item reliability and separation indices were much higher than person reliability and separation indices. The results indicate that the estimated item measures (0.95) are highly reliable with only 5% measure variability attributed to measurement error within each examination. Estimated person measures (0.61-0.73) also indicated good reliability.

Figure 2 shows the Rasch item-person map of student ability and item difficulty for the three examinations. Students of greater ability and more difficult stations are towards the top, and students of lesser ability and easier stations are towards the bottom. A student plotted on the same level as a station has a 50% ($p = 0.5$) chance of passing that station. Students above that level have a greater chance of passing the station and students below have less chance of passing that station. With OSCEs focused on clinical competence, it is desirable to have all stations at a similar level of difficulty. For all examinations, there are differences in student performance between schools on individual stations, however, irrespective of school affiliations, similar patterns were observed in student performance. The items used were of average difficulty and there was a broad range of student abilities. For the 2011 EC exam, there was a broad range of student abilities from -3 to +5. The items were of average

difficulty with the abdominal pain station been slightly easier than the others and the vaccination station being slightly more difficult (Fig 2a). The 2012 EC exam (Fig 2b) shows a spread of student abilities from -2 to +4, and that items were of average difficulty with STI being the easiest. For the 2012 EE (Fig 2c), the cellulitis station was the most difficult and there was a wide spread of student abilities from -5 to +5. The item-person map confirms the suitability of the examinations to differentiate between high and low performers.

Differential Item Functioning (DIF) was measured to determine items which were biased in relation to construct irrelevant factors such as entry program, gender and origin. It indicates the relative difficulty of the items in relation to students' abilities and the higher the score on the Y axis, the more difficult the students within the subgroup found the station. Figure 3 shows DIF differences for two subgroups. In the 2011 EC exams (Fig 3a), graduate entry (GE) students found the groin station more difficult and undergraduate (UE) students found the prostate station more difficult, with a difference of about 0.5 in each case. However, there were no significant differences in the 2012 EC exams. As depicted in Fig 3b for the 2012 exit exam, domestic students found the neck lump most difficult (physical examination with lots of basic science knowledge) and international students found the asthma station most difficult (management plan and communication skills). There were no differences in performance patterns for all other subgroups in the exams (data not shown). The difficulty of each item for all subgroups was remarkably similar with only few discrepancies, indicating that the test items functioned similarly for different subgroups of examinees.

Research question 2: *what are the benefits of the exercise for the participating schools?*

Data collated from the assessors and the QA reports were categorised into three major themes namely: community of practice, learning experience and diagnostic tool.

Community of Practice: Throughout the duration of the study, ACCLAiM committee members met frequently at meetings, via teleconference and medical education conferences. This promoted the emergence of a 'community of practice', where group members could share OSCE experience and ask advice on a range of OSCE academic issues (i.e. standard setting techniques, optimal station length and reading time length, ideal means of student debriefing immediately post-OSCE). As time progressed, the committee members developed other sharing ventures and collaborations alongside this research project. They began sharing ideas and developed OSCE stations from their local station bank, thus decreasing the time-consuming and expensive work of creating new stations while diversifying the range and method of relevant competencies assessed. It was noted that administrative staff from the participating medical schools also began communicating about more practical issues (i.e. optical mark recognition (OMR) sheets, and OSCE station timers).

Learning Experience: The discussion about the content and focus of the shared OSCEs provided a rich learning experience for the collaborating clinicians, but the post exam analysis was by far the most instructive. Each school has been provided with a confidential report containing the data for their students, set in the context of the performance data of all the participating students. There were several instances where OSCEs which an individual school was confident that their students would find easy was in fact experienced as difficult. This gave rise to review of specific content and skills teaching at individual schools, particularly where students from one school performed differently to their peers at the other schools. It was also possible to identify common strengths and feed this back to teaching staff.

