Implementing Common Assessment: Lessons and Models from AMAC

Developed by the Australian Medical Assessment Collaboration

A project funded by the Office for Learning and Teaching, Australian Government Department of Education

The Australian Medical Assessment Collaboration (AMAC) involves sixteen medical schools in Australia and New Zealand and the Australian Council for Educational Research. This collaboration has been funded by the Office for Learning and Teaching, part of the Australian Government's Department of Education.

The writing and compilation of this document was led by Daniel Edwards (ACER), with significant input from Lambert Schuwirth (Flinders University), and David Kramer (The Australian National University). Contributions to the document were made by members of AMAC across the medical schools involved. Further detail about AMAC can be found at www.acer.edu.au/amac.

AMAC partners are:

Macquarie University

Monash University

Australian Council for Educational Research

Flinders University

The University of Queensland

The University of Notre Dame Australia, Sydney

The University of Notre Dame Australia, Fremantle

The University of Wollongong

The University of New England/
University of Newcastle (Joint Medical Program)

The University of New South Wales

Griffith University

Deakin University

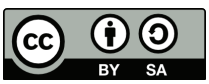
The Australian National University

Bond University

The University of Sydney

The University of Adelaide

The University of Otago



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.

ISBN: 978142862552

CONTENTS

1 INTRODUCTION

This document forms one of the core outputs from the Australian Medical Assessment Collaboration (AMAC), a collaboration funded by the Office for Learning and Teaching (OLT) between 2011 and 2014 (AMAC, 2012; Edwards, Wilkinson, Canny, Pearce, & Coates, 2014; Wilkinson, Canny, Pearce, Coates, & Edwards, 2013). Further details about AMAC can be found at <www.acer.edu.au/amac>.

The aim of this document is to provide insight into the implementation of common assessments in higher education in order to assist in future work on conducting these kinds of projects. The discussion here draws heavily on the AMAC experience, attempting to broaden the learning from this project for use in future collaborations. The focus of this project has been on medical education, and as such, much of the detail is related to this field. However, it is hoped that the general ideas discussed here can be seen as informative for other fields and disciplines in higher education and at least provide some guidance for those considering undertaking collaborations involving common assessment.

This document is focused on the practical implementation of such an initiative and intended as a reference for developing steps in the process of undertaking and administering common assessments across higher education institutions. Two other documents have been produced as part of the AMAC project to complement the detail provided here. One provides further insight into the issue of developing items for common assessment (*Determining the Quality of Assessment Items in Collaborations: Aspects to Discuss to Reach Agreement).* The other discusses issues of governance and dissemination of such projects (*Governance Models for Collaborations Involving Assessment).* One of the key drivers for the development of these three documents is the recognition that although such collaborations start in good faith by colleagues who know and respect each other personally, there is an imperative to formalise and professionalise the collaboration with a clear understanding and written agreements on quality, implementation and governance. This insight has been instrumental in deciding on the work packages for this AMAC project.

In preparing this document, the AMAC members involved reflected on the experiences of running the AMAC assessment in 2011, 2012 and 2013 in medical schools across Australia. In total, 11 medical schools have been involved in AMAC assessment implementation, with 20 different student cohorts tested and more than 1400 students taking part. AMAC developed an Assessment Framework and 120 items spanning a range of medical disciplines which were used in the testing implementations. Further AMAC items have also been developed, but not yet tested. Details of this are contained in other AMAC reports available on the AMAC website.

It is important to remember that in such projects, there is often a large difference between what is originally desired and what is practically feasible. Having comparability of outcomes of schools for benchmarking purposes was one of the key drivers for the establishment of AMAC, and while the nature of the project evolved along the way to be as much about collaborative development of items, sharing of expertise and nurturing of capabilities in assessment development, for the practical implementation of this project, comparability remains an important facet. Therefore, it is important to understand the implementation options and provide flexibility in these choices, while also maintaining some perspective on how these decisions might influence the comparability of outcomes. This document explores a number of these issues and aims to provide information on how AMAC dealt with these issues while still maintaining the aim of developing useful and useable data outputs for schools and students. While the main focus here is on the lessons learnt though AMAC, there is also some reflection on the way in which other collaborations have implemented common assessment projects.

This document examines four important facets of implementation of common assessments in higher education: **models of implementation, population identification, timing of assessment** and **reporting.** It concludes with some synthesis of these four areas. The focus of this document is the experience of developing and administering AMAC and thus the detail relates to the activities of AMAC implementation in medical schools in Australia and New Zealand.

2 MODELS

This section explores two approaches taken by AMAC in administration and implementation of the assessment items developed through the collaboration. In general, schools chose a model of implementation that best suited their needs and the timing of when the assessment items could feasibly be tested. Overall, there were two basic approaches used:

- online administration of a 'full' AMAC assessment
- embedding AMAC items within existing assessments.

These two approaches are discussed in this section. For each of these basic approaches, there were many nuanced differences in implementation across schools; these are detailed and further discussion on the benefits and possible pitfalls of each approach is presented.

A core issue highlighted throughout this document is the fact that in these kinds of projects, a 'one-size-fits-all' approach is not advisable. In undertaking the AMAC collaboration, the importance of recognising diversity across medical schools was paramount. This means that compromises are made in terms of collecting data that are perfectly comparable across schools. It also means ensuring that schools interpret results meaningfully within their own context. However, this does not necessarily mean that the output or the usefulness of the undertaking is reduced – in fact by acknowledging the different contexts within a system a more nuanced understanding of the outcomes should be gained.

A message articulated through the AMAC experience is that as long as some of the key elements discussed here are recognised and dealt with early on in the development, the output and benefits of involvement in the process will provide insight and will have a high likelihood of leading to necessary educational and organisational improvements.

