Cross-Institutional Comparison of Mechanics Examinations: A Guide for the Curious

Thomas Goldfinch
University of Wollongong, tomgold@uow.edu.au

Anna L. Carew
University of Tasmania, carew@uow.edu.au

Anne Gardner
University of Technology, Sydney, Anne.Gardner@eng.uts.edu.au

Alan Henderson
University of Tasmania, alan.henderson@utas.edu.au

Timothy J. McCarthy
University of Wollongong, timmc@uow.edu.au

See next page for additional authors

http://ro.uow.edu.au/engpapers/3818

Publication Details
Authors
Thomas Goldfinch, Anna L. Carew, Anne Gardner, Alan Henderson, Timothy J. McCarthy, and Giles Thomas

This conference paper is available at Research Online: http://ro.uow.edu.au/engpapers/3818
Cross-institutional Comparison of Mechanics Examinations: A Guide for the Curious

Thomas Goldfinch
University of Wollongong, Wollongong, Australia
tom_goldfinch@uow.edu.au

Anna L Carew*
Australian Maritime College, University of Tasmania, Launceston, Australia
anna.carew@utas.edu.au

Anne Gardner*
University of Technology, Sydney, Australia
Anne.Gardner@eng.uts.edu.au

Alan Henderson*
University of Tasmania, Hobart, Australia
alan.henderson@utas.edu.au

Timothy J McCarthy*
University of Wollongong, Wollongong, Australia
tim_mccarthy@uow.edu.au

Giles Thomas*
Australian Maritime College, University of Tasmania, Launceston, Australia
gthomas@amc.edu.au

(* In alphabetical order)

Abstract: This paper describes a process used for, and interim findings of a comparison of final mechanics exam papers for first year engineering mechanics courses at the University of Wollongong, the University of Tasmania, the University of Technology, Sydney, and the Australian Maritime College. The process developed for the purpose emphasized a transparent and sequenced approach to comparing the concepts included in each exam paper, as well as the perceived level of difficulty of exam questions. The exercise was carried out remotely, using readily available communications technology, including telephone, email, and Skype teleconferencing. This process is an example of a simple, easy to implement, and readily transportable approach to cross-institutional peer review of assessments, and an effective way of enhancing collaborative links between engineering educators.

Introduction
Engineering mechanics is a key foundational subject area for numerous engineering disciplines. This topic has consistently proven a challenging area of study for students (Dwight & Carew, 2006; Rezaei, Jawaharlal, Kim, & Shih, 2007), and a cause for much head scratching among the academics responsible for teaching it. Four engineering schools (University of Wollongong, University of Tasmania, University of Technology, Sydney, and Australian Maritime College at UTas) are currently involved in a substantial project to address these problems by identifying factors that may predict poor student performance in engineering mechanics.

As part of this project, it was necessary to analyse student’s actual work in final examinations to quantitatively identify the most common areas of difficulty for students. Four Engineering Statics and/or Dynamics exam papers from the four participating schools were collected,
redistributed and evaluated by the participating academic from each school. An evaluation of the results was also undertaken.

The main purposes for conducting this comparison included:

- Establishing at the very beginning of the research, what variations may exist in mechanics assessment between the four schools.
- Identifying similarities between the exam papers which may enable the comparison of student responses.
- Identifying variations in the academics’ perception of each question, with respect to level of difficulty, concepts assessed etc.
- Comparing differences in terminology, question wording, and types of problems used in each exam.
- Building the collaborative relationship within the research group.

The authors propose that the process described here for evaluating and comparing final examination papers offers opportunities for academics to learn from each other and receive supportive, critical peer review of their assessment approach. The process could also be used as an elementary form of external quality review. This is a practice that is common in many countries such as the UK (QAAHC, 2004), but is not widespread in Australia.

