Benchmarking the attainment of clinical competencies in Australian medical schools: an innovative collaboration

Final report 2016

Lead Institution
James Cook University

Partner Institutions
University of Tasmania, Deakin University, University of Wollongong, Griffith University, The University of Newcastle

Project Leader
Associate Professor Peta-Ann Teague

Project Team
Dr Bunmi Malau-Aduli, Professor Richard Turner, Professor Ian Wilson, Associate Professor David Garne, Dr Karen D’Souza, Professor Nicky Hudson, Associate Professor Clare Heal, Dr Ian Kerr

https://acclaim.edu.au/
Support for the production of this report has been provided by the Australian Government Office for Learning and Teaching. The views expressed in this report do not necessarily reflect the views of the Australian Government Office for Learning and Teaching.

With the exception of the Commonwealth Coat of Arms, and where otherwise noted, all material presented in this document is provided under Creative Commons Attribution-ShareAlike 4.0 International License http://creativecommons.org/licenses/by-sa/4.0/.

The details of the relevant licence conditions are available on the Creative Commons website (accessible using the links provided) as is the full legal code for the Creative Commons Attribution-ShareAlike 4.0 International License http://creativecommons.org/licenses/by-sa/4.0/legalcode.

Requests and inquiries concerning these rights should be addressed to:
Office for Learning and Teaching
Department of Education and Training

GPO Box 9880,
Location code N255EL10
Sydney NSW 2001

<learningandteaching@education.gov.au>

2016

ISBN 978-1-76028-742-9 [PDF]
ISBN 978-1-76028-743-6 [DOCX]
Acknowledgements

On behalf of the ACCLaiM collaboration I would like to thank and acknowledge the following people:

Professor Cees van den Vleuten, University of Maastricht, and Associate Professor Kathy Boursicot, Lee Kong Chian School of Medicine, for their support and advice;

Dr Ian Kerr whom represented Griffith University at the time of the OLT application, for his support and advice;

Mrs Yolanda Kerlen-van der Kruk and Ms Katie Lee for their outstanding administrative support;

Dr Bunmi Malau-Aduli for her support, advice and statistical expertise;

Dr Paula Heggarty, College of Medicine and Dentistry James Cook University, for coordinating the videoing of all the stations;

The “med tech” team at James Cook University for videoing and editing all the stations;

and not least the partner universities who joined the collaboration during the course of the project:

Bond University

The University of Queensland

University of Western Sydney

University of Notre Dame Australia

The University of Adelaide

Monash University

The University of Sydney

Australian National University
### List of acronyms used

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCLAiM</td>
<td>Australian Collaboration for Clinical Assessment in Medicine</td>
</tr>
<tr>
<td>MDANZ</td>
<td>Medical Deans of Australia and New Zealand</td>
</tr>
<tr>
<td>OSCE</td>
<td>Objective Structured Clinical Examination</td>
</tr>
</tbody>
</table>
Executive summary

Project Context
There were two separate but related contexts for this project:

1. The assessment of clinical competencies of Australian medical students;
2. The debate about whether medical students should undertake a national exit or licencing exam, similar to the US model.

These two contexts provided the focus of this project which was to trial an innovative way of enabling Australian medical schools to collaborate in benchmarking and quality improvement in the assessment of medical students’ clinical competence while maintaining the integrity of individual medical schools’ assessment processes.

Project Aim
The aims of this project were:

1. To develop an academic standards framework to monitor the attainment of clinical competencies by Australian medical students. This was to be done by sharing clinical assessment materials in a collaborative and flexible way, with online examiner training, in order to improve the validity and reliability of summative Objective Structured Clinical Examination (OSCE) in a national context.

2. To foster the evaluation of assessment practices at the participating schools, thereby leading to improved student outcomes.

These aims were to be achieved in a cost effective way.

Project outputs, deliverables and resources
The project has achieved all the outputs set out in the original proposal and project extension.
Phase 1: Development of secure website, obtaining ethical approval and invitation to all medical schools

- Ethics approval was obtained for all participating medical schools (JCU Human Ethics Approval HS246, attached as Appendix B) and extended to cover the project extension (attached as Appendix C).
- All medical schools in Australia and New Zealand were contacted to invite them to join the collaboration. As of April 2015 14 medical schools are active participants and a further two are joining shortly, giving a total of 16 out of a possible 19 Australian medical schools. A representative from one of the New Zealand schools has also attended meetings.
- A secure website was established by January 2014 and is in frequent use. It is located at https://acclaim.edu.au/

Phase 2: Blueprinting of assessment items for November 2013 and June 2014 exams (extended to include November 2014 exam period)

- Key clinical competencies were identified by the group and OSCE stations were collaboratively developed.
- By December 2014 28 OSCE stations had been developed; 35 will have been developed and implemented by the end of 2015.

Phase 3: Development of OSCE videos

- OSCE stations were filmed using simulated patients and junior doctors who role played the student (candidate). Two iterations of each OSCE stations were filmed: one showing the candidate performing at the expected level and the second iterations showing a borderline candidate’s performance.
- These videos were then loaded onto the ACCLAiM website to be available for examiner training.

Phase 4: Implementation

- The collaboratively developed OSCEs were embedded in summative clinical examinations at participating medical schools.
✓ Each participating school had an external examiner (from one of the partner schools) attend their clinical exam and provide a QA report on the exam.
✓ De-identified student performance data from each school was collected and analysed. Pooled data provided information about which competencies that students were performing below, at or above expected standards.

**Phase 5: Dissemination**

✓ Formal external dissemination about the project and its’ outcome has occurred in a variety of settings including: Medical Education conferences, workshops and meetings, and publications.
✓ Formal internal dissemination has occurred via presentation of data at the biannual face to face meetings and the confidential individual reports to participating schools.
✓ Informal dissemination has also occurred at many points, and many clinicians and clinical academics at the participating schools are now aware of the project and familiar with the outcomes.
✓ A full list of the formal dissemination is available as Appendix D.

**Project Outcomes**
The project has had a significant educational impact. It has enabled the development of a ‘community of practice’ of medical educators in Australia which is focused on developing best practice in assessment. Secondly, it has provided shared, defensible, reliable and valid standardised assessment items which have produced useful data which is being used to inform curriculum and assessment in the participating medical schools.

**Key Finding**
This collaborative approach to assessment works and is suitable for any discipline that assesses students’ clinical competencies.
# Table of contents

Acknowledgements.................................................................................................................... 3
List of acronyms used .................................................................................................................... 4
Executive summary....................................................................................................................... 5
Project Context ............................................................................................................................ 5
Project Aim .................................................................................................................................. 5
Project outputs, deliverables and resources .............................................................................. 5
  Phase 1: Development of secure website, obtaining ethical approval and invitation to all medical schools .................................................................................................................. 6
  Phase 2: Blueprinting of assessment items for November 2013 and June 2014 exams.... 6
  Phase 3: Development of OSCE videos ............................................................................... 6
  Phase 4: Implementation ................................................................................................... 6
  Phase 5: Dissemination ...................................................................................................... 7
Project Outcomes ...................................................................................................................... 7
Key Finding ................................................................................................................................... 7
Tables and figures ..................................................................................................................... 10
Figures ...................................................................................................................................... 10
Narrative .................................................................................................................................. 11
Chapter 1:  The Context of the Project .................................................................................... 11
Chapter 2: Approach and Methodology .................................................................................. 13
  Overview .......................................................................................................................... 13
  Approach (as outlined in the project application) .............................................................. 14
  Summary of statistical analysis methodology ................................................................. 18
  Phase 5: ............................................................................................................................ 21
Chapter 3: Project Outcomes and Findings ............................................................................. 22
Tables and figures

Figures

Figure 1  The process of the shared OSCE station development and implementation........ 17
Figure 2a  Person-Item Map for 2014 Early Clinical Exam.......................................................... 26
Figure 2b  Person-Item Map for 2014 Exit Exam........................................................................ 27
Figure 3  Student performance across three 2014 early clinical exams............................... 28
Narrative

Chapter 1: The Context of the Project

Australia has a world-class medical education system with a rigorous process of medical school accreditation by a national body, the Australian Medical Council (AMC) [1]. However, all Australian medical schools have different curricula and assessment procedures. Concerns about junior doctor competencies [2-4], has led to a call for greater integration of medical education in Australia [5]. In addition, there has been a move towards outcomes-based medical education with identifiable core competencies [6-8] 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 [9].

Two major proposals have emerged worldwide in response to these pressures.

This first proposal suggests a national medical education curriculum with a mandatory national exit exam to allow medical school benchmarking [7, 10]. 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 [11]. In Australia, a rigid national assessment may not be readily applied across the diversity of medical schools. The vast geography, population density differences and local cultural variation between the centres in Australia hosting medical courses contribute to the diversity in curricula and student experience.