This section explores the two approaches specifically trialled through the AMAC project. For each approach the discussion begins with a basic description of what processes AMAC followed. The next part then reflects on the benefits of this approach in the context of developing common assessments. The final discussion aims to offer some key considerations that future collaborative projects should be aware of during their development.

2.1 Online, formative implementation

2.1.1 AMAC approach

The AMAC project began with the original aim of developing items to be implemented in an online assessment for medical students. As such, initial development in 2011 and 2012 had the online administration as the sole focus for implementation. All schools who participated in 2011 and 2012 used the online version of the AMAC instrument with their students. In 2013, AMAC expanded to other uses with some institutions choosing to run the assessment online and others to embed items in existing exams generally undertaken with pen and paper.

The online administration of AMAC items involved participating students sitting a 100-item test, based on items mapped to the AMAC Framework (AMAC, 2012). Six different online AMAC tests were developed that rotated through the 120 items developed in the project. Following completion of the assessment items, students answered a basic survey, designed to gain insight about the extent to which the items reflected the students' coursework and anticipated future workplace practice.

An online testing platform developed and housed by the Australian Council for Educational Research (ACER) was used for the testing. This platform was accessed through a standard web browser via passwords (a unique password was provided to each participant). Students navigated their way through the test, answering multiple-choice questions. Response data were collected by the ACER servers and stored securely.

Participating institutions were provided with a manual that detailed the specifications required and support

was provided in ensuring computers were enabled for the testing. By using a platform that ran through any standard web browser, issues with installation of software and compatibility were avoided.

All institutions involved in the online implementation of AMAC undertook the test as a formative assessment. Sessions were invigilated and students were told that the test was to be completed under 'exam conditions' – i.e., individually – and without the use of supporting books or other information. In most institutions, participation in the AMAC online assessment was voluntary, with students invited by their school to take part. In general the AMAC assessment was promoted as an opportunity to test oneself on a range of areas of medical education, in a low-stakes environment, and in many cases universities timed the testing to coincide with periods where study for summative exams was occurring.

Detailed population data was collected from institutions in 2013 to enable tracking of student response numbers and the ability to inform schools on the extent to which participation was representative of the target student population. Of the four schools participating in 2013, three had participation rates at 85 per cent or above.

Some universities held the AMAC assessment sessions in the faculty computer laboratories (often in multiple sessions due to the size of the labs). Other institutions undertook the testing in lecture rooms or exam halls and allowed students to use their own laptops to access and complete the test. Invigilators were present in the sittings of the test to ensure that the examination protocols were adhered to. Institutions reported back to AMAC project managers on the running of the sessions, completing a basic online form that described the implementation process and allowed for noting any adverse incidents.

Participating students were provided with detailed individual reports on the outcome of the assessment. These reports were sent to students directly by the ACER members of AMAC and were not provided to schools. Students were sent the report via email addresses they supplied at the beginning of the test process (supplying an email address was not mandatory, but was necessary in order to receive an individual report). In 2013, these reports were generally sent to students within two weeks of completing the test.

2.1.2 Benefits of an online, formative approach

Online implementation has a number of advantages for institutions and in terms of the practicalities of administering a common assessment across multiple schools. The discussion here relates directly to the AMAC experience, but the overall application of these ideas is much more widely relevant.

Overall the online assessment approach in AMAC was useful because it ensured that participating students were all taking a set number of items that spanned the full spectrum of the areas in which items had been developed. This meant that uniform approaches to reporting could be undertaken and that comparative data could be consistent, because everyone had undertaken a rotation of the same items. As detailed in a later section of this document, the reporting possibilities from the online, standard implementation of AMAC were much more varied (and useful) than through the embedded approach undertaken. This was the case particularly for student-level reporting, but also for institution reporting.

From the point of view of managing the project, the online assessment option was more easily administered than other approaches. Once the assessment items were set in the online platform, the process of providing logins and technical information to universities was relatively straightforward. In addition, data collection and scoring was automatic and did not rely on schools to send (or even collect) test-related data. The database created through the online administration was simple for psychometricians to analyse and prepare item-, student- and institution-level data for reporting.

For individual schools, the online application of AMAC enabled an additional avenue for providing students with a formative assessment, in full and relatively easy to access online. There was some administrative burden for the institution in terms of identifying time and space for the assessment, ensuring the adequate information technology infrastructure existed and recruiting students.

Most schools organised for the AMAC assessment to be undertaken during study periods, where students would be most likely to benefit from 'revision' in test conditions. Institutions also benefited by having detailed reporting of aggregated student outcomes from the assessment (discussed in a later section). The online administration ensured that sufficient data across a range of items was collected to enable meaningful comparisons of student outcomes by student groups and across benchmarked institutions.

For students, the online approach enabled access to a range of items that were developed by medical experts from across the country. The reporting (outlined later) produced through the online implementation was detailed enough to provide feedback on a range of areas of the assessment with a short turnaround time. This gave students a tool to help identify areas of weakness to concentrate on for further revision.

For institutions, the use of a ready-made formative assessment instrument provided a relatively inexpensive and validated approach to providing feedback on student learning in the weeks preceding the summative assessments.

Other collaborative assessment projects have taken the feedback possibilities for formative assessment further. The inter-university progress test collaboration involves a group of five medical schools in the Netherlands who jointly produce and administer four summative tests per year, which are sat by all students of all year classes simultaneously. Each of the four tests is different but all use the same blueprint sampling functional medical knowledge and knowledge application attuned to graduate level. That way each student's individual progress towards this end point is measured. There is a massive amount of data (roughly 10 000 student results for each test, so 40 000 per year), and this enables valuable feedback, not only at the level of total scores or blueprint categories, but also momentary results and longitudinal results. Students are able to compare their results to the results of the students of their own university and the combined results of all participating students. To obtain this, students can log on to the database and compile feedback queries (for example: how is my progress on anatomy compared to the progress of all students of all five universities in my year group? Or how is the development of my correct to incorrect ratio on items concerning the respiratory system?). Thus, the feedback is completely student-centred.