**Approach**

To begin the analysis of students’ work, the researchers first had to obtain the necessary resources. Final exams were the most obvious choice for this analysis as transcripts from these examinations were readily available from each institution. Other assessment methods used such as quizzes, assignments, tutorial hand-ins, and lab reports were less consistent between each institution. To conduct a fair and useful comparison, final exams provided the best platform in terms of comparable content, assessment conditions for students (eg. high pressure environment with finite time allowance), assessment weighting (Nightingale, Carew, & Fung, 2007), and assessment task style. In addition, final exams are generally used to test students understanding and ability over a wide range of topics taught during the course and should provide a reasonably comprehensive summary for analysis. It is important to note at this point that in comparing the topics examined between institutions, some differences are to be expected due to the fact that final examinations cannot cover all subject content and thus, provide a sample of student learning (Nightingale et al., 2007).

The project team, consisting of the authors of this paper, was unable to meet as a group due to distances, and time and funding constraints. This is a situation many Australian teaching academics can relate to, and may be an inhibitor to regular inter-institutional peer review of commonly used assessment tasks. Low cost, web-based teleconferencing applications such as Skype (and Skypeout) offer a useful substitute for the time consuming travel necessary to conduct face-to-face meetings. This medium, along with regular email correspondence enabled satisfactory communication throughout the exercise.

An initial teleconference was set up to discuss and agree on what the comparison should aim to identify. The project team agreed that:

- key concepts;
- skills, and;
- the difficulty level

of each examination question should be identified. It was agreed that each teaching academic would perform an analysis of these factors for each of the exams, including their own. This meant that all four exams were examined by all four participating academics.

During early teleconferencing, it became apparent that a roadmap of the exam comparison and analysis would be useful to negotiate and clarify what needed to be done. The roadmap needed to be clear and visual to help the project team maintain focus on what was required from the comparison and its overall purpose. Deadlines and meeting dates were also included.
in the roadmap to provide a single document containing all the relevant information for future reference. The roadmap developed is shown in Figure 1, and shows the basic approach devised for the whole exam analysis process. The comparison exercise described in this paper comprised stages 1 & 2 of the guidelines.

<table>
<thead>
<tr>
<th>Stage 1 – Independently Evaluate Exam Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Analyse each question in each of the four exam papers and identify key concepts considered.</td>
</tr>
<tr>
<td>2. Comment on the difficulty of each concept in each question (this can be compared with transcripts to determine alignment of lecturers’ perceptions, and actual difficulty experienced by students).</td>
</tr>
<tr>
<td>3. Note concepts that are similar in each exam paper (ie. Q1(b) in University of Example paper and Q3(c) in University of Demonstration paper require students to understand concept X).</td>
</tr>
<tr>
<td>4. Email findings to Tom by 9am, 13th Feb 2008.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stage 2 – Agree Common Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Share interpretations of exam papers with the group (during Skype meeting: 11am-1pm, 13th Feb, 2008).</td>
</tr>
<tr>
<td>2. Discuss and agree on concepts assessed by each question.</td>
</tr>
<tr>
<td>3. Discuss and agree on concept similarities between questions.</td>
</tr>
<tr>
<td>4. Discuss and agree on difficulty levels of concepts.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stage 3 – Formulate Matrices to Analyse Exam Scripts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Identify theories/taxonomies that describe the thought/analytical processes needed to complete the exam questions (in accordance with concepts identified in stage 1).</td>
</tr>
<tr>
<td>2. Develop matrices to evaluate the solutions presented by students in exam transcripts.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stage 4 – Analyse Exam Transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Collect Transcripts from each university.</td>
</tr>
<tr>
<td>2. Identify key conceptual, procedural, and other errors made by students (in accordance with concepts identified in stage 1).</td>
</tr>
<tr>
<td>3. Evaluate students’ solution approach against matrices developed in stage 3.</td>
</tr>
<tr>
<td>4. Identify key proficiencies/deficiencies evident in students’ problem solving approaches.</td>
</tr>
</tbody>
</table>

**Ethics Application**

1. Compile and submit to UOW ethics committee. |
2. Submit approval to UTS and UTas ethics committees for ratification.

**Figure 1 Exam Paper analysis roadmap**

Identifying the concepts in each of the exam questions and the similarities between the different exam papers was found by participants to be a fairly straight forward task, with a pleasing degree of consensus between each analysis. However, identifying the degree of difficulty for each exam question, in the initial stage, was less simple. The project team did not initially discuss and agree on standards for identifying the degree of difficulty of questions. This resulted in variation in the way each academic defined difficulty levels, with a few opting not to define these levels in stage 1 due to confusion. A basic scale of 1 – 3 (straight forward – moderate – challenging) for identifying difficulty levels was then established to rectify this in stage 2.