Diagnostic Tool: The Quality Assurance (QA) and examination performance reports provided to each school after the OSCE exam gave external qualitative and quantitative data on the overall examination process, the student experience and performance, the preparation

and performance of role players and assessors and the level of difficulty of each of the stations. Although each school was tasked with the editorial lead of one of the OSCE stations, this did not particularly benefit students of the corresponding school. The reports were able to externally identify strengths and weaknesses of each school's OSCE and provide this feedback in a constructive way.

Discussion

Prevailing assessment theory considers the primacy of construct validity, which draws upon theory and evidence to give meaning to assessment. Typically, evidence for validity is drawn from five areas to support confidence in the inferences made from assessment: curriculum content; data management; statistical analyses of test data; correlational analyses; and effects of assessment (Kane, 2006). In this study, we demonstrate the value of benchmarking and quality assurance processes in the generation of evidence to support validity of assessment scores.

The results of the Rasch analysis demonstrate that the assessment items measured the same underlying construct as evidenced by the fit statistics and the high reliability indices obtained for the three examinations, a measurable proof of how well the items had distinguished between students in terms of their latent trait ability regardless of geographical locations. Generally, the scores followed a normal distribution pattern. This is ideal for OSCEs as they are focused on clinical competence. However, in all three examinations, more or less difficult items could have been included to better distinguish between examinees with very high or very low total scores. The absence of DIF in the subgroups (gender, origin and entry program) suggests that the observed examinee scores were free of construct irrelevance, thus confirming the unidimensionality and local independence of the data.

Although similar performance trends were observed in student performance across all participating schools, the observed variations between students' mean scores on individual stations highlights the challenge of comparing performance between medical schools in 'league table' format. Local differences between medical schools reflect varied student performance and limit the comparison of results (Petrusa et al, 1991; Muijtjens et al, 2008; Chesser et al, 2009). We need to understand these differences between schools much better before we embark on a "one size fits all" assessment benchmarking strategy. This requires considerable further research. While each school had differences in the taught curricula, course duration, student entry requirements, and assessment scoring criteria, the study promoted critical reflection on curriculum areas that potentially need greater emphasis or development. As such, the comparison and the shared learning arising from the study were useful for educational quality improvement in each school. Although each school was tasked with the editorial lead of one of the OSCEs, this did not particularly benefit students of the corresponding school. Origin of the test material written had no effect on the resulting performance of a school and this contrasts with a previous study of written assessment (Muijtjens et al, 2008).

Overall, the time and expense spent in bringing the ACCLAiM committee together was beneficial to the process of ensuring that shared items were relevant to each school's curriculum, and fitted in with each school's overall OSCE blueprint. Retaining local processes of administration and examiner training in conducting the hybrid OSCE of shared and local content facilitated uptake by the wider staff group of each medical school and allowed the necessary academic rigour of each school's assessment process to continue without external interference. Furthermore, the collaboration process has provided the participating schools with valid quality-assured data on the competency of their students in key clinical areas. This

data can be used to identify strengths and weaknesses in curricula, teaching and learning, as well as assessment processes. In addition, it can be used to demonstrate robust assessment processes, including national standard setting, to external stakeholders.

This research has served primarily as a learning exercise, much more than an outcome measurement on curriculum effectiveness. It has aided participating schools to learn how to align test materials, standards, assessors, information analysis, scores interpretation and meaningful utilisation of analysed data, and most importantly, how to learn from each other's assessment practices. The flip side of the coin is that it is difficult to use these data as sole and absolute benchmarks of effective curricula. Other unaccounted sources of random variation and unavoidable side effects need to be carefully considered and monitored, thus necessitating the need for further studies, which our group is undertaking. Further studies could explore the use of G-studies to provide more data on inter-rater reliability and reproducibility of the scores over time.

