Improving Learning and Developing Professional Judgment in Large Classes Through Collaboration and Self and Peer Assessment

Final Report 2014

ALTC Teaching Fellowship

Dr Keith Willey

University of Technology, Sydney

Collaborative Learning Activity Framework

<sparkplus.com.au/factors>
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 <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 <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
Location code N255EL10
GPO Box 9880
Sydney NSW 2001

learningandteaching@education.gov.au

2014

ISBN 978-1-74361-669-7 [PRINT]
ISBN 978-1-74361-670-3 [PDF]
ISBN 978-1-74361-671-0 [DOCX]
Acknowledgements

I would like to thank my wife Helen and sons Aaron, Joel and Daniel who saw me a little less for 18 months. I dedicate this work to them without whom life, while a gift would mean a lot less.

I would also like to acknowledge and thank my colleague Anne Gardner with whom I collaborated while undertaking much of the research disseminated through this fellowship.

Finally I would like to thank Mike Howard and Siobhan Lenihan for their assistance and support.
## List of acronyms used

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABDC</td>
<td>Australian Business Deans Council</td>
</tr>
<tr>
<td>ACED</td>
<td>Australian Council of Engineering Deans</td>
</tr>
<tr>
<td>ACU</td>
<td>Australian Catholic University</td>
</tr>
<tr>
<td>ALTC</td>
<td>Australian Learning and Teaching Council Ltd</td>
</tr>
<tr>
<td>IF-AT</td>
<td>Immediate Feedback Assessment Technique</td>
</tr>
<tr>
<td>RMIT</td>
<td>RMIT University</td>
</tr>
<tr>
<td>SAPA</td>
<td>Self Assessment to Peer Assessment Factor</td>
</tr>
<tr>
<td>SPA</td>
<td>Self and Peer Assessment Factor</td>
</tr>
<tr>
<td>SPARK</td>
<td>Self and Peer Assessment Resource Kit</td>
</tr>
<tr>
<td>SPARK\PLUS</td>
<td>Assessment and Evaluation Software Tool</td>
</tr>
<tr>
<td>UNSW</td>
<td>The University of New South Wales</td>
</tr>
<tr>
<td>USYD</td>
<td>The University of Sydney</td>
</tr>
<tr>
<td>UTS</td>
<td>University of Technology, Sydney</td>
</tr>
</tbody>
</table>
Executive summary

Introduction

Despite recent momentum for assessment to change from ‘assessment of learning’ to ‘assessment for learning’ many academics view assessment primarily as a means for students to demonstrate what they have learnt. Conversely, learning-oriented assessment embeds learning in assessment, reconfiguring its design to emphasise the function of learning. Students should also be involved in the assessment process to develop their judgment and receive feed-forward to improve subsequent contributions and learning. Like any skill their ability to critically evaluate, make judgments and assess will only develop with practice, feedback and reflection. However, many academics feel unable to apply frequent and/or innovative forms of assessment due to the overwhelming administrative burden.

In addition, while a key factor in learning is the provision of feedback to students, the capacity of academics to provide quality judgments and feedback is often taken for granted. Without appropriate consensus around the meaning and understanding of academic standards there is no assurance that assessment standards and practices are valid and/or reliable. Furthermore, if academics and tutors don't understand or can't articulate the standards they are assessing, how can they provide students with quality learning-oriented feedback on their work?

My fellowship focussed attention on, and assisted academics to, adopt, design, and implement collaborative learning-oriented activities incorporating the innovative use of self and peer assessment. In addition, a secondary focus was on assisting academics to improve the quality of grading and feedback provided to students and to develop, construct and maintain academic standards.

A combination of workshops, presentations and individual support and resources were provided to disseminate and adapt tested practices to individual contexts. Particular emphasis was placed on the use of SPARKPLUS, a tool that facilitates self and peer assessment and provides feedback on an individual’s contribution to a team project, individual work or enables participants to benchmark their judgment and reasoning against their peers and/or an expert instructor.

Outcomes and Deliverables

Throughout this fellowship I achieved a range of outcomes including:

- Keynotes: 3 (one international, more than 400 attendees)
- Presentations: 19 (approximately 350 attendees)
  - Including 8 open-invitation state-based presentations (150 attendees from 15 different universities and 4 other education providers)
- Workshops: 19 (approximately 300 attendees)
  - Including 8 open-invitation state-based workshops (132 attendees from 15 different universities and 4 other education providers)
- Video resources: 8 (2000+ hits see below)
- Written resources: 1, entitled ‘SPARKPLUS Supporting Resources’ (46 pages)
- SPARKPLUS webpage <spark.uts.edu.au> 23,000+ views since September 2010
- Fellowship resources <sparkplus.com.au/factors> 2000+ views (in the first seven months since posting)
- Reference Groups: 2
Resources
The resources used in the presentations and workshops were made available to all participants electronically. Additional resources including a selection of instructional videos are available online at <sparkplus.com.au/factors>.

Adoption of resources and SPARK\textsuperscript{PLUS}.
Since commencing my Fellowship the use of SPARK\textsuperscript{PLUS} has increased significantly as shown in the table below:

<table>
<thead>
<tr>
<th>Used by staff at</th>
<th>Three years prior to Fellowship</th>
<th>Two years since commencing Fellowship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of subjects</td>
<td>150</td>
<td>900</td>
</tr>
<tr>
<td>Number of students</td>
<td>12,000</td>
<td>&gt; 80,000</td>
</tr>
</tbody>
</table>

In addition, international users have increased to 9 universities (Canada 1, USA 3, South America 1 and European Union 4).