Stage 2 was commenced with another teleconference. The team hoped to discuss and agree on the common concepts between the questions and determine a common set of difficulty levels for each questions, though this proved to be more difficult than anticipated. Discussion on these topics was lively during the teleconference, however it soon became apparent that teleconferencing was not the ideal medium for focusing ideas into a final outcome. It proved
to be difficult to explain the finer details of each question without graphical aids or sketches, and discussing four different exam papers in the midst of a five-way discussion was simply impractical.

As a result, points raised during the meeting and the documented exam evaluations were compiled into a single document and distributed to the participating academics for comment. Key concepts and skills identified were combined in the document as there was common agreement on these among the academics. Difficulty levels were included as individual contributions rather than being combined into an average as the project team were interested to see how any differences in academics’ perceptions of difficulty related to actual student performance in the subsequent exam transcript analysis (stage 4). This text/email based method of finalizing stage 2 proved to be much more effective in refining and elaborating the final exam paper evaluations.

What did we find?

Differences and Similarities

In general, there were some substantial differences between the four exam papers, particularly considering that all were from first year mechanics courses. The first major difference was that two of the institutions opt to separate Statics and Dynamics into two courses while the others combine the two topic areas into a single course in the curriculum. The exams from schools A and D included a mix of statics and dynamics problems, while the school B exam focused only on statics and the school C exam was predominantly dynamics. This limited the number of similarities across all four exam papers, as demonstrated below in Table 1.

Table 1 Concept similarities across the four exam papers (only concepts existing in more than one paper are shown here)

<table>
<thead>
<tr>
<th>Common Key Concepts</th>
<th>School A</th>
<th>School B</th>
<th>School C</th>
<th>School D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Force/moment resolution</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Force/shear/BM</td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td>✔️</td>
</tr>
<tr>
<td>Centroid &amp; moment of inertia</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td></td>
</tr>
<tr>
<td>Acceleration/velocity/distance</td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear momentum</td>
<td>✔️</td>
<td></td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Stress/strain</td>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td></td>
</tr>
<tr>
<td>Conservation of energy</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Angular dynamics</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Truss analysis</td>
<td>✔️</td>
<td></td>
<td>✔️</td>
<td></td>
</tr>
</tbody>
</table>

The only assessed concepts found to be common across all four exam papers were basic force and moment resolution. Concepts shown to be common between the papers assessing statics were internal forces (shear force and bending moment), and moment of inertia. For the questions examining dynamics, the relationships between acceleration, velocity and distance, as well as conservation of energy were shown to be important to all. This highlights the basic concepts in statics and dynamics that students are commonly expected to have mastered by the end of their first year of study in engineering. In further research, the project team will use these commonalities to investigate whether there are any differences between institutions in students’ ability to master these important concepts, and if so, identify what factors may be at the root of these differences. The researchers would also like to establish whether there are any differences between courses that combine statics and dynamics, or those that separate them.
When comparing such details as question wording, terminology, and problem types, there were no substantial differences noted. It appeared that the types of problems presented in all the exams were of similar nature and variety. This confirmed for the academics that the methods used for setting each of the exams were basically aligned with each other, and that the student responses to exam questions were likely to be comparable.

**Difficulty levels**

When evaluating the difficulty levels of each question, there were differing views between the academics over several questions. Such disagreements in difficulty levels existed for just over a third of the questions overall. As a result, the question has now been raised: Why are the academics perceptions of question difficulty so varied?