2.1.3 Important considerations for an online approach

Through the development of AMAC over the past three years, a number of important issues have arisen in relation to undertaking this kind of exercise online. Some of these issues are raised here in the context of attempting to provide advice or guidance for future collaborations considering online test administration.

Test platforms for online assessment need to be robust, reliable and flexible. Some of the important facets for consideration include:

- technical simplicity
  - avoiding the need for installation of specific software (i.e., a web-based approach that utilises browser)
  - ensuring the flexibility of the platform to adapt to different screen types, resolutions and so on
  - ensuring the platform has the ability to function in a range of institutions (i.e., not limited to single institution/intranet)
- data storage capabilities
  - ensuring data are secure and ideally independent of the institutions involved in the testing
- item flexibility
  - the ability to house different formats for items
  - the ability to support a range of multimedia elements (images, diagrams, videos)
- a 'tried and true' platform
  - reducing technical problems associated with developing from scratch
  - reducing the costs of development.
- security – while it is recognised that the web-based approach to testing can be difficult to guarantee to be free from security issues, there are some steps that can be taken to reduce concern and increase reliability, such as
  - ensuring unique login/password based access to tests
  - use of test invigilators to supervise online testing sessions

- enabling tests to be 'switched' on and off so as not to be accessed outside of testing times
- where possible, the ability to 'lock-down' computers during testing
- initiating measures to appropriately deal with technical failures; for example, ensuring a backup server is available in the event of server failure
- ensuring data collected are stored securely with limited accessibility
- developed protocols to deal with breaches in security – identification of issues, alerting of breaches and processes for resolving such issues
- ensuring invigilation that will verify that the proposed candidate is really who they claim they are.

Practical considerations relate to the kinds of information that are useful for collecting from institutions prior to and after online test implementation. Setting out processes for the collection of this information, and having the information itself, can save a significant amount of time, while also helping to ensure the validity of the test processes. Some suggestions are outlined below.

- Collect a population frame – that is, ensure that prior to testing some information about the size and characteristics of the student population being targeted for the test at each institution is known. This information can inform issues such as response rates (overall and by particular groups), enable the creation of unique logins for each student to increase security and increase transparency through examination of response bias and potential weighting of data.

- Include collection of basic demographic information in the administration of the test – in order to cross reference with the population frame, validate responses and increase confidence in the accuracy of the data collected.
- Collect detailed information at the institution level about the settings for the test administration, that is, the presence of invigilators, the location of test sittings, issues during testing, and so on. For AMAC,
this information was collected in a single, short online survey undertaken by each institution on completion of testing.
- In case of large numbers and summative tests, coordination of time frames of test administration.
- Secure sufficient computers or 'devices' to allow all students to sit the same test simultaneously or in other case ensuring sufficiently equated testlets (different versions of the assessment) to allow for valid comparisons.

2.2 Embedding items in existing exams

The embedding of 'common' items into examinations is a model relatively familiar to medical schools in Australia. In particular, the members of the Australian Medical Schools Assessment Collaboration (AMSAC) have been embedding items in exams for a number of years (Wilkinson, 2014). The recently developed Benchmarking project of the Medical Deans of Australia and New Zealand also aims to embed items rather than undertake stand-alone assessments (MDANZ, 2014). Similarly in the United Kingdom, the Medical Schools Council's Assessment Alliance (MSC-AA) has been undertaking such a project for a number of years <http://www.medschools.ac.uk/MSCAA/>.

For AMAC in 2013, institutions were offered the possibility of utilising AMAC items in their existing examinations rather than running AMAC as a stand-alone online test. The reason for this development was that AMAC leaders believed that given the emphasis on collaboration and versatility, the assessment items should be able to be used flexibly and to the benefit of institutions in any way they deemed appropriate, and that some schools would prefer to utilise the items in existing exams rather than organise an additional test administration.

By 2013, the 120 originally developed AMAC items had robust psychometric data to validate them as a result of the online test administrations in 2011 and 2012. The individual items, the psychometric data and advice in interpreting this (along with further support when required) was provided to institutions interested in using AMAC items in their exams. Four medical schools took up the option of embedding items into exams in 2013. Three of these schools embedded items in the exams of two cohorts, meaning that in total, AMAC items were used in seven institutional-based assessments. Each institution reviewed the AMAC items and chose those which were appropriate to their needs. No specific requirements were given to schools in terms of the number of items or grouping of items that were chosen for embedding, ensuring complete flexibility for the schools involved.

Each of the medical schools chose a selection of AMAC items to embed. The number of items used ranged from 15 items in one institution to 46 in another. For further context, the AMAC items contributed to between 18 per cent and 33 per cent of all items used in the examinations in which they were embedded. All of the examinations in which AMAC items were embedded were summative assessments.

The institutions using the AMAC items implemented them in different ways. The most typical use was in a major exam during the last or second last year of the degree. Other uses included embedding items in smaller 'end-of-rotation' exams, completed after a particular clinical rotation and therefore focused on a certain specialty.

Each institution involved in embedding AMAC items into their exams used a slightly different process for undertaking this task; generally, differences were a result of existing protocols for organising assessments each year.

One example of the basic process followed is outlined in the box below for indicative purposes and insight into the logistics of this process at the school level.

The embedded process – an example provided by University Y

University Y saw merit in the flexible approach offered to institutions participating in the AMAC Project in 2013 as it provided an opportunity to access an appropriate selection of high-quality assessment items that had already been validated and to obtain institution-level benchmarking information on students' performance on the AMAC items. The ability to sample questions from across a range of content, including content areas which the university has not traditionally assessed in multiple-choice question format, was seen as a benefit of drawing items from the AMAC bank.

Prior to participating in this phase of the project, faculty assessment-oversight committees agreed to the inclusion of up to 40 AMAC items to be embedded in the summative end-of-fifth-year multiple-choice papers, which comprise 220 items equally split between two three-hour computer-delivered examinations.