The second proposal promotes voluntary assessment collaboration between medical schools [12]. 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 [7, 13, 14]. The applicability of either proposal merits research.

In Australia there has been a lack of data on the attainment of clinical competencies by medical graduates. OSCEs are widely used at important clinical assessment checkpoints [17-20] and this project hypothesised that benchmarking Australian medical school assessment standards would prove valuable in establishing the attainment of clinical competencies [5, 21], as it has done for written assessment [13]. The project also proposed to provide individual medical schools with feedback about the strengths and weaknesses of their clinical curriculum, teaching and assessment.
This research project was therefore designed and conducted with the following intertwined aims:

1. To develop an academic standards framework to compare and monitor clinical competence across Australian medical schools. This was to be done by sharing clinical assessment materials in a collaborative and flexible way, with online examiner training, in order to improve the validity and reliability of summative OSCEs in a national context.

2. To foster the evaluation of assessment practices at the participating schools, thereby leading to improved student outcomes.

These aims were to be achieved in a cost effective way.
Chapter 2: Approach and Methodology

Overview

A key element of this project was to set up processes which were inclusive of all participants and were simple and cost effective. These processes are outlined below.

1. At the outset of the project, an **Executive Committee** was established, comprising the academic lead from each of the collaborating institutions as well as Dr Bunmi Malau-Aduli (BMA). This committee met monthly by teleconference. A project administrator was appointed and she works closely with the two lead academics (Associate Professor Peta-Ann Teague (PAT) and Dr Bunmi Malau-Aduli) as well as with the executive. She is also in regular contact with all academic project leads at the participating medical schools and has attended the face to face meetings.

2. A **secure website** was developed early in the project. This allowed for the geographically dispersed participants to readily keep up to date with developments and participate in finalising the OSCE stations, as well as in accessing the training videos.

3. The successful pilot project demonstrated the efficacy of **biannual face to face meetings** with representatives from each participating school and these were continued during the project. Initially these meetings were held at a Qantas Club meeting room with the schools taking it in turns to cover the venue and catering costs. As the collaboration has grown in size the meetings have been transferred to medical school sites. The model of rotating the venue and of the hosting school to meet the catering costs has continued.

4. **Data Collection templates** were developed specifying the student performance data required from schools.

5. The **visiting (external) examiner system** which was part of the pilot project was enthusiastically adopted by the project participants and is reported on below. Schools identify the site and exam day that it wishes to have a visiting examiner at. The external examiner observes the running of the exam, the examiner and role player performance as well as the student experience. A confidential QA report is subsequently provided to the school. Each school meets the cost of one of its clinical academic staff taking on the visiting examiner role.
Approach (as outlined in the project application)

Phase 1: Development of secure website, obtaining ethical approval and invitation to all medical schools

- **Ethics approval** was obtained for all participating medical schools (JCU Human Ethics Approval H5246, attached as Appendix B) and extended to cover the project extension (Appendix C). As each new school joined, a copy of the JCU Ethics approval was provided to it and local Ethics approval was obtained.

- **All medical schools** in Australia and New Zealand were formally contacted to invite them to join the collaboration. There have also been frequent informal contacts with all medical schools at different medical education fora. As of April 2015 14 medical schools are active participants and a further two are joining shortly, giving a total of 16 out of a possible 19 Australian medical schools. A representative from one of the New Zealand schools has also attended the face to face meetings.

- **A secure website** was established by January 2014 and is in frequent use. It is located at [https://acclaim.edu.au/](https://acclaim.edu.au/). The website has three different levels of access.

  - **A public site** which is open to anyone and which briefly details the history and aims of the collaboration as well as the current participating medical schools;
  - **Member only** access point which allows the lead academics from each partner school to access:
    - the OSCE site where current and previous OSCEs and mark sheets are located (OSCEs under development have a lead clinician who has version control of the OSCE);
    - the training videos;
    - the minutes from previous face to face meetings;
    - the Visiting Examiner (QA) timetable and report template;
    - the data collection templates;
    - psychometric data from previously used stations.

  - **Executive Committee** access which additionally enables the Executive Committee members to look at meeting minutes and track dissemination activities.
Phases 2, 3 and 4: Blueprinting and implementation of assessment items for November 2013 and June and November 2014 summative exams

The process of blueprinting and development of the OSCE stations comprised much of the time in the face-to-face meetings. Participants divided into two groups with one group focusing on developing OSCEs for the Exit Exam and the other for the Early Clinical Exam. The MDANZ medical student attributes spectrum was developed in 2011 [1] and the ACCLAiM committee drew on it to identify appropriate clinical competencies that should be assessed during the project. Figure 1 summarises the process of OSCE station development and implementation.

A blueprint was developed to track which competencies were assessed in each year, thereby ensuring that as large a number of competencies as possible were assessed.

The competencies were grouped into six domains:

1. Communication Skills (CS)
2. History taking Skills (HTS)
3. Examination Skills (ES)
4. Procedural Skills (PS)
5. Diagnostic Skills (DS)
6. Management Plan (MP)
This grouping allowed for clear identification of what was being assessed on the mark sheets. The use of different locally developed mark sheets by each school posed a significant challenge to data analysis and so the development of an ACCLAiM mark sheet for each ACCLAiM OSCE was an important step in allowing for better analysis of performance data. This is discussed further below. A copy of an ACCLAiM mark sheet is attached as Appendix E.

By December 2014 28 OSCE stations had been developed and implemented and a further 7 stations have been developed for implementation in 2015. A copy of a written ACCLAiM OSCE is shown in Appendix F.

The process of identifying the competencies, developing and implementing the OSCEs and discussing the results was observed on two occasions by external evaluators. Professor Cees van den Vleuten attended the face-to-face meeting on April 4th 2014 and provided educational input to as well as informal feedback about the project. His comments are attached in appendix G.

Associate Professor Kathy Boursicot attended the face-to-face meeting on April 4th 2014 and her report is attached as Appendix H.

**Examiner training**

There are a number of well recognised variables which can impact on the validity and reliability of an OSCE station as an assessment tool. A key factor of reliability in an OSCE is the judgment made by the examiners [2,3]. This becomes a paramount issue when a typical OSCE station is solely reliant on a single examiner.

Many medical schools run examiner training events prior to the OSCE. The data regarding the efficacy of examiner training is limited and contradictory. There is a suggestion that training may not be helpful [4] but others have shown that training of and feedback to examiners about their performance can be helpful in improving inter examiner reliability [5,6].

The challenge of ensuring inter examiner consistency in summative OSCEs has become increasingly complex with the increase in student numbers and the multi-site nature of many medical schools, where students at a single medical school simultaneously sit the same OSCE at different and often remote geographical sites. For reasons of exam security, each site must hold these exams on the same day commencing at exactly the same time. Furthermore, increasing student numbers means that the ideal of having two assessors at one station is a practical impossibility.
This scale of assessment requires large numbers of clinicians as examiners. In regional and remote areas, this is a significant task. In order to achieve the required number of examiners, clinicians from both the private and public sectors are used as assessors. Many of these clinicians, although committed in principle to supporting the training of medical students, do not teach into the medical programme regularly and often lack background knowledge about assessment and curriculum. As all the assessors are practicing clinicians, time available for training is limited and as a consequence not all assessors have ‘ownership’ of the process. Geographical distance of staff precludes all examiners of a single OSCE station meeting face-to-face beforehand to discuss it, and the fact that assessors will be taking a day out of clinical practice (and often longer if there is travel involved) means even more time pressure when considering availability for pre-OSCE training.

In the pilot project the use of online examiner training for shared OSCEs demonstrated that inter-examiner variation was markedly less in the shared stations than in stations where there was no online examiner training and where global rating scores were used [7].

The ACCLAiM project has therefore continued a focus on examiner training using videos. Two iterations of each OSCE station are filmed. Simulated patients and junior doctors are trained in their roles, with the junior doctor playing the role of the student being assessed. The first version of the station has the ‘student’ performing at the expected standard. The second version shows a borderline candidate. These videos are loaded onto the website and are available for examiner training prior to the actual exam. Examiners are blinded to the versions and mark the station using their school’s mark sheet. Examiners then have the opportunity to reach agreement in the expected standard.

The use and efficacy of the training videos is discussed further below.

External Examiner Visit and Report

Medical education is a public trust and whilst medical educators have always sought the best methods for formative and summative evaluation of trainees, the renewed emphasis on patient safety and quality outcomes in the social consciousness necessitates that medical educators use high-quality, reliable, valid and educationally sound assessment methods [1].