An answer proposed by one academic is that assigning a difficulty level to each exam question can be challenging due to the context of the peer review. An examiner sets a question to provide a suitable summative assessment of student learning (Biggs, 1999). The examiner must consider what students have been taught during delivery of the unit, and with this in mind formulates questions at an appropriate difficulty level to allow students to demonstrate what they have learned. This perception of difficulty level by the examiner may also be influenced by student feedback from formative assessment tasks. This creates a degree of uncertainty for an external person reviewing a paper, without knowledge of the unit delivery.

Another challenge in assigning a difficulty level relates to how examinations are used to grade student performance from a minimum level to an exceptional level of achievement. This grading function may not necessarily be done in each exam question, since most of the exam papers required students to answer all questions. In this case some questions may have intentionally been made ‘easy’ while others ‘hard’. However in one paper, students were given a choice of questions to answer: in this case the examiner must be careful to provide questions of equal difficulty. It is possible to structure each question with some parts designed to provide a modest level of difficulty, thus allowing students to demonstrate a satisfactory level of achievement, while more difficult parts allow students to demonstrate a high level of achievement. The style and structure in which the examination questions are written influences the overall difficulty level. Moreover, the style in which each academic tends sets their own exams may influence their perceptions of difficulty in others’ exams.

So here we may raise another question: Given that the academics’ perception of difficulty may rely on previous experience and their own style of teaching, could students’ ability to comprehend the question vary depending on their learning experiences in different tutorial groups?

Another intriguing finding was the variation in the proportion of the questions in each exam paper that elicited differing difficulty ratings. For the school A exam, disagreements existed over six of the eight questions, while with school C there was just one from 14 questions. Schools B and D incurred two and four disagreements respectively from seven questions each. It appeared that the two exams that focused on either Statics or Dynamics (rather than both) elicited greater consensus as to the perceived level of difficulty. The reason for this is variation is unclear at this stage. It is hoped that in depth analysis of student responses to the questions may shed some on light why this particular variation exists.

In terms of comparing the average difficulty rating of each paper, the results are less obvious. Only very small differences exist between the four papers, with all being rated as moderate overall. However, more significant differences do exist between the papers in terms of similar questions (see figure 2). The truss analysis question in the school D paper was rated straight forward to moderate, while the truss analysis question in the school B paper was rated challenging by all academics. The project team intends to identify whether differences such as these are reflected in the students’ performance and their subsequent impact on pass rates.
Figure 2 Average difficulty levels compared between common questions

The Academics’ View

Below are each academics’ personal views on what they have taken from the exam comparison exercise. Each describe opinions of the exercise, and how it may have affected their own work. The views expressed here also echo many view expressed by academics involved in the external examiner system in the UK (Bjørn, Ellen, & Nils Henrik, 2008; Hannan & Silver, 2004), further emphasizing the benefits of exam paper comparison.

Academic A: As a lecturer of engineering mechanics, I found identifying key concepts in stage 1 relatively easy as many of the questions clearly fell into familiar categories of a ‘centroid’ problem or a ‘static equilibrium’ problem, etc. The majority of questions, especially in statics, were of the convergent type where the examiner clearly expected a unique answer by means of a preferred solution method.

I was surprised by the high level of similarity in the style of questions, content and concepts. In a broader sense, it is suggestive of a high degree of commonality between the teaching programs. Establishment of this ‘common’ ground is very encouraging, as the workgroup can now focus on the task of improving teaching and learning of engineering mechanics.

Academic B: I welcomed the opportunity to compare topics included in the various final exams as well as the level of difficulty of the question posed. This process initiated reflection on my own teaching, about what I regard as the most important learning outcomes in this subject and why I regard them as the most important. Because the process used the final exams as the vehicle for comparison, it focused attention on how exam questions were worded – was the question unambiguous? Did the wording lead the student to provide the answer I was expecting or was another interpretation possible? I try to set exams so that an ‘average’ student can earn 50%, but I don’t want this ‘average’ student earning say 80%, so there have to be some discriminators as well. As the only institution teaching just Statics to its Civil students, the benchmarking exercise has lent some weight to a proposed curriculum renewal process in the mechanics area. One of the main benefits of the exercise was the establishment of a small ‘community of practice’ in teaching engineering mechanics to, hopefully, generate
mutual capacity building in this subject area which is fundamental to so many subsequent engineering subjects.