Recommendations
I learnt a great deal during my fellowship. As a result I adopted different dissemination strategies than outlined in my original proposal. In regard to this I make the following observations that may assist future fellows.

1. I found it difficult in a one year fellowship to undertake Australia-wide dissemination activities. Firstly, there are only a number of windows available in the academic year (typically at either end of the academic semester) when academics have the time to engage in developing new activities. Secondly, there is little lead time between finding out you have been awarded a Fellowship and commencing your Fellowship. To address these issues I recommend allowing fellows the option of commencing their fellowships in January. This would allow for a smoother transition of work, sufficient time to find a replacement to teach their courses and to complete other activities that may otherwise place demands on their fellowship time.

2. It is difficult to form new communities of practice to adopt disseminated activities. Often new communities start enthusiastically but fail to continue after the completion of a Fellowship/project. I found dissemination was more successful by identifying an existing community/group of innovative practitioners whose objectives or outcomes could benefit from adopting the practices you want to disseminate. Then, after sharing how your practice could benefit their objectives, provide the necessary training and support to help them achieve their goals.

My final recommendation is in regard to how the impact of grants and fellowships are assessed. I think the true success of dissemination practices can only be determined in retrospect hence I think it is worth considering preserving a small amount of funding to conduct evaluations one year or years after a project/fellowship has been completed.
Table of Contents

Acknowledgements .................................................................................................................. 3
List of acronyms used .............................................................................................................. 4
Executive summary .................................................................................................................. 5
Table of Contents .................................................................................................................. 7
Chapter 1 ................................................................................................................................. 8
  Disseminated Tools ............................................................................................................... 9
  SPARKPLUS .......................................................................................................................... 9
  SPARKPLUS Screenshot Exemplars .................................................................................. 10
  Quizzes using IF-AT Cards .................................................................................................. 15
Chapter 2: Fellowship Resources ............................................................................................ 16
  Framework 1: Collaborative Learning Activity Cycle ....................................................... 17
  Framework 2: Opportunity and Disposition Framework ................................................... 18
  Framework 3: Developing Standards .................................................................................. 20
Chapter 3: Fellowship Activities .............................................................................................. 28
  List of Presentations and Workshops Conducted Throughout this Fellowship ............... 28
Chapter 4: Impact ..................................................................................................................... 31
Chapter 5: My Evaluation ......................................................................................................... 34
Chapter 6: Fellowship Reflections ........................................................................................... 37
Chapter 7: External Evaluation ................................................................................................. 39
  Context ................................................................................................................................. 39
  Overall Impressions ............................................................................................................ 39
  The Fellowship Proposal .................................................................................................... 40
  Fellowship Activities and Deliverables .............................................................................. 41
  Final Report ........................................................................................................................ 42
  Conclusion ........................................................................................................................... 43
References ................................................................................................................................ 44
Appendix A: Selected Research Papers .................................................................................. 47
  Getting tutors on the same page ......................................................................................... 47
  Scratch that itch to learn: a comparative study .............................................................. 56
  Threshold exams to promote learning and assurance of learning .................................. 65
Chapter 1

Despite recent momentum for assessment to change from ‘assessment of learning’ to ‘assessment for learning’ (Torrance 2007) many academics view assessment primarily as a means for students to demonstrate what they have learnt. Conversely, learning-oriented assessment embeds learning in assessment, reconfiguring its design to emphasise the function of learning (Keppell & Carless, 2006; Keppell et al, 2006).

Students should also be involved in the assessment process to develop their judgment (Boud & Falchikov 2007) and receive feed-forward to improve subsequent contributions and learning (Carless, 2007). Like any skill their ability to critically evaluate, make judgments and assessments will only develop with ongoing practice, feedback and reflection.

Feedback, while arguably having the most potential to affect future learning and student achievement, is often not well planned but rather an artefact of the subsequent marking process. Too often it consists of comments that justify the awarded grade rather than assisting students to appreciate strengths, address weaknesses and improve their work. Furthermore, feedback is often provided long after the assessable work has been completed at which time students may be more engaged in their next assessment tasks (Rust et al, 2005).

In addition, many academics feel unable to apply frequent and/or innovative forms of assessment due to the overwhelming administrative burden. As a result, in large classes the main assessment is often a high-stakes assessment, typically an exam. This is a lost opportunity as while instructor engagement and enthusiasm may motivate students to learn, arguably it is our assessment (summative and formative) design and methods that will have the biggest impact on their learning (Ramsden, 2003).

The thoughtful combination of collaborative learning activities, self and peer assessment and educational technology can assist in addressing all of these issues.

While self and peer assessment is frequently associated with providing fairer assessment of group work, it has the capacity to facilitate collaborative learning-oriented assessments, develop students’ judgment and to efficiently facilitate frequent opportunities for students to practise, assess and provide feedback on their own and their peers development (Freeman & McKenzie, 2002, Willey & Freeman 2006a, 2006b, Boud & Falchikov 2007, Willey & Gardner. 2009, 2008a, 2008b, 2010b). Using self and peer assessment also has the advantage of requiring students to reflect and use their judgment in deciding how to respond to the feedback they receive, as feedback from peers represents an opinion to be considered as opposed to expert feedback which has an imperative to action.

Collaborative learning provides opportunities for students to develop interpersonal and critical evaluation skills in addition to professional judgment. The ability to critically evaluate and clearly articulate your point of view is a requisite skill for participation in most professional practice. Despite this, students often receive little training and infrequent opportunities to develop such skills during their academic studies. If educators rely on
transmission primarily telling students what they need to know, they encourage a reliance on memorisation of facts. “For students to make cognitive changes, the learning experience must begin with each student becoming aware of his or her own present understanding” (Hagstrom, 2006, p28).