In this project, the use of the shared OSCEs provides a dataset on student performance in the national context but not on the clinical exam as a whole. In order to address this, as part of the project each participating school was offered the option of a visiting examiner
who would attend and observe one of the clinical exams and write a confidential report on it.

One of the Executive Committee members (PAT) developed a guide for QA examiners as well as a template for the report (Appendices I and J). This confidential report is then submitted to the school. The reports are held separately to other ACCLAIaM data and there is no open access to them.

The impact of these External Examiner reports is discussed below.

Collection and analysis of student performance results

Student performance data was collected from each participating school. The school de-identified the data prior to sending it for analysis. The school data sets are held securely with access to the data available only to two people: the project administrator and BMA who has undertaken the statistical analysis.

Summary of statistical analysis methodology

Students’ total (checklist and global) scores were collated for data analysis and they were converted into z-scores to standardise scores from the different data groups. The z-score enables comparison of scores that are from different normal distributions [8]. The dataset was analysed using Statistical Analysis System (SAS) software [9]. Descriptive summary statistics were calculated and means, standard deviations, and minimum and maximum values for each station were assessed for transcription errors and outliers. Two-way ANOVA analysis was then run with school, shared OSCE station and their second-order interaction fitted as fixed effects and total z scores as dependent variables. To minimise Type I errors, the level of significance was set as $P<0.001$ and Duncan’s multiple range tests were used in identifying differences.

Evidence for validity was drawn from five areas to support confidence in the inferences made from assessment: test content; response process; internal structure of test data; correlational analyses; and effects of assessment [10, 11]. The evidence used to validate the assessment data were test content, response process, internal structure of test data and impact/benefit of the benchmarking assessment process on curriculum. Evidence for test content was obtained through the use of items which were chosen from prospectively reviewed blueprints of the specific clinical skills and medical problems which represent a fair and reasonable assessment and which mapped to the MDANZ medical competencies project. Evidence for data management was obtained through the use of on-line examiner training to standardise marking as well as quality control and validation of scores. Partial Credit Rasch Model (PCRM) [12] and Generalisability Theory (G-theory) [13] were used to
provide evidence of internal structure for the OSCE data i.e. to validate the psychometric data. Consequential evidence for the beneficial impact of the benchmarking process was obtained by the judgmental evaluation of the quantitative data and its impact on teaching and learning.

To answer the question as to what extent the OSCE performance data form a unidimensional and locally independent construct according to the Rasch measurement 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 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 [14]. 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 [12]. The raw scores for each station were first collapsed into 10 categories – zero to nine, to be fitted into the PCRM, using the Winsteps software [15]. 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 [16]. 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 [12] e.g. clinical competence. A measure of person separation (PSI) is calculated to indicate the efficiency of the items in separating the persons measured [12]. The item measure reliability (IRI) indicates how well the scale can distinguish between items, on the basis of their difficulty [12], 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 [17]. 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 [17].

**Differential Item functioning (DIF)** tests measurement invariance by detecting test items biased towards different subgroups of test takers according to construct irrelevant factors [18]. 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 [12].

When analysing the data, the second question that was asked was what the estimation of the reproducibility of the observed test scores was. To address this question generalisability analysis was used to measure the reproducibility of the observed OSCE scores by evaluating the different sources of variation affecting the measurement. The magnitude of the variations was also calculated using analysis of variance and variance component estimation. In addition, the estimated variance components were also used for decision studies; to estimate generalisability coefficients (G coefficients) i.e. reproducibility of scores as a function of examination cases. Variance components for each station within each school were estimated separately and the estimates were pooled across sites in a nested OSCE cases by persons by schools design (o:p:s). Nesting of stations was appropriate because not exactly the same stations were used across schools and with schools several circuits of stations were used.

**Results**

The results have been presented in a variety of settings, including at the biannual face-to-face ACCLAiM meetings as well as in workshops and publications as described in the dissemination list (Appendix D). Each participating school receives a confidential report where their school is identified. This is submitted to the school along with the relevant External Examiner report/s. The 2013 data has been published in *Medical Teacher* [2] (Appendix K). The impact of the confidential reports as well as the external examiner reports is currently being formally evaluated by the ACCLAiM executive.

The data and process of dissemination is discussed further below.
**Phase 5: Dissemination**

The approach to this has been three pronged:

- **Formal dissemination** via conferences, workshops, medical education meetings and publications;
- **Confidential dissemination** to individual participating schools;
- **Informal dissemination** via medical education networks and other settings.

A full list of formal and confidential dissemination activities is available in Appendix D.
Chapter 3: Project Outcomes and Findings

Outcomes

Two related but distinct outcomes have emerged from this project. Firstly, a ‘community of practice’ of medical educators in Australia has emerged and is focused on developing best practice in assessment. Secondly, useful quantitative data has been produced which is being used to inform curriculum and assessment in the participating medical schools.

These two outcomes will be discussed separately.

1. Development of a Community of Practice

Throughout the duration of the project, 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 including:

- standard setting techniques;
- optimal station length and reading time length;
- the ideal means of student debriefing immediately post-OSCE;
- inter assessor reliability.

This sharing of best practice has had a significant impact on participating schools as it has prompted the review of current assessment practices and how these could be improved.

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).

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. 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.

The examiner training videos have been a useful resource to help prepare both assessors and role players for the OSCE. There has been variable uptake of the videos with some schools using them extensively. The schools that used them reported positive feedback from assessors and the opportunity for assessors to calibrate their marking prior to the OSCE, thereby increasing inter examiner reliability.

The External Examiner reports provided to each school after the OSCE exam gave external qualitative data on the overall examination process, the student experience, the preparation and performance of role players and assessors and the level of difficulty of each of the stations. These reports were able to externally identify strengths and weaknesses of each school's OSCE and provide this feedback in a constructive way.

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 benchmarked 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.

The strength of this community of practice as led to the organisation of a national Assessment Masterclass for health professionals who are involved in the assessment of health professional students. The Assessment Masterclass will be held on 2-3rd July 2015 in Cairns. Further information about it can be found at http://www.jcu.edu.au/smd/medicine/events/healthassessmentcourse2015/index.htm
Data Analysis

The analysis of the 2011 and 2012 student performance data is available in the publication of Malau-Aduli et al. (2015) (Appendix K). Both the 2013 and 2014 performance data has been analysed and dissemination of such has taken place at a number of formal and informal settings, including in the confidential individual school reports. Figures 2a and 2b provide an example of such presentations. These figures show the Rasch item–person map of student ability and item difficulty for the 2014 Early Clinical and Exit exams respectively.

Students of greater ability and more difficult stations are towards the top, and students of lesser ability and easier stations are towards the bottom. Students are indicated by #. The OSCE stations are noted on the right hand side. In figure 2a ‘opthal’ = ophthalmology; ‘gastro’ = gastroenterology; resp = respiratory. In figure 2b ‘pre op’ = pre-operative assessment; ‘MSK’ = musculoskeletal; ‘paeds’ = paediatrics.

Further detailed performance data has not been included in this report due to its sensitive nature, but can be made available on request.

Figure 2a. Person-Item Map for 2014 Early Clinical Exam.
There are several discussion points arising from the data, two of which are briefly outlined below.

The first point is that the project has identified that students at all schools tend to have similar strengths and weaknesses in their clinical abilities. In addition, there appears to be no impact of gender on performance. As more schools have joined the collaboration there has been more variation in performance. This was expected and has not impacted negatively on the outcomes as the project has focused strongly on quality improvement of assessment rather than a ‘league table’ approach. Individual schools have thus been able to
reflect on their students’ performance in the context of trends rather than of ‘school A versus school B’.

An example of this is an ophthalmology station that was run in 2014 in the early clinical exam in several schools (as shown in figure 2a). Student performance at this station was unexpectedly varied with many students at all schools experiencing it as a difficult station. Analysis of the results using PCRM confirmed the poor fit as compared to the stations that examined clinical competencies relating to the respiratory and gastrointestinal systems (Figure 3). Figure 3 indicates the fit of the items (construct validity) to the Rasch model, with an acceptable range of -2 to +2. It shows that the ophthalmology station is a misfitting and poorly discriminating assessment item and as such may need revision. This data has allowed schools to revisit their teaching of examination of the eyes and assessment of vision.

Figure 3. Student performance across three 2014 early clinical exams.

The second observation is that the different mark sheets used at the different participating schools, whilst very similar, are different enough to potentially confound the data. In order to address this, we have nested a small arm of the project comparing the use of a single ACCLAiM mark sheet with individual school’s native mark sheets. Assessors double mark an
individual student using both sheets. The locally developed mark sheets counts for the medical school’s assessment processes. This part of project will be reported as a separately when it is complete.