A selection of items for potential inclusion was made from the original AMAC item bank. Items were selected to complement the examination blueprint, both in terms of content coverage and task being assessed, e.g., diagnosis, management investigations, management treatment and so on. The selected items were subject to the faculty's usual review and standard setting processes prior to a final selection of 40 AMAC items being made.

The embedding process reduced the burden on faculty staff to develop new items and obviated the administrative time required to solicit new items and undertake a preliminary editorial review of those items before entering them into the faculty items database. From the university's point of view, embedding shared items in our summative end-of-year assessments was far easier to facilitate than administering a stand-alone formative online assessment (i.e., the 2012 AMAC pilot) as it did not necessitate the administration of an additional test.

One of the challenges of this process was merging the look, feel and structure of AMAC items into the existing formalised format of the university's own items. For example, the university recently adopted a policy of including the patient's age, gender and ethnicity as standard descriptors in all clinical scenarios in our assessments. The AMAC items did not have the same level of detail in every scenario (for example, many did not include ethnicity) and as such were potentially identifiable as having been drawn from a different source to the other items in the exam.

Related to this issue was how to respond if student performance on the shared items differed significantly from their performance on the 'native' items. Given the AMAC items were subjected to the university's standard selection and review process and linked to its examination blueprint, the likelihood of this being a problem was probably small. There was also some discussion around whether ethical approval was needed at the institutional level prior to administering the shared items to participating students as part of a summative (i.e., compulsory) assessment and around the need to inform students prior to their sitting the examination and/or receiving their individual feedback.

2.2.1 Benefits of an embedded approach

Based on reflections from the work described above, a number of benefits from employing an approach that offered institutions the option of embedding items in their own exams were identified. In summary, the benefits of sharing and involvement in an embedded approach are:

- efficiency in item development
- potential for higher quality
- an additional layer of transparency through benchmarking
- robust comparable outcomes based on summative assessment (where students are 'trying').

The benefit of sharing exam items is clear. Examinations are costly and burdensome to develop. Having access to a bank of items that have been developed by peers and colleagues from across the sector, have been used in testing on similar cohorts of students, and have robust psychometric data to accompany them is a potentially significant time and cost-saver. The inter-university progress test collaboration in the Netherlands has specifically identified the time saved through administration of their common assessment. Their collaboration identified that the time spent by academic and administrative staff on assessment development and implementation has more than halved following the formation of their collaboration (Schuwirth, Bosman, Henning, Rinkel, & Wenink, 2010).

In addition to this, while not covered specifically in this report, there are enormous benefits to institutions and educators from a professional development point of view through the involvement of the development of common assessment items in projects such as AMAC. Similar experiences in the long standing Netherlands collaborative project have also been identified, with members of that collaboration identifying an increase in assessment literacy in the participating institutions which has had its effect on raising the quality of other examinations in those institutions (Schuwirth, et al., 2010). The Dutch group has developed agreed-upon criteria for quality of items and organisation of a balance between decentralised and centralised quality assurance procedures. This idea has been mirrored by AMAC through the development of the item quality document for this project.

Propagating exams using well developed, common items is arguably a significant benefit in terms of transparency. The results allow internal benchmarking, and external validation of outcomes if necessary. The overall process also tends to build trust across institutions and partners. In the case of AMAC, institutions involved in the embedded assessment showed trust in the AMAC development process by agreeing to use the items in their summative examinations. Trust in the administration of the collaboration was also demonstrated through the willingness of institutions to provide the outcomes data back to the collaboration at the conclusion of the exams so that comparative data across schools could be developed.

In terms of the benefits of embedded over the formative online approach, from a practical consideration the embedded option does not add an additional test for schools to administer and students to participate in. As discussed above, in reality the benefit of a formative assessment with benchmarking is valuable, but in practice for some schools and students the option of using an existing exam for the implementation of common items is logistically a more realistic option.

A further benefit of embedding is that it allows for pre-testing items for scaling. This is a principle used by some assessment organisations and it allows for good item parameter estimates in reasonably authentic situations. As the students do not know which items are the real ones on the test and which are the pre-tested ones they can be assumed to take all the items with a similar degree of seriousness.

Data collected in a common assessment undertaken through an existing summative exam also provides a benefit in that schools are guaranteed full participation (in the summative AMAC online assessment participation was generally voluntary). This ensures the full picture is gathered in terms of the overall outcomes of the student cohort. Summative conditions in examinations also tend to 'bring out the best' in students; that is, students know that the results count and therefore are in the most part making a considered effort to perform well. Some basic analyses of the AMAC items undertaken summatively and formatively suggests this is the case, although further analysis on this is necessary.

2.2.2 Important considerations for an embedded approach

There are key challenges and possible pitfalls from using an embedded approach for collecting data for common assessment purposes. The discussion here draws on the experiences from AMAC, in which there was substantial flexibility in approach. This flexibility was provided to institutions to ensure ease of use for medical schools, and also to explore the different approaches and evaluate the benefits and challenges that arose. Core issues are outlined below:

- **Have an agreed set of common items.**
  This is necessary for enabling basic cross institutional comparison. In AMAC, baseline data for each item existed as a result of the use of the items in the online formative implementations, thus allowing for comparisons and providing the schools participating in the embedded approach with some flexibility in choice. But in reality it is clear from this exercise that in order to gather robust comparative data year-on-year through embedding in summative exams, a common set of questions for all participating institutions is the best approach.
- **Few items = limitations to output.**

While an agreed set of common items is important, gaining agreement across participating schools on a common set for everyone to embed is difficult. This is not impossible (and being displayed in the current Medical Deans Australia and New Zealand Benchmarking project and in the Australian Medical Schools Assessment Collaboration), but it tends to result in a limited number of items being used. A smaller number of items means that a more narrow perspective of student outcomes is available for benchmarking. Inevitably in this situation, the wider benefit of the process is limited to the specialities or disciplines focused on (in the case where a set of common items are concentrated in one area), or the data cannot provide a representative analysis of outcomes because the items were spread too thin (in the case where a range of areas are attempted to be covered by only a small number of items).