In my fellowship I promoted tested collaborative learning and self and peer assessment practices that facilitate regular learning-orientated assessments (both formative and summative), even in large classes without undue administrative burden. My aim was to provide real assistance to academics to adopt, design and implement these activities. I promoted assessment tasks that encourage students to take more responsibility for their own learning, learn from their mistakes, explore their learning through peer conversations and to develop their professional attributes including judgment, reflection and critical evaluation.

Training, support and resources were produced and made available to academics Australia wide to develop, implement and evaluate these tasks. These resources were disseminated through a combination of presentations (including three keynotes) and workshops. These presentations and workshops were targeted at specific audiences. They included a series of open state-based presentations, presentations for specific organisations including the Australian Council of Business Deans, the Australia Council of Engineering Deans and NSW/ACT Promoting Excellence Network. As well, I received invitations to present at numerous universities to individual institutions and faculties. To further assist the successful dissemination and development of these activities early adopters from several universities were identified. These early adopters were provided with follow-up support to assist them in designing and evaluating these activities in their own context. In addition, I collaborated with early adopters from three different institutions to produce conference papers and in one case a journal article reporting the results of this collaborative research.

**Disseminated Tools**

During my fellowship I placed particular emphasis on supporting learning-orientated collaborative activities with the software tool SPARKPLUS and Immediate Feedback Assessment Technique (IF-AT) cards. I chose these products because their flexibility enables instructors with even a little imagination to use them in numerous contexts.

**SPARKPLUS**

SPARK (circa 2008) was a self and peer assessment software tool resulting from a joint research project between UTS and USYD. It built on the concepts developed in the original SPARK (Self and Peer Assessment Resource Kit) developed by Associate Professor Mark Freeman. The redevelopment team consisted of Associate Professor Mark Freeman, Dr Keith Willey and Darrall Thompson.

In 2009 Dr Keith Willey assisted by Mike Howard and Anne Gardner further developed the program which is commonly now known as SPARKPLUS. Among many major additions to the tool were a number of new collaborative learning and benchmarking modes and an extension of the program’s capacity to provide learning oriented feedback. SPARKPLUS assists students to make their self and peer assessments by requiring them to rate
Improving learning and developing professional judgment in large classes through collaboration and self and peer assessment

The program has the capacity to produce several assessment factors. Two factors are used in norm-based assessment activities where students evaluate their own and their peers’ contribution to a team task or project. These are:

- a weighting factor used to change a team mark for an assessment task into an individual mark.
- a formative feedback factor allowing students to compare their perceptions to the perceptions of their peers thus providing feedback for learning and development.

We have found that the sharing of these feedback factors between group members not only provides developmental feedback but, combined with students being aware that SPARK PLUS can identify saboteurs (students who deliberately enter dishonest assessments), free-riders (students who do not contribute their fair share to a team project) and over raters (students who over rate both themselves and their peers), promotes honest assessment.

The SPARK PLUS benchmarking modes were designed specifically to facilitate collaborative learning activities. They allow students to compare their assessments and reasoning to that of their peers and or an instructor, highlighting differences to provide feedback, develop judgment and facilitate discussions for learning. The benchmarking modes have a high degree of inbuilt flexibility enabling them to be used effectively for many different types of activities including:

- developing and maintaining academic standards
- improving the quality of multiple-marker grading and feedback
- engaging students with and providing feedback on their understanding of assessment criteria
- designing activities to support students to test and develop their critical evaluation and judgment.

SPARK PLUS provides feedback through a combination of numerical factors, comments and graphic interfaces including radar diagrams, statistical analysis and assessment variations. The capacity to provide this feedback across diverse types of assessment assists students to identify and receive feedback on their individual strengths and weaknesses in different contexts.

SPARK PLUS challenges academics to carefully consider their assessment criteria and, used thoughtfully, has significant potential to improve curriculum and subject design. For example, having to link assessment criteria to attribute development categories can motivate the design of assessment tasks and criteria aligned to learning outcomes.

**SPARK PLUS Screenshort Exemplars**

While SPARK PLUS development was not part of my fellowship, supporting people to use it to improve student learning was. In this report I do not intend to discuss in detail the many different functions, applications and uses of SPARK PLUS. This information can be found in the user guide and other resources available at <spark.uts.edu.au/> and <sparkplus.com.au/factors/>. However, to assist readers of this report who are unfamiliar
with SPARKPLUS I will provide the following screenshots and brief functional explanation of the main modes that were most adopted during my Fellowship.

**Norm-Based Assessment Mode**

This mode is typically used for the self and peer assessment of an individual’s contribution to a team task or project.

![Figure 1: SPARKPLUS norm based assessment results screen](image)

When assessing team contribution it is recommended that a norm-referenced scale like the one shown in figure 1 is used. Here the scale varies from WB (well below the average contribution to the team) to WA (well above the average contribution to the team). SPARKPLUS allows participants to identify and receive categorised feedback on the strengths and weaknesses of their contribution. Feedback is provided on each criterion by the slider triangles (Figure 1). The upper blue triangle shows the self rating submitted by a student for each criterion. The lower orange triangle shows the average rating submitted by a student’s peers.

The SPA or performance factor is used to convert a group mark into an individual mark while the SAPA or feedback factor indicates differences between students’ self perception of their contribution and the average perception of their contribution by their team peers. The radar diagrams provide a graphical summary of a student’s contribution for each category.
Self and Peer Assessment of Individual Work

The process is similar to the assessment of an individual’s contribution to a team task except that a grade scale is used. For example, as shown in figure 2 the assessment scale varies from Z (pass) to HD (high distinction).