McCrorie and Boursicot (2009) suggested that national qualifying level examinations should be considered to ensure formal quantitative comparisons of clinical competence. However, Australian Universities Quality Agency (2009) reported that increased formal standards will reduce the incentive for institutions to develop new methods of teaching, new curricula and general improvements to their operations. They stated that over time, this will damage the sector rather than enhance it. We believe that we are still far away from an absolute use of instruments for the purpose of benchmarking and there is still scope to learn new things about teaching, learning and assessment methods. In addition, the evidence from this current study supports the proposal for increased use of shared assessment by medical schools for quality assurance purposes and also in order to provide a more robust assessment system of clinical competence. The process also serves as a diagnostic tool for improvement

of learning and teaching. Given that OSCEs are expensive to run and developing high quality test material is resource intensive, it may be misconstrued that including shared OSCEs for quality assurance and benchmarking would make the assessment process more complex and expensive to organise. However, we argue that the benefit of this process far outweighs any cost implications with the added value of collective development and sharing of high quality assessment. This brings increased validity, accountability and insights to the curriculum and assessment procedures.

Limitations

As this was a new collaboration, student scores were collected over only two years, however, the volume of data collated and the observed high reliability indices confirm the construct validity of the assessment. Furthermore, the DIF studies could have been confounded by the choice of role player and examiner (both across circuits and across sites), although standardised on-line examiner training and thorough role player training sessions were conducted at all examination sites. There were also minor differences between schools in the implementation of the examination, in relation to timing and organisation of examinations. These limitations are expected to be remedied as this collaboration continues to improve in subsequent years.

Conclusion

This research demonstrates the validity of the psychometric data and benefits of evaluating clinical competence across medical schools without the enforcement of a prescriptive national curriculum or assessment. This study has a significant educational impact as it supports the use of shared OSCEs by medical schools to benchmark clinical competence, providing a more robust yet flexible assessment system. It demonstrates that

sharing of assessment materials can provide common, defensible, reliable, valid, robust and standardised assessments which in turn, enhance transparency and accountability. The economic benefits and collective wisdom gained by such collaboration provide ample justification for its ongoing application.

Acknowledgements

The authors would like to thank the administrative and academic staff from all participating schools, who supported the organisation and implementation of the examinations.

Declaration of interest: The authors report no conflicts of interest.

Ethical approval: All participating schools obtained ethics approval from their local Ethics Committee. All information was de-identified before data analysis.

Practice points

- Collaborative OSCE development by medical schools can enable benchmarking of key nationally required clinical competencies.
- The sharing of assessment materials can provide common, defensible, reliable, valid, robust and standardised assessments which in turn, enhance transparency and accountability.
- Improving assessment practice through cross-institutional collaboration fosters development of communities of practice, rich learning experiences and serves as a diagnostic tool.

Glossary

Clinical Competence: The mastery of relevant knowledge and the acquisition of a range of relevant skills at a satisfactory level including interpersonal, clinical and technical components at a certain point of education, i.e., at graduation.

Reference: Wojtczak, A. 2003. Glossary of Medical Education Terms. AMEE Occasional Paper No 3. Dundee: AMEE

Notes on contributors

BUNMI MALAU-ADULI, BSc, MSc, PhD, is a Senior Lecturer in Medical Education at the College of Medicine & Dentistry, James Cook University

PETA-ANN TEAGUE, MBChB, DRCOG, MRCGP, Dip Med Ed, FRACGP, is an Associate Professor and the Lead Clinician at UniHealth Medical Centre and the Director of Clinical Studies at the College of Medicine and Dentistry, James Cook University

RICHARD TURNER, MBBS, BMedSc, FRACS, PhD, is a Professor of Surgery and the Associate Head of the Hobart Clinical School at the School of Medicine, University of Tasmania

BEN HOLMAN, BSc. PhD, is a Research Officer at New South Wales Department of Primary Industries, Australia

KAREN D'SOUZA, MBBS, FRACGP, is a Senior Lecturer in Medical Education (Clinical Skills) and the Coordinator, Doctor and Patient Theme for the School of Medicine at Deakin University

DAVID GARNE, MBChB, DCH, MIPH, is an Associate Professor and the Associate Dean of Community, Primary, Remote and Rural Health at the Graduate School of Medicine at the University of Wollongong

CLARE HEAL, MBChB DRACOG, FRACGP, MPH&TM, Dip GU Med, PhD, is an Associate Professor of General Practice and Rural Medicine for James Cook University in Mackay