- **Scaling?**
  The challenge apparent from the points above is to use enough items to have some robust and detailed comparative data but to avoid complete uniformity and the 'spectre' of the summative assessment becoming a 'National Exam'. One way out of this problem is to undertake a detailed scaling study, which has the potential to produce a scale (for example an 'AMAC Scale') against which new items can be calibrated. This is a time-consuming research project, but would allow for a greater variety and number of items to be used for comparison.

  Scaling has been identified as an important issue for the Dutch progress test collaboration as well for various reasons. First, there is a home effect; items that are produced at a certain medical school tend to be answered better by students of that particular school (Muijtjens, Schuwirth, Cohen-Schotanus, & van der Vleuten, 2007). Currently this influence is mitigated by ensuring an equal contribution to the joint test by each participating institution. Scaling in the Dutch collaboration is also important to ensure equivalence of subsequent tests in the progress test. As it is very difficult to ensure that all tests are equally difficult some measures have to be taken to dampen the effects of variations of difficulty. Fortunately medical schools in the Netherlands are quite large, with 300 to 450 students per year class, and such cohorts of medical students are quite comparable. Therefore a norm-referenced scoring system based on the mean and standard deviation ensures that equivalent conclusions are drawn based on the results. Yet, as norm-referenced scoring systems are generally less palatable to stakeholders, currently studies are being conducted into scaling procedures based on item response theory, though these studies have seen limited success. Items that demonstrate sufficient fit with the models constitute only roughly 40 per cent of the total number of items passing the quality control processes. The relationship between the relevance of the item and item response theory scalability is currently under investigation in the Dutch collaboration, but preliminary results suggest that the more relevant the item the better the model fit with item response theory.

- **Recognise there is an administrative burden for institutions.**
  From a more practical perspective, the differences between the embedded option and the online full formative approach can be burdensome in terms of collating the results from the common items and

submitting them for the analysis and benchmark reporting. While this may seem a trivial task, in the context of providing results from a summative exam within deadlines for students, the additional burden of creating a stand-alone file for benchmarking purposes can be considerable. If summative exams were to be undertaken on computers rather than pen and paper, this would be much less of an issue.

3 POPULATIONS

One of the important facets of undertaking common assessments across schools is that there is the potential for the collection of data to be used for comparative and benchmarking purposes. Therefore, a key consideration in ensuring that such comparisons are valid is to ensure that the population of focus in the testing is consistent across the institutions involved.

In AMAC, population identification and uniformity across medical schools was particularly complex. Medical schools run medical programs at the undergraduate and graduate starting points, have different numbers of years required to completion, have different timing as to when key stages in the degree are undertaken, and have different points at which formal examinations are conducted. AMAC originally aimed to implement testing on final year medical students. However, given the structure of medical degrees, generally having a pre-clinical period followed by clinical years, with the later involving significant time off campus and on various rotations, the student population specified for the AMAC testing was generalised to those in the clinical years.

Data was collected from participating institutions to provide context to outcomes and for benchmarking and comparisons. In 2013, the populations of students ranged considerably across the eight institutions involved in the AMAC testing. Overall, the schools involved implemented the AMAC test in either the penultimate year or the final year of the Bachelor of Medicine/Bachelor of Surgery. The main reason given by institutions not running the AMAC assessment in the final year was that they had no formal examinations or graded assessment in that year.

The lesson from this project is that it is practically very difficult to identify cohorts across institutions that should be at exactly the same point in their studies. This has been a particular issue for this project, given the aim of focus on assessment of learning outcomes at the 'end point' of a medical degree. Given the clinical focus of the final years of medicine and the different approaches of medical schools to final year assessment (for example, some don't have exams in the final year), assessing a student at a point close to graduation has been difficult. An alternative approach in medicine is that taken by the Australian Medical Schools Assessment Collaboration (AMSAC), which develops assessment items for the end of pre-clinical training in a medical degree – essentially a common time in which assessment is uniformly taking place across medical programs. This is a more practical approach in terms of identifying a common cohort across schools, but does not operate as an 'end-point' assessment in the way that the AMAC project was focused.

However, this issue should not necessarily prevent the undertaking of such initiatives, but rather be understood and noted as a caveat – with appropriate information collected on the cohorts tested to facilitate interpretation of outcomes within the constraints of the implementation.

In disciplines other than medicine, it may be easier to define a 'graduating student' or to identify a practical time for implementing a common assessment. This makes such exercises less complex, but nonetheless, it is still important to identify such a point early-on and gain agreement from participating institutions.

4 TIMING

Linked to the decision making about which populations to involve in a common assessment process is the issue of the timing at which the assessment should be undertaken. Practical issues of the setting of

examinations, the existing timetables for result reporting and the course structures contribute to difficulties in identifying times in which the selected population can be assessed in each institution. Ideally for reporting purposes (especially for the students) it is best to have assessments across institutions taking place at a similar time in the year so that data can be collated, and reporting can be provided within a timeframe that ensures the feedback provided from the exercise is of use.

In AMAC, this was a constraint in the first year of testing. Institutions undertook the assessment at times in the year that best suited their programs, but essentially the schools and students had to wait until all participating institutions had completed testing before comparative statistics could be generated. In the second year of AMAC this was less of an issue because there was already baseline data for the AMAC items, enabling indicative reporting prior to the completion of testing across the schools. However, this was only possible because the same set of items was used in both years.