Figure 2: A student’s SPARKPLUS results screen for a task where each student had to self assess their own and their peers individual submissions.

Multiple Assessor Mode

The multiple-assessor mode is quite flexible and can be used for numerous applications. During my fellowship I ran a number of workshops and provided training and support to use this mode for:

- tutor benchmarking: to promote understanding of assessment criteria and improve the quality of marking and feedback provided to students.
- development of academic standards: this mode has been adopted by the Achievement Matters: External peer review of accounting learning standards project
- students to self and peer review and provide feedback on each other's work
- committees to assess grants, awards and other types of applications and proposals

The multiple-assessor mode allows participants to rate work and provide written feedback on categories of criteria. After the activity, participants can compare their rating and feedback to those of other participants anonymously if required. Individual ratings are displayed by using colour-coded triangles superimposed on a rating slider (Figure 3). Written feedback from different participants is displayed anonymously, if required, in viewing windows provided for each category of criteria.
The following screen shots were taken from a tutor benchmarking activity published in the proceedings of the Australasian Association of Engineering Education 2011 conference in Perth (Willey and Gardner, 2011a). Here the ratings of the subject coordinator/expert instructor are shown on the top of each slider while those of the subject’s tutors are shown on the bottom.

![Benchmarking results screen in SPARKPLUS](image.png)

Figure 3: Benchmarking results screen in SPARKPLUS: Upper triangle shows coordinator’s marking, lower triangles show individual tutors’ marking of this report.

Participants also receive feedback via the rating and feedback summary screens (Figures 4 and 5 respectively, next page). The rating summary screen shown in Figure 4 provides histograms (that expand when clicked) showing the distribution of ratings across the rating scale frequency bins. An associated slider also shows the minimum, maximum, average and standard deviation of participants’ ratings, making it easy to identify the criteria where there is general agreement and those where participants have quite different opinions. This allows discussion to focus quickly on those areas that need to be addressed. The feedback summary screen allows participants’ comments for each category to be viewed for either all or one particular rating level (Figure 5).
Improving learning and developing professional judgment in large classes through collaboration and self and peer assessment
Quizzes using IF-AT Cards

Immediate feedback assessment technique (IF-AT) cards (Figure 6) developed by Epstein allow students to immediately identify if they have answered multiple-choice questions correctly. These cards require the students to scratch off a covering over the response they think is correct (hence they are often referred to as ‘scratch cards’). If students have selected the correct response a star is revealed (Figure 6). If they selected incorrectly, they consider the remaining options and try again.

IF-AT cards allow students to assess their mastery of material and indicate areas of potential misconception. For example, topics for which questions require multiple attempts before arriving at the correct answer (Q3 in figure 6) identify an area that needs revision.

In my fellowship I used IF-AT cards in the collaborative learning workshops to highlight the characteristics to be considered when designing collaborative learning activities.

I discussed the use of these cards to facilitate many activities including:

- individual summative and formative exams and quizzes
- collaborative summative and formative exams and quizzes
- collaborative discussions and activities
- identification of common misconceptions and threshold concepts
- facilitation of two-stage feed-forward learning oriented examinations
- facilitation of threshold and/or hurdle assessments

In our studies students reported that collaborative activities using IF-AT cards promoted high engagement and that the conversations they had with their group peers helped them identify and subsequently address gaps in their knowledge. This is revealed by the following student comments from a range of IF-AT card activities:

“The heavy use of group/teamwork was very effective in consolidating everyone's knowledge and kept us on the ball.”

“Having an individual quiz followed by a group quiz forced discussion of the questions which was often productive in terms of learning”

“Groups allowed me to ask others when I was stuck both in and out of class. .... I learnt more.”
Chapter 2: Fellowship Resources

Most professionals need to make critical judgments involving decisions that extend beyond their traditional discipline boundaries, requiring them to undertake ongoing learning. Often this learning is informal, learnt on the job from peers. Hence, to prepare students for professional practice educators must provide opportunities for students to experience, practise, reflect and improve their ability to work in collaborative learning environments.

In an educational context, collaboration is generally described as an approach involving joint intellectual efforts between students, or between students and the instructor (Smith & MacGregor, 1992). While few would argue the benefits of collaborative learning, these benefits are not automatic. Kirschner et al (2010) suggest that group learning could be more effective than individual learning if the complexity of the material to be learnt is high. Based on cognitive-load theory the collective working-memory effect predicts sharing the load of processing complex material among group participants enables more effective processing and easier comprehension of the material to be learned. Groups, in effect, form a distributed working memory which had greater capacity than the working memory of an individual. However, one must be careful to avoid the issue of “collective ability” reported by Willey and Gardner (2011b, 2011c) who found that some students, as part of a collaborative team, appeared to understand the subject learning outcomes, however, individually, (without the support of their peers) had gaps in their understanding.

Thus planning a successful collaborative learning activity is a nontrivial task. Thoughtful design including scaffolding to motivate desired approaches and behavior is required. In my fellowship I disseminated the results of several studies I conducted with Anne Gardner investigating the components of successful collaborative learning activities (Willey and Gardner 2012). These studies include investigating both formative and summative assessment activities in which the collaboration was specifically designed to provide opportunities to learn.

In these studies, we found that:

- when activities are summative, students, with some justification, tend to strategically focus on how to achieve the best mark.

- formative collaborative activities provide a low-risk environment (Irons, 2008) allowing students to push their learning boundaries, make mistakes, identify gaps in their learning and have these addressed by their peers and, if necessary, the teaching academic.