PAULA HEGGARTY, MBBCh, FRACGP, is a Senior Lecturer and the Academic Coordinator for the Year 4 MBBS Program at the College of Medicine, James Cook University

JUDITH NICKY HUDSON, BSc, MSc, MBBS, PhD, is a Professor and Director of Rural Health at the University of Newcastle, and Chair of Assessment for the Joint Medical Program (JMP) of the Universities of New England and Newcastle

IAN WILSON, MBBS, PHD, MASSESS&EVAL, FRACGP, is a Professor, Dean and Head of School at the Graduate School of Medicine at the University of Wollongong

CEES VAN DER VLEUTEN, MA, PhD, is the Scientific Director of the Graduate School of Health Professions Education at Maastricht University

References

Australian Medical Students' Association. 2010. Policy Document: National barrier exam (online) February 2010. Retrieved from http://www.amsa.org.au/sites/default/files/Policy-National%20Barrier%20Exam.pdf.

Australian Universities Quality Agency. 2009. Setting and Monitoring Academic Standards for Australian Higher Education. Retrieved from http://repository.unimelb.edu.au/10187/8770.

Brailovsky CA, Grand'Maison P, Lescop J. 1992. A large-scalemulticenter objective structured clinical examination for licensure. Acad Med 67(10 Suppl.), pp. S37–S39.

Bond TG, Fox CM. 2012. Applying the Rasch model: fundamental measurement in the human sciences. Mahwah NJ: Lawrence Erlbaum Associates; 2nd Edition.

Boursicot KAM, Roberts TE, Pell G. 2006. Standard setting for clinical competence at graduation from medical school: A comparison of passing scores across five medical schools. Adv Health Sci Educ 11(2):173-183.

Chapius P, Fahrer M, Eizenberg N, Fahrer C, Bokey L. 2010. Should there be a national core curriculum for anatomy? ANZ J Surg 80(7-8):475-477.

Chesser A, Cameron H, Evans P, Cleland J, Boursicot K, Mires G. 2009. Sources of variation in performance on a shared OSCE station across four UK medical schools. Med Educ 43(6):526-532.

Downing SM. 2004. Reliability: on the reproducibility of assessment data. Med Educ 38:1006-1012.

Goldacre MJ, Lambert T, Evans J, Turner G. 2003. Preregistration house officers' views on whether their experience at medical school prepared them well for their jobs: national questionnaire survey. Brit Med J 326(7397):1011-1012.

Gustafson JE. 1980. Testing and obtaining fit of data to the Rasch model. Brit J Math Stat Psychol 33:220.

Hagquist C, Andrich D. 2004. Is the sense of conherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. Personality and Individual Diffrences, 36:955-968.

Harden, R. M. 2009. Five myths and the case against a European or national licensing examination. Med Teach 31(3), 217–220.

Harden RM, Crosby JR, Davis MH. 1999. AMEE Guide No.14: Outcome-based education: Part 1—an introduction to outcomebased education. Med Teach 21:7–14.

Kane, M. 2006. Content-related validity evidence in test development. In S. M. Downing, T. M. Haladyna (Eds.), Handbook of test development (pp. 131-153). Mahwah, NJ: Lawrence Erlbaum Associates.

Linacre, J. M. 2009. Winsteps (Version 3.68) Computer Software. Beaverton, Oregon: Winsteps.com.

Malau-Aduli BS, Mulcahy S, Warnecke E, Otahal P, Teague PA, Turner R, Van der Vleuten C. 2012. Inter-rater reliability: Comparison of checklist and global scoring for OSCEs. Creative Educ J, Special Issue, 3: 937-942. DOI:10.4236/ce.2012.326142

Marais I, Andrich D. 2008. Effects of varying magnitude and patterns of local dependence in the unidimentionsional Rasch model. J Appl Measurement 9(2): 105-124.

Masters GN. 1982. A Rasch model for partial credit scoring. Psychometrika, 47:149-174.

McCrorie P, Boursicot K. 2009. Variations in medical school graduating examinations in the United Kingdom: Are clinical competence standards comparable? Med Teach 31, 223-229.5.