If the testing in a common assessment exercise is undertaken over a wide window of time, it is likely that institutions and (especially) students who wait the longest for their comparative results will find the process least useful. Therefore, the benefits of common assessment exercises, from the point of view of reporting and usefulness of results, can be heavily reliant on the length that the testing window across institutions is open. Identifying a short time within a semester that works for all institutions involved is therefore a key consideration in the development of collaborations focused on benchmarking through common assessment.

5 REPORTING

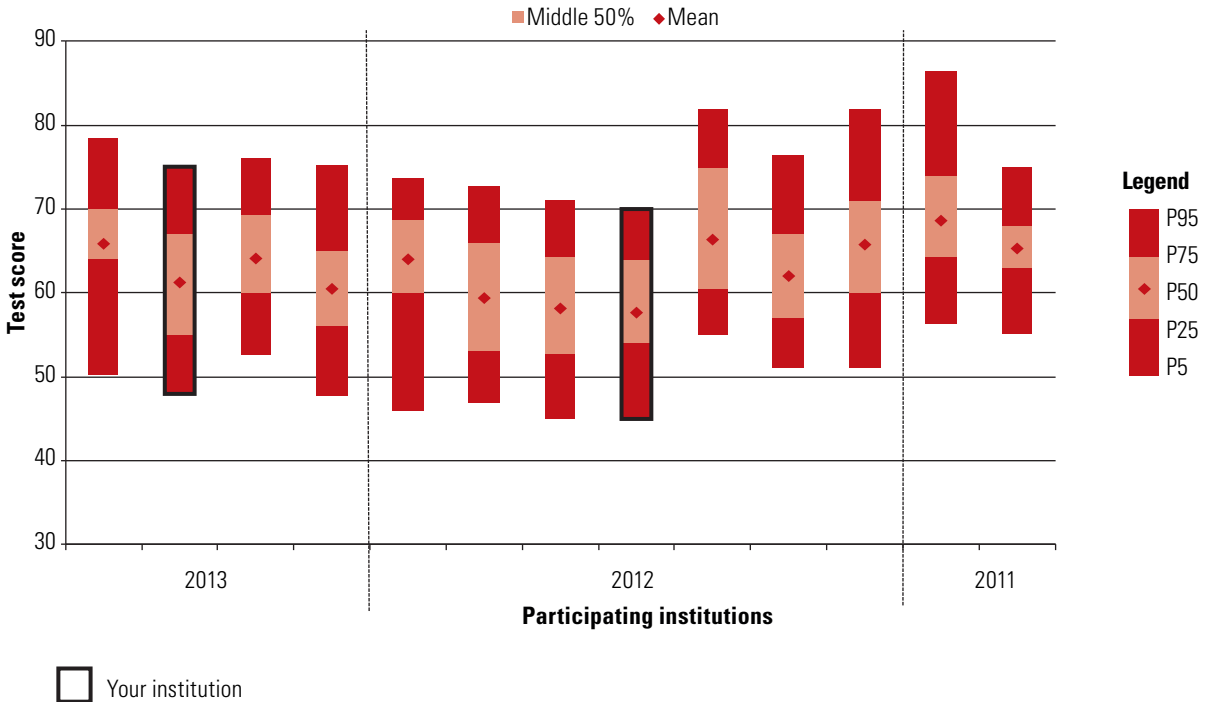
Reporting the outcomes of common assessment is an important facet of such a project. Another document produced for this project (*Governance Models for Collaborations involving Assessment*) covers details relating to data sensitivities in dissemination and governance issues – these are essential considerations when it comes to reporting, but not dealt with here. This discussion explores some ideas relating to effective reporting of outcomes that have been identified through the development of the AMAC project. This section essentially provides examples from AMAC in relation to student reporting and to institution reporting.

A key focus for future collaborations, regardless of discipline, is to begin to identify what schools and faculties would benefit from most through reporting at a very early stage in the development process. This is important for:

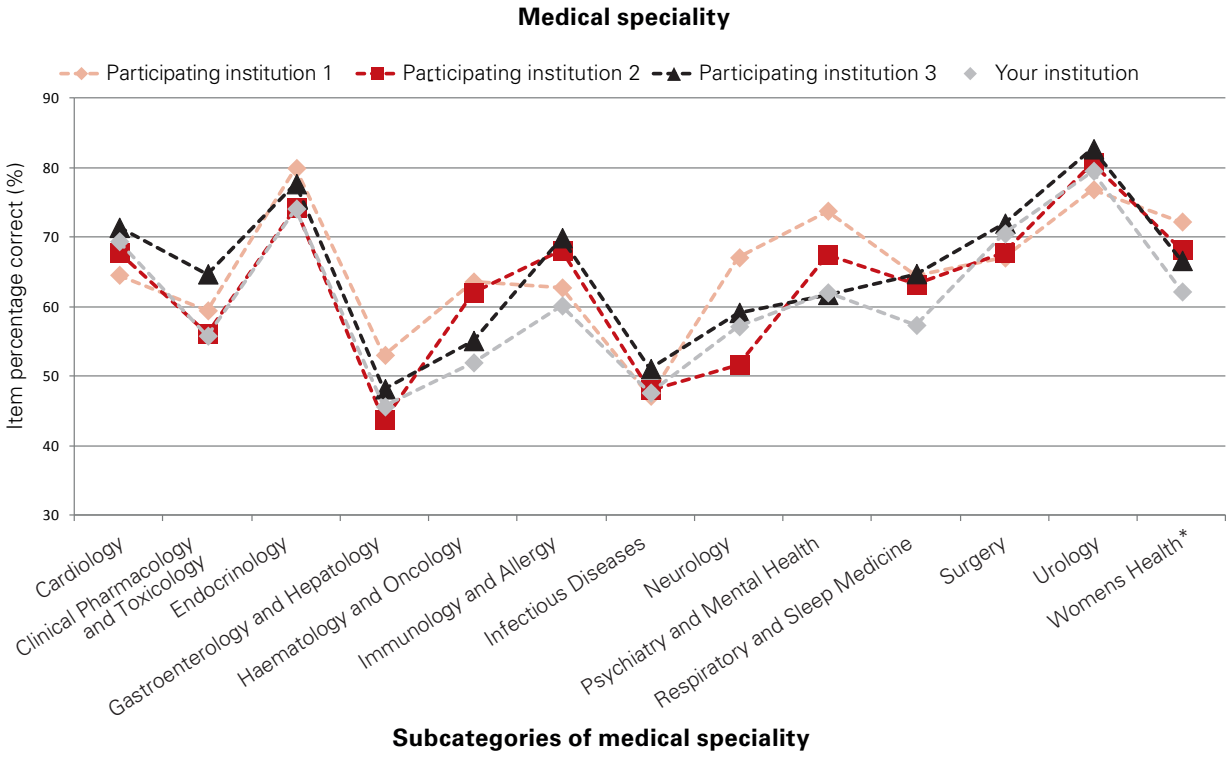
- ensuring that all involved understand what the collaboration will produce
- informing the design and development of the assessment.