- to assist students to make the most of the opportunities provided by formative-learning activities their design should include scaffolding to promote a learning-focus as opposed to a task-focused disposition in students. For example, we constantly remind students that “mistakes compress learning” and to benefit most from formative activities they should push their learning boundaries until they make mistakes and/or discover what they do not know (Svinicki, 2004). Furthermore, to be valued by students the formative activities should be one of the best opportunities for students to assess and receive
feedback on their learning in preparation for their eventual summative assessment.

- collaborative activities should include several cycles with increasing complexity
- collaborative activities should include variation for learning to check understanding.
- collaborative activities should be followed by an individual confirmation task allowing students to confirm their learning to address the issue of “collective ability” (Willey and Gardner, 2011b, 2011c).
- when activities are open-ended (have more than one answer), context dependent and have no specific endpoint, for example, a design problem, students’ learning is difficult unless it is supported by a discourse and feedback that allows students to evaluate the extent of their understanding.

The results of these studies informed the development of the following two frameworks to assist in the design of effective collaborative learning activities.

**Framework 1: Collaborative Learning Activity Cycle**

This framework (Willey and Gardner 2012) was designed to assist in the development of collaborative learning opportunities. The first step in every collaborative learning cycle should be an individual activity that allows participants to identify gaps in their learning/understanding. Often these individual activities are undertaken out of class. Fink reports that: “The key to getting students to do the necessary work and reading before class seems to lie in devising the right kind of in-class activities. Students need to know that the reading done beforehand will be absolutely necessary to do the in-class work and that the in-class work is an important and valuable kind of work” (Bowden and Marton, 1998). The individual activity is followed by a collaborative task where participants have their learning gaps addressed and extend their understanding through conversations with their peers within the collaborative activity. The dialogue within this collaborative activity not only provides the social dimensions important to learning but also provides a discourse to challenge participants’ understanding and judgment, convert tacit understandings to explicit explanations and socially construct meaning, language and standards. Groups are then brought back together, and an instructor clarifies any outstanding issues.

To discern a difference, learn and develop judgment, one must have experienced a variation from previous experience (Irons, 2008). Hence, to verify understanding we recommend that instructors next vary an aspect of the activity to change the outcome and have participants complete this first individually then collaboratively to check their understanding. Finally, students should individually undertake a confirmation task that applies the learning in a new context and/or a more complex situation to confirm understanding after which the cycle is repeated.
We also recommend that a review be undertaken after all activities including final exams. In final exams students are often able to identify gaps in their learning/understanding when undertaking their exam, for example, not being able to answer a question. Without a review these gaps typically go unaddressed with students who pass the exam moving onto next semester's courses. We hold reviews for all our subjects a week before the start of the following semester. (Typically international students who may have returned home during the semester break are back in the country by then.) While attendance at these reviews is voluntary, we have found by listing them in the subject outline a large proportion of students take the opportunity to attend. We make it clear that these reviews are not an opportunity for a re-mark but an opportunity to review their paper to address/clarify any gaps in their learning.

Framework 2: Opportunity and Disposition Framework

In this framework (Willey and Gardner 2012) we suggest that learning is maximised when an assessment activity provides a well-designed learning opportunity and participants (students) approach the activity with a learning focus. Carless (2007) describes learning-oriented assessment as assessment designed to meet both certification and learning purposes. He characterises learning-oriented assessment as having three major components:

- Assessment tasks as learning tasks
- Students involvement in the assessment process (self and peer assessment) and
- Feedback as feed forward

In the case of formative activities the certification purpose could be considered an evaluation purpose. That is, certification implies meeting the requirements of the third party whereas evaluation can solely be for the purpose of the students evaluating their learning and identifying any gaps and/or areas that need to be addressed. Through discussions with academics about their assessment tasks we identified a common tendency to include compliance measures to encourage students to engage with these tasks. For example, academic comments included: “if I didn't give the students marks then no one would complete the exercise” and “if I didn't have a quiz each week students wouldn't come to the lecture.”
As discussed by Sadler (2012) we regularly found instances of instructors awarding marks for activities that were not related to learning outcome achievement. Even when they were, these marks were often below the level required to demonstrate satisfactory achievement for the subject. Examples include:

- attendance (not participation) at tutorials
- revision of prerequisite material
- frequency of contributions to a discussion forum (without regard to the quality of these contributions)
- quizzes containing simple questions that assessed material at a level lower than required to satisfactorily meet the specified learning outcomes
- participation in an activity, for example self and peer assessment or peer-review (without regard to the quality of their contributions)

In discussions with students we found evidence that compliance measures may result in over-assessing or at least over stressing conscientious students who strive to obtain the maximum grade in all activities. Some described their semester learning experience as going from “assessment to assessment”, “revising material I already know to prepare for assessments” (summative) and being so focussed on so many summative assessments that there is “no time to really learn” by engaging deeply with the material. While poorer students conceded such measures did force them to at least participate more with such activities, we suggest that often this engagement may only be at a surface level.

We would argue that instructors should focus on developing a good learning opportunity and then design scaffolding aimed at moving students towards approaching the activity with a learning focus. Scaffolding measures could be described as being persuasive rather than punitive in that there is no summative penalty for non-compliance. Subsequently we
recommend that for all assessment activities (both summative and formative) academics should explain to students (Willey and Gardner 2012):

- why they designed the assessment activity the way they did.
- what learning opportunities the activity provides the students
- how students can evaluate their learning from the activity
- how the activity is going to impact on their reality (enable them to see the world differently)

Framework 3: Developing Standards

While not considered in my initial fellowship proposal, it became apparent as I worked with academics throughout Australia that there was a need for a means to benchmark academic standards. My colleague Anne Gardner and I had been using SPARK PLUS to run benchmarking exercises aimed at improving the consistency of grading and the quality of feedback when assessments were marked by multiple markers, for example, multiple tutors in large classes.