McGrath BP, Graham IS, Crotty BJ, Jolly BC. 2006. Lack of integration of medical education in Australia: the need for change. MJA 184(7):346-348.

Medical Council of Canada. 2002. Qualifying Examination Part II, Information Pamphlet Ottawa, Ontario, Canada, Medical Council of Canada.

Medical Deans Australia and New Zealand. 2011. Developing a National Assessment Blueprint for Clinical Competencies for the medical graduate. Competencies project stage 3. Retrived from http://www.medicaldeans.org.au/wp-content/uploads/Medical-Deans-Competencies-Project-Stage-3- Final-Report-FINAL.pdf.

Muijtjens AMM, Schuwirth LWT, Cohen-Schotanus J, Thoben AJNM, van der Vleuten CPM. 2008. Benchmarking by cross-institutional comparison of student achievement in a progress test. Med Educ 42(1):82-88.

Norcini JJ, Holmboe ES, Hawkins RE. 2008. Evaluation challenges in the era of outcomes-based education. In E.S. Holmboe, R.E. Hawkins (Eds.) Practical Guide to the Evaluation of Clinical Competence. pp 1–9. Philadelphia, PA: Mosby/Elsevier.

Petrusa ER, Blackwell TA, Carline J, Ramsey PG, McGaghie W, Colindres R, Kowlowitz V, Mast TA, Soler N. 1991. A multi-institutional trial of an objective structured clinical examination. Teach Learn Med 3(2):86-94.

Roberts C, Newble D, Jolly B, Reed M, Hampton K. 2006. Assuring the quality of high-stakes undergraduate assessments of clinical competence. Med Teach 28(6):535-543.

Turner JL, Dankoski ME. 2008. Objective structured clinical exams: A critical review. Family Med 40(8):574-578.

Whelan GP. 1999. Educational commission for foreign medical graduates: clinical skills assessment prototype. Med Teach 21(2):156–160.

Wilkinson D. 2010. Medical lesson from a Russian doll: common core curriculum. The Australian. Retrieved from http://www.theaustralian.com.au/news/health-science/medical-lesson-from-a-russian-doll-common-core-curriculum/story-e6frg8y6-1225834513820.

Wilkinson D, Schafer J, Hewett D, Eley D, Swanson D. 2014. Global benchmarking of medical student learning outcomes? Implementation and pilot results of the International Foundations of Medicine Clinical Sciences Exam at The University of Queensland, Australia. Med Teach 36(1):62-67.

Wright BD, Masters GN. 1981. The measurement of knowledge and attitude (Research Memorandum No 30). Chicago: University of Chicago, MESA Psychometric Laboratory