5.1 Student reports

Providing student feedback was considered a vital element for AMAC. Students who participated in the formative online AMAC assessment were each provided with an individualised report detailing outcomes from the assessment in a range of areas and benchmarked to their fellow classmates as well as to other participants across medical schools in Australia and New Zealand.



Given the number of items and the fact that each AMAC item was mapped to the framework and 'tagged' in various ways, there was good opportunity for reporting to students on a range of areas. An example of a



*Women's health includes Obstetrics and Gynaecology

table that summarised a student's outcomes on the assessment is provided below (Table 1). As part of the report, information on interpreting the output is provided, as are caveats to this interpretation and a contact name and number for students to access should they have further queries.

It is not so much the detail in the numbers that is useful for the purposes here, more the overall indication

on the type of output that was generated for individual students from the online implementation of AMAC.

Feedback from the students who participated and received a report was overwhelmingly positive. For some, the fact there was a report at all was a benefit. The importance of providing student feedback is critical. Lessons from other projects, such as the OECD's Assessment of Higher Education Learning Outcomes (AHELO) Feasibility Study (Edwards, 2013) showed that motivating students for participating in a voluntary common assessment can be difficult.

As mentioned in the section above, timing of the testing in institutions did influence the delivery of reports to students. In the first year of AMAC testing, some student cohorts had to wait a month or two to receive their reports due to the fact that benchmark data needed to be collected from other participating schools before it could be reported. By the second year of testing, the baseline outcomes from the first year were established and students were sent their reports within a couple of weeks of testing. This was important because for a number of institutions the AMAC assessment was conducted as a 'practice' and 'stocktaking' exam formatively in the weeks leading up to summative exam periods. Therefore, the reporting from AMAC could be used by students as one tool for identifying potential areas for revision focus.

Ideally, in the case of formative, online assessment, such reports would be generated instantly for students on completion of the test. This kind of feedback opportunity is possible once baseline data has been collected and should be considered as an important tool to be developed in future collaborations.

In case of summative examinations, feedback value may be limited as students typically tend to ignore the feedback and look at whether they passed or failed only. Combining formative feedback with summative tests is therefore tricky. Two lessons have been learnt from the Dutch progress test experience. First, if you want students to really take up the feedback, their own analysis of the results with learning plans need to be part of the examination system. At Maastricht University, for example, students will have to analyse

Table 1 Example output from AMAC Student Report based on online sitting, 2013

Category	Subcategory	Your Score	Your % correct	Your school % correct	Full cohort* % correct [95% confidence]
Medical speciality	Cardiology	5 of 7	71%	65%	63% [61, 65]
	Clinical Pharmacology and Toxicology	8 of 12	67%	56%	57% [56, 58]
	Endocrinology	8 of 9	89%	73%	72% [71, 73]
	Gastroenterology and Hepatology	6 of 9	67%	42%	42% [41, 43]
	Haematology and Oncology	9 of 11	82%	62%	59% [58, 60]
	Immunology and Allergy	3 of 4	75%	68%	67% [65, 69]
	Infectious Diseases	2 of 5	40%	54%	52% [50, 54]
	Neurology	4 of 10	40%	53%	55% [54, 56]
	Psychiatry and Mental Health	4 of 5	80%	77%	74% [72, 76]
	Respiratory and Sleep Medicine	2 of 4	50%	63%	63% [61, 65]
	Surgery	1 of 3	33%	61%	62% [60, 64]
	Urology	4 of 4	100%	78%	77% [75, 79]
	Women's health (includes Obstetrics and Gynaecology)	6 of 10	60%	63%	62% [61, 63]
	Other (incl. Dermatology, Emergency Medicine, Neonatal) Ophthalmology, Rheumatology)	6 of 7	86%	55%	58% [56,60]
	Total	**68 of 100**	**68%**	**60%**	**60% [59,61]**
Clinical context	Decision making	16 of 22	73%	62%	62% [61, 63]
	Making a diagnosis	33 of 48	69%	62%	61% [60, 62]
	Medical knowledge recall	6 of 10	60%	55%	55% [54, 56]
	Medical testing	10 of 12	83%	64%	65% [64, 66]
	Prescriptions	2 of 6	33%	46%	48% [46, 50]
	Other (incl. Interpreting data, Patient assessment)	1 of 2	50%	43%	47% [44, 50]
	Total	**68 of 100**	**68%**	**60%**	**60% [59, 61]**

Category	Subcategory	Your Score	Your % correct	Your school % correct	Full cohort* % correct [95% confidence]
Professional practice	Emergency management	12 of 14	86%	60%	59% [58, 60]
	Patient assessment	36 of 56	64%	60%	59% [58, 60]
	Patient management	13 of 20	65%	63%	64% [63, 65]
	Other (incl. Patient interaction Not applicable)	7 of 10	70%	58%	57% [56, 58]
	Total	**68 of 100**	**68%**	**60%**	**60% [59, 61]**

* Full cohort scores are based on ll administrated items for this student's particular tst rotation. mean percentage correct are listed alongside 95% confidence intervals (displayed in square brackets)

the breakdown of their scores and relate them with other elements of the examination program to define realistic and effective learning goals. Therefore students cannot escape using the feedback provided (a query-based feedback system is therefore built to facilitate this). Second, students are allowed to criticise questions and ask for them to be removed from the test, provided they do so with clear argumentation and support from the scientific literature. Typically two groups of students take up this challenge: those who score high but see it as their academic duty to contribute to the content validity of their examinations, and those who need some extra points because they fear they might fail. Regardless of the motive, students are incentivised to re-read the questions and learn from their mistakes.