**Academic C:** Academics can tend work very much in isolation. The background to our expectations of required teaching material and knowledge can derive from our own experiences as undergraduates. Are we then stuck in a time warp? Do we teach what we were taught; at the level that we were taught (though usually lamenting that it was harder back in our day)? So this project was a great opportunity to see what others are teaching at universities around the country and what their expectations are of students at the end of the unit.

Ranking the difficulty of the questions was the most challenging part of the study. I was a little nervous that I would rank a question as extremely difficult only to have all my co-researchers mark it down as a very easy! Though the whole project has progressed in a very supportive way and this initial study has helped establish good collaboration for further work. Once all the results were circulated amongst the team it was clear that generally the exam papers have been set at a similar level, but each contains questions with a range of difficulty.

I found taking part in this study to be a valuable experience and it will help me with my teaching in the future; for example I’ve picked up some nice ideas from my co-researchers that I’ll certainly be using.

**Academic D:** The opportunity to have colleagues from other institutions comment on my exam questions has been very valuable. It has also been informative to review papers from these colleagues. Comparing difficulty levels of the questions has led me to better understand how to create different levels. The exercise has also shown that the background of the exam setter influences both the type of question set and how they perceive its difficulty. It is comforting to note that while there are subtle and significant differences between the examinations papers, there is general agreement that they represent an equivalent set of tests of student’s knowledge and understanding of Engineering Mechanics.

Sharing our reflections on what we have done during the benchmarking process has added to my understanding of how to ask the right question. It takes considerable design effort to formulate questions that truly test a student’s grasp of a concept without being blurred by other aspects of the problem. Predicting what mistakes demonstrate specific misunderstandings is another factor in question design.

**Discussion**

While the overall outcomes of the exercise have been positive, and the findings useful, there are some additional exercises that could further enhance the outcomes. One such addition to the comparison would be the marking criteria used to grade each exam. A comparison of the approaches used, the marks allocated for each question, and where marks are deducted (eg. incorrect or missing units on answers, scrutiny of free body diagrams, quality of shear force and bending moment diagrams etc.) would contribute to a more complete comparison of the whole final examination process. It may also provide more insight into the differences in performance of students at each institution, and indeed, what is expected of them.

The comparison may also be easier to conduct through face-to-face meetings. While it is possible to conduct these remotely, the team found that comparison of the more visual aspects of the papers was difficult without the opportunity to explain things graphically. Ideally, each academic would review the papers independently, compile notes and points for discussion, then meet once to finalise and agree on the similarities and differences between the exams. It may also be possible to enhance teleconferencing with other technologies such as electronic whiteboards, document cameras or sketch pads as a substitute for face-to-face meetings.
Conclusions

The main motivation for the cross-institutional comparison of mechanics exams described in this paper was to provide a platform for further research on difficulties experienced by students in introductory mechanics. This first stage of the research, however, developed into a useful process in itself. In particular, the process allowed the four participating academics to give and receive considered peer analysis and critique of their approach to assessing students’ learning in the subjects they teach. Given Australia has little culture of cross-institutional benchmarking at the level of examinations, this collective review allowed participating academics to reflect on the assessment design. They could identify where exam questions may have been assessing too many concepts at once or providing too little opportunity for students to demonstrate their breadth (or narrowness) of understanding. The process described above worked for the situation and objectives of the participating project team, but others might chose to adapt it to their own needs and contexts.

The general question that is provoked by all of the findings noted above is ‘what impact do these differences and similarities have on learning’? Given that each of the exams presented here constitute around half of the students’ final mark in the course, the impact of the differences found could be significant for students. With such a reliance on final exams as a reflection of what students have learnt, ensuring that these exams are structured in the most appropriate way is crucial.

References