**Factors contributing to the success of the project.**

The key factor that has contributed to the success of the project is that there has been ‘buy in’ from the participants. This is because:

1. The project is relevant to the participants;
2. The project complements work that participants already do;
3. The benefits of the project have been evident from early on.

The project slowly expanded as medical schools were attracted by its’ utility. Assessment is a key part of our work as medical teachers and participation in the project offers access to high quality clinical assessment items that are relevant to the Australian context and that have been subjected to rigorous peer review. It also offers the opportunity for schools to contribute to the development of these assessment items and influence assessment practices nationally.

The project has been very fortunate to have many local champions who have ensured the successful participation of their medical schools.

**Key Finding**

**The key finding is that this approach to assessment works.**

1. Collaborative development and sharing of clinical assessment items is a cost effective way of providing common, defensible, reliable, valid, robust and standardised assessments which in turn, enhance transparency and accountability;
2. This can be done without compromising local processes and decisions on student progression;
3. It can be done without the development of unidimensional league tables.

The true test as to the utility of this project is that it is continuing without OLT funding. Medical schools have adopted it as an important part of their assessment processes and will continue to participate in the project.

It is the opinion of the ACCLAIM collaboration that this approach to assessment can be successfully applied to any discipline that requires clinical assessment of its students.
References

Chapter 1


Chapter 2


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


**Chapter 3**


Appendix A

Certification by Deputy Vice-Chancellor (or equivalent)

I certify that all parts of the final report for this OLT grant provide an accurate representation of the implementation, impact and findings of the project, and that the report is of publishable quality.

Name: .......................................................... Date: 13/5/15
<table>
<thead>
<tr>
<th>Human Research Ethics Committee</th>
<th>Application ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>APPROVAL FOR RESEARCH OR TEACHING INVOLVING HUMAN SUBJECTS</td>
<td>H5246</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PRINCIPAL INVESTIGATOR</th>
<th>Peta-Ann Teague</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCHOOL</td>
<td>Clinical School</td>
</tr>
<tr>
<td>CO-INVESTIGATOR(S)</td>
<td>Bunmi Malau-Aduli</td>
</tr>
<tr>
<td>SUPERVISOR(S)</td>
<td></td>
</tr>
<tr>
<td>PROJECT TITLE</td>
<td>Benchmarking the attainment of clinical competencies in Australian medical schools: an innovative collaboration</td>
</tr>
<tr>
<td>APPROVAL DATE:</td>
<td>29/07/2013</td>
</tr>
<tr>
<td>EXPIRY DATE:</td>
<td>31/07/2014</td>
</tr>
<tr>
<td>CATEGORY:</td>
<td>1</td>
</tr>
</tbody>
</table>

This project has been allocated Ethics Approval Number H5246, with the following conditions:

1. All subsequent records and correspondence relating to this project must refer to this number.

2. That there is NO departure from the approved protocols unless prior approval has been sought from the Human Research Ethics Committee.

3. The Principal Investigator must advise the responsible Human Ethics Advisor:
   - periodically of the progress of the project,
   - when the project is completed, suspended or prematurely terminated for any reason,
   - within 48 hours of any adverse effects on participants,
   - of any unforeseen events that might affect continued ethical acceptability of the project.

4. In compliance with the National Health and Medical Research Council (NHMRC) “National Statement on Ethical Conduct in Human Research” (2007), it is MANDATORY that you provide an annual report on the progress and conduct of your project. This report must detail compliance with approvals granted and any unexpected events or serious adverse effects that may have occurred during the study.

Human Ethics Advisor: Woodward, Lynn

Email: lynn.woodward@jcu.edu.au

This project was Approved by Executive on 29 Jul 2013

Dr Anne Swinbourne

Chair, Human Research Ethics Committee
### Human Research Ethics Committee

#### APPROVAL FOR RESEARCH OR TEACHING INVOLVING HUMAN SUBJECTS

<table>
<thead>
<tr>
<th>Application ID</th>
<th>H5246</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PRINCIPAL INVESTIGATOR</strong></td>
<td>Peta-Ann Teague</td>
</tr>
<tr>
<td><strong>SCHOOL</strong></td>
<td>Clinical School</td>
</tr>
<tr>
<td><strong>CO-INVESTIGATOR(S)</strong></td>
<td>Bunmi Malau-Aduli and Michael Wan</td>
</tr>
<tr>
<td><strong>SUPERVISOR(S)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>PROJECT TITLE</strong></td>
<td>Benchmarking the attainment of clinical competencies in Australian medical schools: an innovative collaboration</td>
</tr>
<tr>
<td><strong>APPROVAL DATE</strong></td>
<td>29/07/2013</td>
</tr>
</tbody>
</table>

This project has been allocated Ethics Approval Number H5246, with the following conditions:

1. All subsequent records and correspondence relating to this project must refer to this number.

2. That there is NO departure from the approved protocols unless prior approval has been sought from the Human Research Ethics Committee.

3. The Principal Investigator must advise the responsible Human Ethics Advisor:
   - periodically of the progress of the project,
   - when the project is completed, suspended or prematurely terminated for any reason,
   - within 48 hours of any adverse effects on participants,
   - of any unforeseen events that might affect continued ethical acceptability of the project.

4. In compliance with the National Health and Medical Research Council (NHMRC) “National Statement on Ethical Conduct in Human Research” (2007), it is MANDATORY that you provide an annual report on the progress and conduct of your project. This report must detail compliance with approvals granted and any unexpected events or serious adverse effects that may have occurred during the study.

**Human Ethics Advisor:** Woodward, Lynn  
**Email:** lynn.woodward@jcu.edu.au

This project was Approved by Executive on 29 Jul 2013

**Dr Anne Swinbourne**  
*Chair, Human Research Ethics Committee*
Appendix D

ACCLAiM Dissemination List

25th June 2013: Medical Deans of Australia and New Zealand (MDANZ) Assessment meeting: Presentation by Associate Professor Teague describing the project and inviting all medical schools to participate; panel discussion including leads from Australian Medical Assessment and the Australian Medical Schools Assessment collaborations, the head of the United Kingdom Medical Schools Council Assessment Alliance and Chair of the USA National Board of Medical Examiners.

2013 – 2014 Participation by Associate Professor Teague (representing ACCLAiM) in the Steering Committee of the HWA funded MDANZ Collaboration project.

6th December 2013: ACCLAiM Meeting in Sydney attended by representatives from James Cook University, University of Tasmania, University of Wollongong, Deakin University, Bond University, Notre Dame, University of Sydney, University of Western Sydney, and Newcastle/Joint Medical Program. Apologies received from the University of Queensland, Griffith University, University of Adelaide, University of Western Australia, Flinders University & Otago University (NZ). Analysed project data was presented. Debrief of 2013 OSCEs took place and 2014 OSCE blueprinting was discussed. All the new collaborators indicated their intention to continue participating in the project.

February 2014: Faculty Development Meeting for James Cook University clinical academic staff: presentation of ACCLAiM project and student performance data (JCU students identified). Dr Malau-Aduli, A/Prof Teague.

4th April 2014: ACCLAiM Collaborator meeting in Sydney attended by representatives from JCU, UTAS, Deakin University, Griffith University, University of Wollongong, Joint Medical Program (University of Newcastle/University of New England), University of Notre Dame Australia, University of Western Sydney, University of Sydney, University of Queensland, Otago University and Bond University. Apologies received from Flinders University and Australian National University. Guest attendee Professor Cees van den Vleuten, University of Maastricht. Analysed data was presented, 2014 QA process was reviewed, IMS was introduced and 2014 OSCEs were discussed and finalized. All existing collaborators indicated their intention to continue project participation. Otago University noted they would not be able to participate this year, but would discuss future options.

April 2014: Faculty Development Assessment meeting for University of Wollongong clinical academic staff: presentation of ACCLAiM project and student performance data (UOW student data identified) A/Prof Teague.

April 2014 Ottawa 2014 Medical Education Conference:

i. Oral paper (Dr Bunmi Malau –Aduli)

ii. Workshop (Professor Richard Turner, A/Prof Teague, Dr Bunmi Malau-aduli)
April 2014: National Medical Educational Forum University Western Sydney: Presentation on ACCLAiM project Associate Professor Teague.

July 2014: Australia and New Zealand Association for Health Professional Educators (ANZAPHE) annual meeting Gold Coast:
   iii. Oral presentation on ACCLAiM with focus on the External Examiner (QA) process: Dr Ian Kerr
   iv. Oral presentation on ACCLAiM presenting student performance data: A/Prof Teague

1st August 2014: ACCLAiM Collaborator meeting in Sydney (hosted by University of Sydney) attended by representatives from JCU, UTAS, Deakin University, Griffith university, University of Wollongong, Joint Medical Program (University of Newcastle/University of New England), University of Notre Dame Australia, University of Western Sydney, University of Sydney, University of Queensland, Bond University, University of Adelaide, Monash University and Australian National University. Guest: Professor Katharine Boursicott (Project Evaluation). Verbal update OSCEs so far and recommended changes. Generic QA report template found to be helpful. Presentation on the work-based assessment model in the intern year. Blueprinting for 2015 OSCEs.