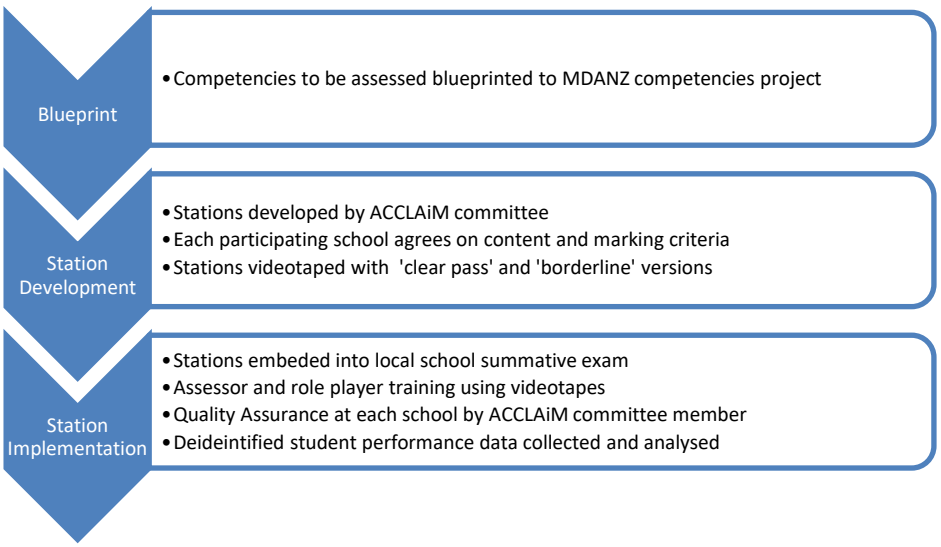

Figure 1: Process of Development and Implementation of Shared OSCE Cases

Table 1: Reliability Measures

Measurement	2011 EC	2012 EC	2012 EE
Person reliability index (PRI) is similar to Cronbach's alpha, although usually smaller. It indicates the reproducibility of person ordering, if same sample of persons are given a parallel set of items which measure the same construct.	0.61	0.73	0.70
Person separation index (PSI) indicates the efficiency of the items in separating the persons measured.	1.25	1.63	1.52
Item reliability index (IRI) is a measure of the consistency of inferences made on item difficulty.	0.95	0.95	0.95
Item separation index (ISI) indicates the efficiency of the sample of persons in separating the items used.	4.31	4.58	4.30