Early in my fellowship Associate Professor Mark Freeman from The University of Sydney introduced me to Professor Phil Hancock from The University of Western Australia. Phil Hancock’s team was about to commence its Achievement Matters: External peer review of accounting learning standards project partially funded by the ALTC. This project built on the earlier work done by Mark Freeman in his role as the ALTC discipline scholar. I saw this as a unique opportunity to disseminate both the activities and tools we had developed and was grateful for Mark and Phil’s willingness to collaborate.

I subsequently, through a series of presentations and workshops, introduced their team to the SPARK PLUS multiple assessor mode and the methods we had used for developing and benchmarking academic standards (Willey & Gardner 2011a, 2013). I am pleased to say that they have not only used the tool successfully but have built on our approach to develop their own processes and techniques suited to their purposes.

I was subsequently invited to hold a number of workshops and make presentations on developing academic standards and improving the quality of grading and feedback provided by instructors. While each one of these activities was tailored for the specific audiences I will now outline the basic concepts in this report.

Background: (from “Getting tutors on the same page”, Willey & Gardner (2011a)
Australasian Association for Engineering Education (AAEE) annual conference, Perth 2011)

As a result of changes in the last two decades, Australian and UK universities have seen a reduction in staff-student ratios often resulting in large classes. Furthermore, research funds are often used to buy permanent academic staff out of teaching, resulting in an increasing number of less-experienced casual or sessional teaching staff being used to conduct core teaching activities such as tutorials and marking of student work (Price 2005, White, 2006).
Improving learning and developing professional judgment in large classes through collaboration and self and peer assessment

Grading is often an activity that results in anxiety for both teachers and students. This is especially so for less-experienced staff when holistic marking is used, in part due to the difficulties in justifying grading decisions to students. This issue is further complicated in large classes by the fact that often a number of staff are used to mark the same activity for different students. Even experienced staff differ in their understanding of academic standards. The fact that increasing grading is being undertaken by less-experienced sessional teachers and tutors only compounds this problem.

For consistent marking amongst tutors, it is important for all assessors to share a common view of the value of a given grade. Tomkinson & Freeman (2007) suggest that some form of induction, for example, a small number of ‘yardstick’ assessments be used as a basis for discussion about standards.

Several researchers, including Price, report that: “An assessment standards discourse is needed to support the functioning of assessment communities of practice…” (Price, 2005, p. 226). That is, tutors develop their understanding of the assessment criteria and language of feedback by discussing marking with other academics.

**Framework:** After conducting a number of studies associated with benchmarking, grading and feedback activities we developed the following framework (Willey & Gardner 2013). This framework, explained below, outlines an efficient and effective process to:

- improve tutors’ understanding of assessment criteria including the factors to consider and their relative importance when assessing against the criteria.
- benchmark their judgment and reasoning against other tutors and instructors.
- assist tutors to convert their tacit judgments into explicit explanations improving feedback to students.
- develop a shared descriptive language improve feedback comprehension.

While we developed this process as a result of studies into grade and feedback variations amongst multiple tutors in large classes, it can easily be applied to the development of virtually any academic standard.

---

**Figure 9: Developing Standards Framework**
Prior to Assessment meeting
Stage 1 (individual assessment)
Assessors/tutors are provided with a copy of two pieces of work to mark against specified criteria, entering their assessment (marks and feedback comments) into the multiple assessor tool in SPARKPLUS. The subject co-ordinator also entered their assessment (marks and comments explaining their reasoning) into SPARKPLUS.

During Assessor/Tutor Meeting
Stage 1 (collaborative discussion and assessment)
Tutors either log on to SPARKPLUS or are provided with a printout to compare their marking and comments to that of the other tutors (displayed anonymously) and reflect on any differences.

Tutors are formed into small groups and asked to discuss their individual marking (previously recorded in SPARKPLUS) and subsequently to collaboratively re-mark and provide feedback comments on each report. That is, they are required to reach a consensus about the appropriate grade for each assessment criterion, agree on an overall holistic grade for the submission and provide comments to explain their reasoning (if you only have time to review one report we suggest that you choose the one with the most variation in ratings and feedback comments).

Stage 1 (facilitator led discussion)
The subject co-ordinator explains their marking of the reports and facilitates a discussion in which the reasons for any grading or opinion (reasons given for awarding grade) differences are explored.

Stage 1 (Vary activity):
To discern a difference, one must have experienced a variation from their previous experience (Runesson, 1999). Hence variation is needed for discernment, which is an important step in learning (Bowden and Marton, 1998). To assist participants to clarify and reflect on their judgment, facilitators should vary an aspect of the activity to change the outcome. For example, you might choose to modify the assessment criteria or task objective.

Let’s assume that the original task was to produce a five-minute video to teach a professional audience CPR, and the assessment criterion evaluates the video’s capacity to engage the target audience. The introduced variation could be to change the target audience to teenagers. Hence the video is now assessed for its capacity to engage a teenage audience. This variation now requires the assessor to apply their judgment, drawing on their experience and understanding of what would engage a teenage audience. By simply changing the target audience the requirements to meet the assessment criteria have changed providing a new context to confirm understanding and or identify and differences and uncertainties.
The cycle is now repeated with tutors reassessing the work considering the variation first individually then collaboratively to broaden their understanding.

Stage 1 Repeat
If time permits and or there were numerous differences in the first exercise to discuss, we recommend that the Stage 1 cycle be completed again using a new piece of work.