September 2014: Association of Medical Educators of Europe (AMEE) annual conference Milan, Italy: oral presentation on benchmarking of clinical assessment A/Prof David Garne.

3 December 2014: Professor Richard Turner presented the ACCLAiM project to the Peer Review of Assessment Network (PRAN) at the University of Tasmania Newnham Campus. PRAN is an OLT-funded project encompassing various disciplines in the university sector.

20 March 2015: ACCLAiM Collaborator meeting in Brisbane (hosted by University of Queensland) attended by representatives from JCU, UTAS, Deakin University, Griffith university, University of Wollongong, University of Notre Dame Australia, University of Western Sydney, University of Sydney, University of Queensland, Bond University, University of Adelaide, University of Newcastle and Australian National University. Apologies received from Joint Medical Program (University of Newcastle/University of New England) and Monash University. Presentation of analysed 2014 data. Changes made regarding streamlining data collection and analysis process. Examiner training process discussed. 2015 OSCEs discussed and finalised, as well as 2015 exam dates and locations. All existing collaborators indicated their intention to continue project participation.


April 2015: Medical Deans of Australia and New Zealand (MDANZ) Assessment Forum, Sydney: oral presentation on ACCLAiM; participation in panel discussion. Associate Professor Peta-Ann Teague
Appendix E

ACCLAIM Early Clinical Exam 2014 Gastroenterology

Student ID: ___________________________ School Code: ___________________________

<table>
<thead>
<tr>
<th>A- COMPETENCIES</th>
<th>Fall</th>
<th>Borderline Pass</th>
<th>Borderline Fail</th>
<th>Clear Pass</th>
<th>Exceptional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication skills</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clear pass: Washes hands, introduces self and gives explanation of their role. Establishes good rapport and empathy with the patient. Explains differential diagnoses to examiner in appropriate terms.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exceptional: Demonstrates a confident and empathic approach. Offers patient an opportunity to ask questions Borderline: Some empathy and rapport absent. Explanation of examination fail: Does not wash hands and has little or no communication/ empathy with the patient.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Examination: General &amp; Signs of Hepatic Dysfunction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clear pass: Will lie the patient flat. Checks for palmar jaundice, bruising. Vitals, Peripheries for CRT, perfusion. Dilated veins, spider naevi, bruising. Caput medusa and other signs of liver dysfunction the student will mention several of these signs but not an exhaustive list. Exceptional: Considers resuscitation before examination rechecks vital/ postural drop. Borderline: Student considers vitals but there is little focus or organisation to general examination and there are deficiencies in technique. Fall: Examination is not focused or relevant.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abdominal Examination</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clear pass: Abdominal distension, tenderness, guarding or rebound Haphtomegaly, Splenomegaly BS Exceptional: PR for melena, Ascites Borderline: there is a focus on organisation to GI examination and there are some deficiencies in technique. Fall: Examination is not focused, marked deficiencies in technique.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Differential Diagnoses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clear pass: Bloods for FBC, U and E, LFT Coagulation studies, GP and Hold.[Would accept X matches 2 units.] Considers urgent endoscopy, considers surgery Gastroenterology referral. At expected student will consider most of these investigations. Exceptional: Considers another IV line, Erect CPR Borderline: Ilogial and unfocused approach to investigations. May over investigate with irrelevant investigations Fall: will request 1 or 2 tests only with little focus.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B. GLOBAL RATING

1  2  3  4  5  6  7

Very poor performance Well short of expected standard Short of expected standard Expected standard Better than expected Much better than expected Exceptional performance

COMMENTS (Please provide comments especially if you have rated the candidate as Unsatisfactory or Borderline.)

EXAMINER NAME: ___________________________ SIGNATURE: ___________________________
Appendix F

<table>
<thead>
<tr>
<th>Year Level:</th>
<th>Early years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of Station:</td>
<td>8 minutes</td>
</tr>
<tr>
<td>Discipline/Topic:</td>
<td>Respiratory Medicine</td>
</tr>
<tr>
<td>Overview of station:</td>
<td>Respiratory history and interpretation of spirometry in a patient presenting with shortness of breath.</td>
</tr>
<tr>
<td>Year objective/s being assessed:</td>
<td>History taking skills, communication skills, interpretation of investigation skills</td>
</tr>
</tbody>
</table>

**STUDENT INFORMATION**

**Scenario:**

You are an intern working under the supervision of a General Practitioner. You are reviewing patients in the local practice.

50 year old Jack Lang presents with progressive shortness of breath on exercise over the last five years.

You are required to take a history relevant to the patient’s respiratory symptoms.

At six minutes, you will be shown spirometry results, and asked to explain them to the patient.

**Tasks:**

1. Take a focused respiratory history
2. Explain the spirometry results to the patient.
ROLE PLAYER INFORMATION

Scenario

You are Jack a 50 year old presenting to the respiratory clinic with increased shortness of breath. This has gradually increased over the last 5 years but what really forced you to come in today was that you were told by your partner that you have to do something about your cough!

You cough a small amount of grey sputum each day, you thought this was related to a virus you had, but the cough has not gone away. You get several bouts of coughing each day, which annoys your partner, and makes them worry about your health.

You have:

- Daily wet cough, productive of a teaspoon of grey sputum several times per day
- Once or twice per year, usually in winter, worse symptoms with lots of yellow phlegm which have responded to antibiotics
- Shortness of breath when going up a hill, even when it is not too steep. Stairs also make you short of breath after a couple of short flights. You can still manage to make your way around the golf course on the flat, though you do puff noticeably after a longer walk.
- Been immunised to influenza and whooping cough

You DO NOT have:

- A runny nose, sore throat or headaches
- Wheeze or noisy breathing
- Any blood in your phlegm
- A fever, fatigue, malaise, night sweats
- Unexplained weight loss
- Chest pain, palpitations, nausea, sweatiness, or a personal or family history of cardiac problems.
- Any shortness of breath when lying flat and you do not wake at night short of breath.
- Swelling in your legs
• A history of asthma/eczema/hayfever/allergy in yourself or your family
• A history of travelling overseas or local contact with tuberculosis.

You have been smoking for a long time, you started at age 18, and you smoke 40 per day. You quit smoking six months ago, but this has not helped your shortness of breath or cough.

Other medical history:
• Right knee ACL repair when you were 32 after a soccer injury.
• No regular medications, no known allergies.
• No diseases run in your family, mother and father both alive, well in their late 70s.

Social History

You work as an accountant in a usual office environment. Your symptoms are not worse at work. You do not have any pets. You live with your partner, who is a lifelong non-smoker.
You drink a small bottle of mid strength beer several times a week after work.
You play golf about once a week as your major form of sport and recreation.
Your daughter has finished high school and is travelling overseas on a gap year.

Systems review:
You do not have any other symptoms in other bodily systems.

TASK:

At six minutes the examiner will stop the student.

Ask the student

“What do these results mean?”
**EXAMINER INFORMATION**

*At six minutes*, stop the student, give them the spirometry results and say “The patient has a question”

**Spirometry Results**

<table>
<thead>
<tr>
<th></th>
<th>Pre-bronchodilator (% predicted)</th>
<th>Post-bronchodilator (% predicted)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FEV1</strong></td>
<td>1.1 L (44%)</td>
<td>1.2 L (47%)</td>
</tr>
<tr>
<td><strong>FVC</strong></td>
<td>3.0 L (92%)</td>
<td>3.1 L (97%)</td>
</tr>
<tr>
<td><strong>FEV1/FVC</strong></td>
<td>0.33</td>
<td>0.39</td>
</tr>
</tbody>
</table>
Station Requirements

<table>
<thead>
<tr>
<th>Item</th>
<th>Quantity</th>
<th>Available from / provided by</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Patient Details

**Role Player:**

**Approximate Age:** 50

**Gender:**  M/F

**Patient name:**  Jack/Jean

**Presenting complaint:** Shortness of breath
Appendix G

43

To whom it may concern

On 4 April 2014 I have attended an ACCLAIM meeting in Sydney. One of my goals during my sabbatical is to become informed on the Australian situation on the issue of national exams and/or benchmarking initiatives. Many countries in the world are thinking about moving towards national systems of assessment. The debate is often carried out with many naïve arguments. Benchmarking is often used as one of the motivations. Unfortunately the state of affairs in assessment is such that easy benchmarking is not really possible. Single test administrations are not very informative for a long list of reasons. ¹ Unwanted side effects need to be carefully monitored.