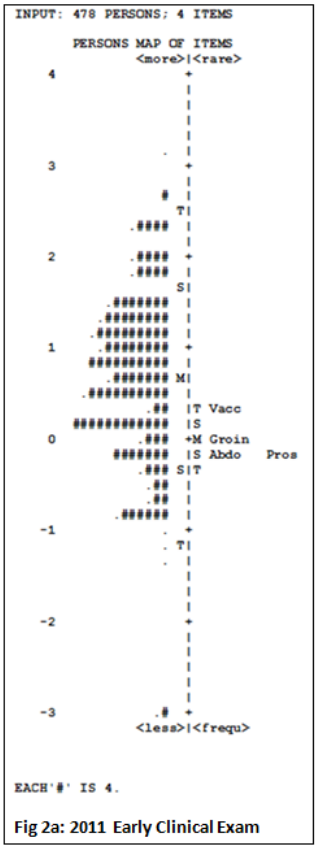

Fig 2a: 2011 Early Clinical Exam

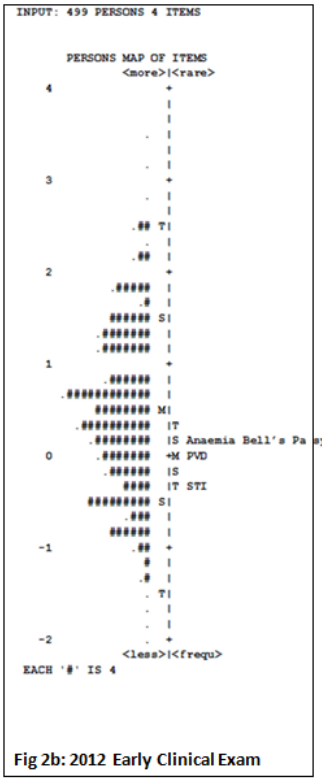

Fig 2b: 2012 Early Clinical Exam

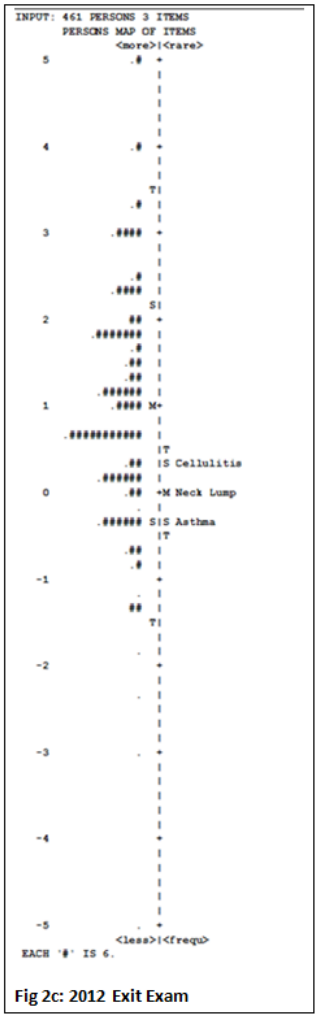

Fig 2c: 2012 Exit Exam

Figure 2: Item-Person Maps

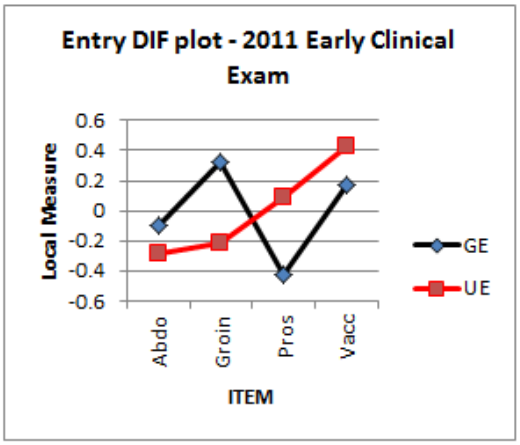

Fig 3a: DIF for Entry program - 2011 EC

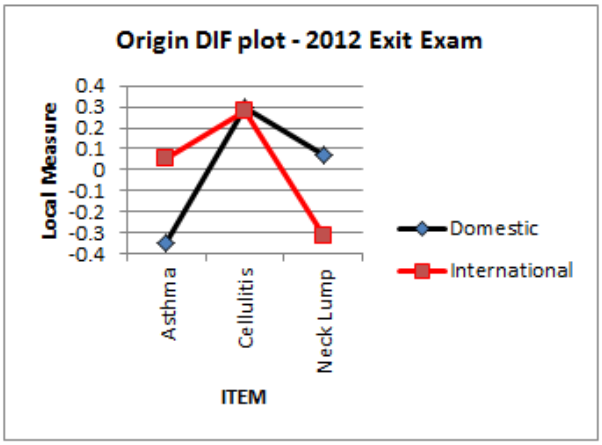

Fig 3b: DIF for Origin - 2012 EE

Figure 3: Differential Item Functioning by entry program and origin

Appendix 1: Description of the Shared OSCEs

2011 Early Clinical Examination		
Station Title	**Description**	**Assessed Competencies**
Abdominal Pain	Focused and relevant clinical GI examination. Formulation of differential diagnoses and first line management	Physical examination, diagnostic, and management skills
Groin Pain	Focused history and examination of patient with possible inguinal hernia or groin strain	History taking, physical examination, and communication skills
Prostate Cancer	Focused history of patient with difficulty passing urine	History taking, diagnostic, and communication skills
Vaccination	Vaccination general and specific knowledge of pertussis	History taking and communication skills, and knowledge of immunisation: Basic science and population health aspects
2012 Early Clinical Examination		
Station Title	**Description**	**Assessed Competencies**
Anaemia	Interpret abnormal full blood count in a 56 year old man, focused history taking to establish likely cause of anaemia	History taking skills, Interpretation of abnormal haematological Investigations and formulation of differential diagnoses
Bell's Palsy	Targeted neurological examination of cranial nerves 6-12, interpret diagram of patient's face and give most likely diagnosis	Physical examination, interpretation and formulation of differential diagnoses
Peripheral Vascular Disease	Focused history and vascular examination of right leg, diagnosis and investigation plan	Physical examination, history taking, formulation of differential diagnoses and investigation plan
Sexually Transmitted Infections	Focused sexual history for vaginal discharge and appropriate investigations	Sexual history & communication skills, differential diagnosis and investigation plan
2012 Exit Examination		
Station Title	**Description**	**Assessed Competencies**
Asthma	Management plan for severe asthma	Communication skills and management plan
Cellulitis	Patient with severe cellulitis of the right lower limb is being discharged home to a nearby Aboriginal community. Arrange discharge plan, which includes a telephone call to discuss the case with a health worker at the local clinic	Interprofessional communication, patient-centred care, and cultural competency / safety
Neck Lump	Perform a systematic head and neck examination, with a thyroid focus on a patient who thinks she may have felt a lump in her neck and seeks reassurance	Physical examination, communication and diagnostic skills, investigation plan