Stage 2
Finally, a confirmation task is undertaken were an additional piece of work is assessed to confirm tutors’ understanding and capacity to articulate their reasons for their assessments. Again this confirmation task is undertaken first individually, then collaboratively, followed by a discussion exploring any outstanding differences in grading and/or reasoning.

Discussion
At first glance this may appear to be a long drawn-out process, but before rushing to judgment we suggest that you consider the following:

Even when descriptive marking rubrics are provided there is often a difference in either the assessment and/or combination of components tutors use to arrive at a particular grade decision. Variations may result from different interpretations or ambiguities in the assessment criteria and/or rubric. However, even in the absence of ambiguities, we found because of their diverse experiences tutors often used different combinations and weighting of components in their grading deliberations. That is, tutors didn’t necessarily consider the same components when assessing against any particular criterion. Furthermore, even when they did, their feedback and reasoning for awarding even the same grade often referred to different aspects of the work.

By way of example let’s assume that a group of tutors is assessing student submissions against the criterion: “The students demonstrate that they can justify and communicate advice and ideas to their professional peers.”

A descriptive rubric is provided to the tutors that describes different grade choices in terms the following components:
1. Spelling and grammar
2. Suitability of chosen communication style for specified audience
3. Clarity of argument and reasoning

Let us further assume that post-grading interviews with three tutors revealed that they had different opinions as to the importance (sensitivity or impact) of each of these three components in awarding their grade. Furthermore, these discussions revealed that each tutor also considered one or more of three additional components (listed below) that their experience and judgment told them was important to consider in making their grade choices:
4. Quality of written expression
5. Critical analysis and evaluation
6. Level of jargon used
Improving learning and developing professional judgment in large classes through collaboration and self and peer assessment

Figure 10: Differences in the components considered by tutors in assessing a piece of work

These assumptions are described in figure 10. The figure shows that all tutors considered the three components (1, 2 and 3) described in the rubric. In addition, tutor 1 who considered a total of five components placed an emphasis on quality of written expression (component 4) and critical analysis and evaluation (component 5), while tutor 2 who only considered four components considered the level of jargon used (component 6) when arriving at their grade decision.

These differences, recorded in SPARKPLUS, are revealed during the collaborative discussion and assessment stage. It is quite common in this stage for a tutor to introduce a component that, after discussion, all agree is important but, prior to the collaborative discussion, not all considered. Hence one of the aims of the collaborative discussion and assessment stage is to get tutors to consider the same components in making their assessments.

This however is not the end of the process. Even after tutors have had a conversation and collaboratively decide that a piece of work deserves a certain grade, the assessment weighting of the individual components that led to this decision are probably different.

As depicted in Figure 11 while each tutor now assesses the same components and generally agrees on the awarded grade (represented by the height of each column) their opinions on the relative rating of each individual component (represented by the height of each coloured component) still vary. These variations often result from tutors having different sensitivities to the importance of aspects that contribute to different components.

Figure 11: Differences in tutors grading of individual components
These differences become evident when assessing a new piece of work with different strengths and weaknesses. For example, let us assume that the piece of work assessed to produce Figure 11 contained good ‘spelling and grammar’ represented by the bottom (blue) component. As shown in the figure, the tutors agreed in their assessments of this component (components are the same height). However, the second piece of work assessed contains some occasional but obvious spelling and grammar errors. The tutors now have different opinions (Figure 12) as to how much these errors affect the grading of this component. The figure shows that tutor 1 considers these errors have a significant detrimental impact and grades the lowest on this criterion, while tutor 2, who grades the highest on this criterion, is less worried about these errors.

Similarly, when these differences in individual weightings and sensitivities are extrapolated across all components the overall grade awarded by individual tutors varies (different column heights in Figure 12). Hence in developing a standard there are two main things to be considered:

- the components considered when assessing against criterion (what is being assessed)
- the rating of these components across the possible grade range (the level of quality to be met for different grades).

In our experience most tutors have a higher tendency to agree when the work to be graded is not satisfactory (Fail) or distinctive (D or HD), so to effectively run benchmarking activities we recommend choosing pieces of work that are typically in the high satisfactory or credit range.

When I started as an academic I remember agonising over my grading decisions for student work particularly in my large classes. My judgment improved in subsequent semesters where I was confident in my grading, and I rarely needed to review my decisions. I attribute this improvement to the fact that as I graded I was practising this activity and receiving feedback for improvement from my students when I returned their work.

This analogy provides insight as to why the tutor benchmarking activities have proved to be so effective. They provide an opportunity for tutors to practice and receive feedback. The conversations tutors have in the benchmarking activities provide feedback on their
judgment, challenge their judgment through alternate opinions and provide an opportunity to ask questions and debate their views all in a short period of time. Hence these activities compress the learning cycle. These same benefits can be made available to students by well-designed collaborative learning activities.

Furthermore, using SPARKPLUS to record assessments and feedback promotes inclusiveness of less experienced and less confident tutors by allowing all points of view to be considered. The anonymous reporting through the software means each comment is considered on its merits as there is no way of knowing if it was made by a new or an experienced marker. Furthermore, having differences so clearly reported by SPARKPLUS discourages group think or being distracted by and subsequently focusing discussions on only one or two issues.

**Informed judgment not measurement**

While it is common to see a reduction in grade variations between tutors that have participated in a benchmarking exercise, it would be naive to believe that additional benchmark activities would eliminate these variations completely. In my opinion this is not a problem. I believe the focus of benchmarking activities should be on informing judgment and improving feedback from assessors rather than a primary aim of reducing variation, just as the focus of assessment should be on learning and opportunities to demonstrate this learning not measurement.