Another motivation is the sharing of test resources across institutions. I am a great believer in that strategy. Developing high quality test material is resource intensive. Making a collective effort to produce high quality assessment and then share the benefit across schools is a definite way to go.

The impression I had from the ACCLAIM meeting was an extremely positive one. Benchmarking in itself was not the prime purpose, more or less a by-product that requires careful consideration in terms of interpretation. One of the ACCLAIM members said: “What we do is one big learning exercise, no more than that”. I could not agree more. Learning how to align test materials across schools, how to align standards, how to align assessors, how to analyse the information, how to interpret scores, how to use the resulting information meaningfully, how to learn from each other’s assessment practices, are all great questions and learning issues. I also understood ACCLAIM members visit each other’s assessment practices for quality assurance and mutual learning. Again, a very commendable strategy.

During the meeting I have presented the Dutch collaboration on progress testing. Of the 8 medical schools in The Netherlands, 5 collaborate in producing and administering concurrently 4 progress tests per year to all the medical students in the school. A progress test is a sample of items across all disciplines representing the end objectives of (undergraduate) medical training. It is not blueprinted against any curriculum, but to the generic end objectives. The progress test provides a formidable amount of (longitudinal) feedback, both to learners as well as to schools. Through an online system, learners, teachers and administrators have access to performance scores in any way they want, with the (near) national performance as the standard. For further explanation of this testing procedure see Wrigley et al., 2012. I am always prepared to give you more information when you are interested.

Overall, I am most impressed by the work of the ACCAIM network. There are more assessment alliances in the world, but in my view ACCAIM is an exceptional one. The project is an exemplary approach to a very mindful collaboration in assessment across schools.

Yours sincerely,

Cees van der Vleuten
Professor of Education
Director School of Health Professions Education
www.maastrichtuniversity.nl/she
Appendix H

Review of ACCLaIM (Australian Collaboration for Clinical Assessment in Medicine)
July 2014

Introduction
This review of the ACCLaIM project is from an external viewpoint and has been compiled by A/Professor Katharine Boursicot, Assistant Dean for Assessment and Medical Education Research at the Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore.

General Comments
This review is based on documentation provided and a meeting I attended in Sydney in August 2014.

Overall, the plans for collaborative assessment across Australia and New Zealand, in the area of testing clinical skills using OSCEs, is a very forward-looking and ambitious project: to attempt to bring equivalent benchmarking in clinical examinations across all medical schools in two countries. The issue of equivalence of graduation assessments is one which challenges medical schools across the world, especially in those countries without national licensing examinations. It is especially difficult to find agreement on clinical testing, as compared to knowledge assessments, where several item banks are in existence.

For the ACCLaIM project, there was engagement with most of the medical schools in Australia. Some schools appeared to have taken on the project with enthusiasm, while challenging issues were present for some of the other schools, where it appeared that more senior support for the project was lacking. It was unclear to me if exit examinations across all the Australian medical schools could currently be considered to be equivalent.

The cross-site visits undertaken as part of the ACCLaIM activities appeared to be very constructive in terms of QA and OSCE-literacy enhancement. Both the visits and reports seemed to be very useful in advancing standards as well cross-collaboration and shared understanding. It is commendable that so many people took the time to visit other sites and conscientiously attend OSCEs, using a structured template for reports (and differentiating between ACCLaIM and non project stations).
OSCE-specific comments

There appears to be considerable variation in course structure and assessment points across the different medical schools. In some schools, integrated OSCEs were conducted, while in others, discipline-embedded OSCEs were prevalent. While it is difficult to achieve consistency across the country within each course and across different medical school curriculums, it might be that a graduation-level OSCE would present the best opportunity for the use of common stations and scoring, with the test mapped to ‘Intern Readiness Skills’.

- **Level of challenge of tasks**
  The differences in course structure across the medical schools makes it challenging for devising appropriate level of station design tasks which could be used across all medical schools at the same levels. The most obvious common level would be graduation level assessments, and indeed many of the stations being designed are aimed at graduation level skills.

- **Design of OSCE circuits/timings**
  There is variability in the design of OSCEs in different medical schools, in terms of numbers and length of stations: numbers of stations varied between 4 to 16 stations while length of stations vary between 5 and 12 minutes. While there is no ‘perfect’ OSCE design, it would be useful to agree on a common length of station timing as well as numbers of stations used in any particular equivalent level examination (especially at graduation).
  Generally, the total testing time is suggested to be between 2 and 3 hours, while station length should be appropriately authentic to the task, with feasibility taken into consideration. In practice, stations up to 15 minutes duration can be used effectively, while 5 minute stations can only test the most basic of practical skills.

- **Blueprinting of skills tested** could be more consistent, but as course structures are variable, each school’s blueprint will necessarily be differently mapped to the learning objectives. However, it would be useful to agree the domains of skills which should be tested i.e.
  - History taking
  - Explanation/Information giving
  - Clinical examination
  - Practical skills
  Blueprinting to ‘Intern Readiness Skills’ would be good way forward, especially if each school’s final year learning outcomes were aligned with these.
• Scoring
  It is worth considering different scoring schemes i.e. checklist versus rating scales, as it is becoming increasingly accepted (and supported by the literature, e.g. Hodges 1999) that rating scales are more appropriate for increasing levels of clinical expertise. In some medical schools, checklists are used only in the early years of the course, while rating scales (with BARS - Behaviourally Anchored Rating Scales) are used for the later clinical years. Most clinicians are comfortable with the use of rating scales, as their expertise is engaged in the decision making processes around competence levels.

• Standard Setting
  I was unclear about the standard setting methods used in the different schools. The current ‘gold standard’ for standard setting for OSCEs is the Borderline Regression Method, but I was unsure if this was being used in any of the medical schools involved in the collaboration. This would be an important aspect of the validity of any OSCE.

• Reliability and Validity
  While reliability of any test is important, the pursuit of high reliability must be balanced against the authenticity of the tasks being tested. Overall, a reliability index of 0.65 is acceptable for an OSCE (as compared with at least 0.80 for an MCQ paper); this can be achieved not just by increasing the number of stations, but also by increasing the length of each station. Of course, the use of Standardised Patients (as in the USMLE) increases reliability but the use of real patients with signs is much more authentic.
  More modern thinking about validity (Kane 1994, 2013) suggests that reliability is only part of the overall validity of any tests, and that the use and interpretation of the results are more important considerations.
  Thus, it is not that any particular station is ‘valid’ but how that station is used in a particular context, and how the results are interpreted and administered (e.g. progression or not), which are the contributors to the ‘validity’ of the test.
  Sharing and use of the same stations across different medical schools, while commendable and a great step forwards in equivalence, does not guarantee equivalence of graduate skills.

• Test security issues
  There were some concerns about test security in relation to examinees sharing test/task challenge content, especially if schools hold their examinations at different times. There is no clear evidence that knowing the upcoming tasks in an OSCE leads to better examinee performance. This is understandable if the OSCEs are testing skills, which cannot be learned in a few hours or days: clinical expertise requires years of ‘deliberate practice’ (Dreyfus 2005).
Another solution is the use of rating scales with BARS, which may be disclosed to the examinees, so they are all aware of what is required in terms of a sufficiently good or excellent performance, rather than just performing to a checklist of ‘done’ or ‘not done’.

- Examiner training
  The use of online videos for examiner training is highly commendable, as many of the examiners are located in remote areas. Expansion and further standardising of examiner scoring is to be encouraged, especially if BARS are introduced.

Summary

The ACCLAIM project is an impressive plan to bring equivalence in clinical testing by the use of shared OSCEs. Encouragement for all schools to sign up to the ACCLAIM project would be a major step forward in ensuring the equivalence of medical school graduate skills. Unfortunately, in some institutions, the senior management may not be up-to-date with modern assessment literacy and do not support the use of OSCEs, maintaining the traditional long case and short case scenarios.