This brings about the issue of whether or not to release the items after the test. Naiveté in test developers often leads them to believe that if they don't release the items there will not be a black market in old items, whereas invariably these black markets exist. The general mantra is that the most secure item bank is the one that holds so many items that it does not pay for a student to memorise all of them. Collaborative test development is the optimal way to quickly build databases that are really secure (instead of only being believed to be secure).

5.2 Institution reports

Institutions, and more specifically the schools and faculties that participate in common assessment exercises, are the core constituents for reporting of outcomes. In AMAC, institution reports were developed based on the data available, with an emphasis on providing important benchmark information but with care to ensure de-identification of institutional outcomes and to avoid over-emphasis of statistically insignificant differences.

Care was taken to clearly re-state for institutions the cohort that they had involved in the testing and to provide various caveats around results, including differences in populations and timing of the implementation of assessments across participating institutions.

Some examples of the institutional report from 2013 output are provided below. Figure 1 shows a boxed distribution chart highlighting the outcomes of an institution in comparison with other institutions involved in AMAC. This particular example shows the outcomes for students from this institution in 2013 and in 2012. The box-plot allows for the display of the spread of student results rather than a sole focus on the average outcomes for schools involved. This is intended to reduce the likelihood of simplistic conclusions being drawn about overall performance of a school and rather to promote a more considered appreciation of the range in student performance within and across participating schools.

As with the student reports for AMAC, consideration was given in the institution reports to providing some insight to schools about the relative outcomes of students across an agreed set of disciplines and specialities. Figure 2 displays how this was done by medical speciality for participating schools in 2013. As with the other reporting in this document, the figure here is used for illustrative purposes to highlight the potential of such reporting.

Figure 1 Comparison of test score distributions between institutions 2011-2013 (n. 940 students, online formative implementation)

Figure 2 Comparison of item percentage correct between institutions, medical speciality sub-categories, AMAC 2013

6 CONCLUSIONS

The purpose of this document is to provide insight into the processes undertaken in the AMAC project for implementing the common assessment items that were developed. The intention is that the lessons that were learnt through this collaboration can be adopted by other collaborations (regardless of discipline) and adapted to help in building a foundation for a successful project. The pathway taken by the AMAC group is not necessarily meant to be displayed here as best practice, but the outcomes from the work and the processes followed can be adapted to suit other common assessment projects that may perhaps lead to best practice and a more streamlined model for implementation.

As a conclusion to this document, some of the key decision-making steps in implementation of a common assessment are highlighted below. While some of the steps described here are slightly beyond the realm of the detail contained in this document, the focus is on assessment implementation issues. The overall picture is intended to be a guide for decision-making processes of future assessment collaborations.

1 **Purpose**	• Identify purpose of collaboration • Determine desired output and outcomes
2 **Assessment design**	• Develop robust, well-targeted design with intention to produce assessment capable of achieving desired outputs
3 **Models**	• Identify best form of test implementation (online, paper, embedded, etc.) • Formative, summative, or both? Explore the pros and cons of each
4 **Population**	• Identify target populations/cohorts for participation – know the cohort • Identify appropriate timing for assessment/s
5 **Support institutions**	• Provide support with administrative tasks involved with implementation • Provide support in interpretation and understanding of outcomes
6 **Reporting**	• Conduct base reporting on initial desired outputs • Recognise the limitations, provide context and caveats to interpretation

7 REFERENCES

AMAC. (2012). *Australian Medical Assessment Collaboration: Assessment Framework*. Melbourne: Australian Medical Assessment Collaboration.

Edwards, D. (2013, December 2013). *The AHELO experience – implementation, outcomes and learning from an Australian perspective.* Paper presented at the NIER Tuning-AHELO Symposium, Tokyo.

Edwards, D., Wilkinson, D., Canny, B., Pearce, J., & Coates, H. (2014). Developing outcomes assessments for collaborative, cross-institutional benchmarking: progress of the Australian Medical Assessment Collaboration. *Medical Teacher, 36*(2), 139–147.

MDANZ. (2014). Assessment Benchmarking MDANZ website. Retrieved April, 2014, from http://www.medicaldeans.org.au/projects-activities/assessment-benchmarking

Muijtjens, A., Schuwirth, L., Cohen-Schotanus, J., & van der Vleuten, C. (2007). Origin bias of test items compromises the validity and fairness of curriculum comparisons. *Medical Education, 41*(12), 1217–1223.

Schuwirth, L., Bosman, G., Henning, R., Rinkel, R., & Wenink, A. (2010). Collaboration on progress testing in medical schools in the Netherlands. *Medical Teacher, 32*(6), 476–479.

Wilkinson, D. (2014). A new paradigm for assessment of learning outcomes among Australian medical students: in the best interest of all medical students? *Australian Medical Student Journal, 4*(2).

Wilkinson, D., Canny, B., Pearce, J., Coates, H., & Edwards, D. (2013). Assessment of medical students' learning outcomes in Australia: current practice, future possibilities. *Medical Journal of Australia, 199*(9), 578–580.