So far, the ACCLAIM project has gone further than any other endeavour to

I would suggest further work for the collaboration should include

- Review of scoring systems: checklist versus BARS
- Blueprinting: mapping to ‘Intern Readiness Skills’
- Implementation of formal standard setting procedures
- Using the Kane model of validity to evaluate the OSCEs conducted in different medical schools
- Enhancing assessment literacy
- Standardised examiner training
References:

Dreyfus, H. L. and S. E. Dreyfus 2005
Peripheral vision: Expertise in real world contexts
Organization Studies 26(5): 779-792

Hodges, B., et al. 1999
OSCE checklists do not capture increasing levels of expertise
Academic Medicine 74(10) pp 1129-1134

Kane, MT, 1994
Validating interpretive arguments for licensure and certification examinations
Evaluation and the Health Professions, 17(2) pp 133-159

Kane, MT, 2013
Validating the interpretations and uses of test scores
Journal of Educational Measurement, Spring 2013, 50,(1) pp 1-73
Appendix I

Guidelines for ACCLAiM Quality Assurance (QA) Examiners 2013

Introduction

The Australian Collaboration for Clinical Assessment in Medicine (ACCLAiM) was formed in 2009. An absence of accessible national data regarding student performance in clinical exams led to the recognition that there was a need to benchmark students’ attainment of core clinical competencies in a way that usefully informed medical schools about the efficacy of their teaching and assessment.

ACCLAiM examines the performance of Australian medical students in summative Objective Structured Clinical Examinations (OSCEs) and provides feedback on this to participating medical schools.

ACCLAiM members collaboratively develop shared OSCE stations which are then embedded within individual medical school summative OSCEs at either the early clinical or exit exam stage. There is sufficient flexibility within the process to allow modifications of the shared OSCE stations to suit participating medical schools’ particular curricula and assessment protocols.

De-identified data about student performance in the shared stations is subsequently made available to the participating schools.

Each school is also provided with a Quality Assurance (QA) report.

Overview of the QA process

Each medical school in ACCLAiM participates in providing QA feedback pertaining to the summative OSCEs. The feedback covers not only the quality of assessment items, but also the organization of the OSCE, the performance of the assessors and the performance of the simulated or role player patients. The student experience is also reviewed. The QA process recognises the importance not only of ensuring high quality assessment items but also the process of the OSCE itself including consistency of assessors and simulated patients.

The visiting examiner’s institution meets the costs of travel and accommodation for this activity.

The collaboration aims to provide a minimum of two QA examiners per OSCE, and will attempt to meet the need for further examiners if medical schools request this (e.g. for schools that hold concurrent OSCEs at a number of geographically distant sites).

Guidelines for QA Examiners

Visiting examiners are requested to allow sufficient time to cover all the QA tasks. This may require a ‘two night’ rather than a ‘one night’ stay.
The QA examiner double marks one or two ACCLAiM stations. The QA examiner then spends time in as many of the other (‘non-ACCLAiM’ stations) as possible.

**Duties prior to the exam**

1. Liaise with the school being visited to get advice on when the examiner meeting/training will be, and to plan to attend that wherever possible (can be done remotely if feasible e.g. the Deakin model).
2. Book flights and accommodation – the local ACCLAiM academic can be helpful here in advising on the best accommodation options. Ensure that the school knows that you are coming (and when).
3. Where applicable, liaise with the other QA examiner/s to divide up the ACCLAiM stations, as well other duties e.g. discussion with examiners, discussion with students. The more this can be shared the more likely it is that all the OSCE stations will be reviewed.
4. Do the online examiner training for the applicable ACCLAiM stations; it is advisable to do this at least 1/52 before the actual OSCE so that you can address any queries you might have about the station.

**Duties on the day**

1. Attend the pre exam examiner meeting, if applicable.
2. Ensure that students and assessors know who you are and what your role is; usually the academic coordinating the day does this introduction.
3. Double mark the ACCLAiM stations using the common marking sheet as well as the medical school’s marking sheet (nifty marking skills required here!). Aim to double mark at least 6 iterations of each ACCLAiM station. Consider the following:
   a. The level of difficulty of the station
   b. The competencies being tested
   c. Student performance at the station
   d. The simulated patient performance (believable/consistent etc)
   e. The assessor performance (consistency/how well prepared etc)
   f. Overall evaluation of the station + any other relevant comments
4. Observe (and double mark if feasible) as many of the non-ACCLAiM stations as time allows, following the domains in point 2 above. Sometimes observing just one or two iterations of each station will be sufficient; in other instances (e.g. concern re an examiner not following guidelines) it may be useful to observe more iterations. If the same station is being run on more than one floor, try to observe all the assessors examining at the same station (this can involve nifty foot work to run up and down stairs between bells).
5. Be careful not to “stalk” any one student; when moving to a new station, best to go in the opposite direction to the students.
6. Talk to the relevant administrative and assessment staff about the organization of the exam, and observe how this is implemented and how smoothly it runs. Consider the impact of the organization of the OSCE on the student experience e.g. amount of reading time, ease of access through the stations etc.
7. Chat informally to as many assessors as possible. Ascertain the level of “buy in”, knowledge of the course and preparation for the OSCE.
8. Wherever feasible, talk to the students as a group after the OSCE. Ask for their feedback about the stations and the exam overall.
Report Writing

Where applicable, liaise with your fellow QA examiners and submit one report. This works if the OSCE is at one site; for those OSCEs held at multiple sites it is more useful for the medical school to have a QA report for each site.

Use the pro forma attached to this guide. Aim to make the report as helpful as possible, highlighting strengths and high quality stations as well as those aspects of the day that could be improved. The pro forma is in MS Word so you can use as much or as little space depending on your comments.

Submit the report to the appropriate ACCLAIIM academic within 1/52 of the OSCE.
Appendix J

Template Visiting QA Examiner Report (OSCE)

Medical School + Site of exam: 

Year Level: 

Date: 

Organisation of OSCE

Student Experience

Examiners

Role Players/Simulated patients

Station X (?ACCLAiM or not)

a. The level of difficulty of the station

b. The competencies being tested

c. Student performance at the station

d. The simulated patient performance (believable/consistent etc)

e. The assessor performance (consistency/how well prepared etc)

f. Overall evaluation of the station + any other relevant comments
Appendix K
with benefits such as improved cost-effectiveness, transparency, accountability, and standardisation of process and relevant content among participating medical schools (Mujiqje et al. 2008; Wilkinson 2010; Wilkinson et al. 2014). 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 (Barsky et al. 1992; Whelan 1999; Medical Council of Canada 2002; Turner & Dankski 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 been done for written assessment (Wilkinson et al. 2014). Given this context, there is a 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 evaluative 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 four different states in Australia. Two of the schools run four-year undergraduate-entry medical 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 educational 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) (2011) medical competencies project. 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 11 stations used across the three distinct OSCE cycles, their descriptors and the competencies that each of them assessed. The final 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 EC exam, as well as three new stations used in the exit level exams (EL). 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 blueprints, 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 (Melusi-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 visiting 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 partitioning the variance in raw scores into variance due to item difficulty and student ability (Mannis & Andrich 2008). The Rasch model serves as a QA 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 (Gustafsson 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 (PRD) 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 (PSD) is calculated to indicate the efficiency of the items in separating the persons measured (Bond & Fox 2012). The item measure reliability (IRD) indicates how well the scale can distinguish between items, on the basis of their difficulty (Bond & Fox 2012), and the item separation index (ISO) 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 <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 logits, thereby converting ordinal data to equal-interval data, allowing 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 (Haggard & Andrich 2004). DIF was used in this study to examine whether the OSCE stations functioning differently by entry program (graduate vs undergraduate entry), gender (males & females) and origin (domestic & international students). A value of $<0.43$ is not significant, $\geq0.43$ indicates slight to moderate difference and $\geq0.64$ 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 QA reports provided to each school after the examinations, the assessors deliberated on and documented the benefits and impact of the benchmarking exercise on teaching and learning in their respective schools. Over the two years study period, 15 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 outfit values for person and item ranging from 0.88 to 0.90 and 0.95 to 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 $-5$ 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 (Figure 2a).

Table 1. Reliability measures.

<table>
<thead>
<tr>
<th>Measurement</th>
<th>2011 EC</th>
<th>2012 EC</th>
<th>2012 EE</th>
</tr>
</thead>
<tbody>
<tr>
<td>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</td>
<td>0.61</td>
<td>0.73</td>
<td>0.70</td>
</tr>
<tr>
<td>which measure the same construct.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person separation index (PSI) indicates the efficiency of the items in separating the persons measured.</td>
<td>1.25</td>
<td>1.63</td>
<td>1.52</td>
</tr>
<tr>
<td>Item reliability index (RI) is a measure of the consistency of inferences made on item difficulty.</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>Item separation index (SI) indicates the efficiency of the sample of persons in separating the items used.</td>
<td>4.31</td>
<td>4.58</td>
<td>4.30</td>
</tr>
</tbody>
</table>
Figure 2. Item-person maps.

Figure 3. Differential item functioning by entry program and origin.
station most difficult (management plan and communication skills). There were no differences in performance patterns for all 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 (e.g. 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 (e.g. 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 student was said to have found 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 QA 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 (Pettus et al. 1991; Mulijens 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 OSCE stations, 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 (Mulijens 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. These 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 & 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 QA 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 QA 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.

Ethical approval

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

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.


Notes on contributors

RUNUH SHERIFF MALAU-ADULL, BSc, MSc, PhD, is a Senior Lecturer in Medical Education at the College of Medicine & Dentistry, James Cook University.

PETE-ANN TEAGUE, MChBr, DRCOG, FRACGP, 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, DMedSci, FRCS, PhD, is a Professor of Surgery and the Associate Head of the Rural Clinical School at the School of Medicine, University of Tasmania.
Appendix L

Attributes Spectrum

Appendix E

(1) Scientific method relevant to biological, behavioral and social sciences at a level adequate to provide a rational basis for present medical practice, and to acquire and incorporate the advances in knowledge that will occur over their working life.

(2) The normal structure, function and development of the human body and mind at all stages of life, the factors that may disturb these, and the interactions between body and mind.

(27) Recognition that the doctor should have the necessary professional support, including a primary care physician, to ensure his or her own well-being.

(30) An appreciation of the systems approach to health care safety, and the need to adopt and practise health care that maximizes patient safety including cultural safety.

(32) A commitment to ease pain and suffering.

(7) The principles of health education, disease prevention and screening.

(38) A desire to achieve the optimal patient care for the least cost, with an awareness of the need for cost-effectiveness to allow maximum benefit from the available resources.

(31) A realisation that it is not always in the interests of patients or their families to do everything that is technically possible to make a precise diagnosis or to attempt to modify the course of an illness.

(10) Systems of provision of health care in a culturally diverse society including their advantages and limitations, the principles of efficient and equitable allocation and use of finite resources, and recognition of local and national needs in health care and service delivery.

(9) The sociology, pathology, symptoms and signs, natural history, and prognosis of common mental and physical ailments in children, adolescents, adults and the aged.

(35) An appreciation of the responsibility to contribute towards the generation of knowledge and the professional education of junior colleagues.

(34) An appreciation of the responsibility to maintain standards of medical practice at the highest possible level throughout a professional career.

(24) The ability to interpret medical evidence in a critical and scientific manner and an understanding of the epidemiology of disease in differing populations and geographic locations.

(29) Respect for community values, including an appreciation of the diversity of human background and cultural values.

(26) Recognition that the doctor's primary professional responsibilities are the health interests of the patient and the community.

(33) A realisation that doctors encounter clinical problems that exceed their knowledge and skills, and that, in these situations, they need to consult and/or refer the patient for help, in clinical, cultural, social and language related matters as appropriate.

(12) The principles of ethics related to health care and the legal responsibilities of the medical profession.

(29) Respect for every human being, including respect of sexual boundaries.

(25) The ability to use information technology appropriately as an essential resource for modern medical practice.

(9) Factors affecting human relationships, the psychological, cultural and spiritual well-being of patients and their families, and the interactions between humans and their social and physical environment.

(40) A realisation that one's personal, spiritual, cultural or religious beliefs should not prevent the provision of adequate and appropriate information to the patient and/or the patient's family, or the provision of palliative care including referral to another practitioner.

(8) The principles of prevention of suffering and disability, rehabilitation and the care of the dying.

(6) Normal pregnancy and childbirth, the more common obstetric emergencies, the principles of antenatal and postnatal care, and medical aspects of family planning.

(20) Communication skills, including being able to listen and respond, as well as being able to convey information clearly, considerately and sensitively to patients and their families, doctors, nurses and other health professionals and the general public.

(21) The skills needed to work safely as an intern, as outlined in the National Patient Safety Education Framework developed by the Australian Council for Quality and Safety in Health Care.

(22) The ability to counsel patients sensitively and effectively and to provide information in a manner that ensures patients and families are fully informed when consulting to any procedure.

(37) A commitment to communicating with patients and their families, and to involving them fully in planning management.

(39) A preparedness to work effectively in a team with other health care professionals.

(11) Indigenous health, including the history, cultural development and health of the Indigenous peoples of Australia or New Zealand.

(13) The ability to construct, in consultation with a patient, an accurate, organised and problem-focused medical history.

(14) The ability to perform an accurate physical and mental state examination.

(32) An appreciation of the complexity of ethical issues related to human life and death, including the allocation of scarce resources.

(23) The ability to recognize serious illness and to perform common emergency and life-saving procedures, including care for the unconscious patient and cardiopulmonary resuscitation.

(15) The ability to choose, from the repertoire of clinical skills, those that are appropriate and practical to apply in a given situation.

(16) The ability to interpret and integrate the history and physical examination findings to arrive at an appropriate diagnosis or differential diagnosis.

(4) Common diagnostic procedures, their uses and limitations.

(17) The ability to select the most appropriate and cost effective diagnostic procedures.

(16) The ability to interpret common diagnostic procedures.

(5) Management of common conditions including pharmacological, physical, nutritional and psychological therapies. A more detailed knowledge of management is required for those conditions that require urgent assessment and treatment.

(19) The ability to formulate a management plan, and to plan management in concert with the patient.

For Attributes 1 - 12 “Graduates completing basic medical education should have knowledge and understanding of”

For Attributes 13 - 25 “Graduates completing basic medical education should have developed the following skills and abilities”

For Attributes 26 - 40 “At the end of basic medical education students should demonstrate the following professional attitudes that are fundamental to medical practice”

© Medical Deans Australia and New Zealand, Competencies Project
Appendix M

Abstract, Ottowa 2014

UNIFORM DIVERSITY: HOW TO ACHIEVE STANDARDIZATION IN LIVE CLINICAL ASSESSMENTS, WHILE STILL RETAINING LOCAL AUTONOMY

Background

Objective Structured Clinical Examinations (OSCEs) are integral to assessment in medical education. Creating and delivering quality OSCE items can be time-consuming and intellectually demanding. Sharing items among like-minded institutions alleviates some of this burden, while also addressing the desire for standardization of assessment at a national level. This often must be harmonized with certain content and procedural specifications imposed at a local level.

The ACCLAIM project was founded by a number of Australian Medical Schools with these goals in mind. Those who have been involved in the project are now in a position to share their acquired expertise.

Intended outcomes

Participants will emerge with a framework for developing, implementing and evaluating collaborative OSCE items. They will also acquire experience in skills required by examiners and item creators.

Structure

Presenters will facilitate discussion around various topics relating to inter-institutional sharing of live assessment items. They will also provide illustration from their own experiences as part of a collaborative demonstration project and evidence gleaned from the literature. Topics covered will include inter-examiner calibration, objective scoring, standard setting, curriculum alignment, and iterative item improvement.

The session will be interactive, with practical exercises in inter-examiner calibration and consensus-driven item improvement.

Intended audience

This workshop is intended for all medical educators who engage in live clinical assessments, and wish to share assessment items in order to maximise intellectual capital and continuously improve item performance.

Level of workshop (introductory, intermediate or advanced)

Intermediate: Familiarity with the OSCE paradigm is a pre-requisite for this workshop.
Appendix N

Ottawa 2014 workshop report

A 90-minute workshop was conducted with the title *Uniform diversity: How to achieve standardization in live clinical assessments, while still retaining local autonomy*. It was facilitated by Richard Turner, Peta-Ann Teague, Bunmi Malau-Aduli and Karen d’Souza.

There were approximately 12 participants, from countries including the UK, Poland, Egypt and Australia. Experience with OSCE assessment varied from none to considerable.

The principles of OSCE station development were outlined, particularly with respect to what needs to be considered when writing and implementing across medical schools with varying curriculum specifications and cultural contexts. There was also a brief presentation of the comparative analysis of ACCLAiM stations thus far.

This was followed by a small-group activity that involved developing a station in the broad competency domain of Ethics and Law. The lead-in stem was “Please don’t tell my partner that (s)he has cancer”. Participants were asked to consider all aspects of station development, but especially what checklist fields they felt were important and what anchor criteria should be allocated to these.

From the ensuing discussion, it was evident that prevailing cultural mores may define different desirable outcomes from such an encounter. It was nevertheless demonstrated that consensus could be achieved on the principles of an ethical patient-centred discussion.

Finally, a brainstorming session was convened to identify issues that need to be addressed for successful collaboration in OSCE station development. These included:

- Clinically grounded, authentic topics
- Good communication
- Central administration or project support
- Respect and trust among institutions
- Prospective blueprinting in accordance with a universally accepted framework
- Acknowledgment of cultural nuances
- Comprehensive training of examiners and simulated patients

By the end of the session, all participants indicated that they found the experience very useful, and also had the first draft of an OSCE station that they could apply within their own curricula. Furthermore, the facilitators benefited greatly from the interchange of ideas and sharing of experience.