Personally I think that grades should be awarded, not marks. In the case of the marking scheme used at UTS this would make a 64/100 a P+ (pass plus) and a 65 a C- (credit minus). The 64 is not one mark short of the 65. The 64 reflects work that has the characteristic of a passing grade at a high level. The C- has the characteristics of a credit grade. We found an added benefit of this approach is that students review their work more holistically rather than focusing on identifying an opportunity for an extra mark to improve their grade.

Finally, standards change or emerge over time. While benchmarking participants have agreed on a standard, once they return to their normal work each individual is subjected to different experiences and opinions that potentially influence the way they interpret that standard. By this we mean that their rating of components might vary or they may even introduce additional new components. Ubiquitous changes in society can also change our expectations of submitted work. Prior to desktop computers most assignments were handwritten, graphs hand drawn and text was regularly corrected using corrector fluid. Such an assignment would be unacceptable in most institutions today. Our standards and expectations have changed, in this case because of technology. Given that standards are socially constructed to maintain agreement on a standard we suggest that regular, at least an annual, benchmarking activity should be conducted.

**Language Activity**

Another activity we like to include in these sessions is to benchmark the language tutors used to provide feedback to student.

In a previous study (Willey K. & Gardner A. (2010a & b)) we found that even when there were only minor differences between the grades awarded by different tutors for the same pieces of work, there were often substantial differences in the language used by tutors.
when providing feedback. We subsequently found that these variations in feedback were a significant contributor to students’ perception that grading was unfair.

For example, let’s consider the case of two students in tutorials taught by different tutors. On one student’s report their tutor has written ‘good report’ and given a mark of 80% Distinction. On the other student’s report their tutor has written ‘good work’ and given a mark of 65% Credit. The two students later compare their grades and all they see is that both reports were judged as being ‘good’ but one received a higher grade. The students’ perception is that the marking is unfair and that one tutor marks ‘easier’ than the other. In reality when each tutor was given the other report to grade in a double-blind marking exercise they awarded marks within the same grade range.

In conjunction with our benchmarking exercises we also run an activity where the facilitator takes a sample of the feedback comments made by the tutors, tabulates them and asks each tutor independently to indicate what grade they would associate with each feedback comment. The results of one such activity is shown in Table 1. The variation in the grades tutors associated with the feedback comments was far greater than the variation in the grades they awarded for the associated work. For example, for the second comment – “Overall this is a fair report leveraging what we expect students to do” – tutors associated this comment with a rating of Satisfactory (S) to Very Good (VG) representing a maximum difference from the mean tutor rating of 25%.

<table>
<thead>
<tr>
<th>Overall</th>
<th>Max Difference from Mean Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good report let down by weak requirements. Students have demonstrated understanding of the main components</td>
<td>S to G 14%</td>
</tr>
<tr>
<td>Overall this is a fair report leveraging what we expect students to do</td>
<td>S to VG 25%</td>
</tr>
<tr>
<td>Good report and good understanding of customer needs and concept of operation. Became too specific and requirements started to specify the solution</td>
<td>S to G 15%</td>
</tr>
</tbody>
</table>

Table 1: Excerpt from feedback language activity

Tutors are asked to reflect on how they would feel as a student receiving an apparent mismatch between their grade and feedback. The exercise concludes with tutors discussing the feedback they would provide against each criterion for the two reports they marked in the benchmarking exercise. This discussion assists tutors to unpack their tacit understandings so they can explicitly articulate their judgment when providing feedback to students. Again our intention here is not for tutors to use exactly the same language but rather for them to appreciate that the language they use in feedback also implies a standard. The aim being to assist tutors to provide developmental feedback for students to feed-forward to their next similar assessment task rather than writing comments to justify the grade they have awarded.
Chapter 3: Fellowship Activities

List of Presentations and Workshops Conducted Throughout this Fellowship

Keynotes

**Collaborative Learning, Assessment and Feedback**

1. Australasian Association of Engineering Education Annual Conference, December 2010
2. Collaborative Learning Conversations Peer Assisted Educational Programs (PEAP) Symposium, Monash University, 7 June 2011
3. NZ Engineering Education Leaders Forum, Manukau Institute of Technology, 29 August 2011

Presentations

**Collaborative Learning, Assessment and Feedback**

1. Western Australia presented at the University of Western Australia, 30 November 2011
2. Australian Council of Engineering Deans (ACED), Fremantle, Western Australia, December 2011
3. South Australia presented at the University of South Australia, 15 February 2012
4. New South Wales presented at the University of Technology, Sydney, 21 February 2012
5. College of Fine Arts (COFA), University of New South Wales, 22 March 2012
6. Victoria presented held at Royal Melbourne Institute of Technology, 29 November 2012

**Assessment: Design, Methods, Validity**

1. ALTC Fellows Forum Australian Catholic University (ACU), 24 April 2012

**SPARK**<sup>PLUS</sup>: **Only limited by your imagination**

1. Australian Council of Engineering Deans (ACED), Sydney, December 2010
2. Australian Council of Business Deans (ACBD), Sydney, February 2011
4. Western Australia presented at the University of Western Australia, 1 December 2011
5. South Australia presented at the University of South Australia, 16 February 2012
6. New South Wales presented at the University of Technology, Sydney, 22 February 2012
7. Victoria presented held at RMIT, 30 November 2012
**SPARK**<sup>PLUS</sup> **Power To Learn. Benchmarking with SPARK**<sup>PLUS</sup>


**Other Presentations**

2. Interpreting SPARK<sup>PLUS</sup> results, 2011 UTS Teaching and Learning Forum, 21 June 2011
3. Inclusive Assessment Academic Liaison Officer Network, University of Technology Sydney, 22 October 2010
4. Using SPARK<sup>PLUS</sup> to support collaborative learning activities, NSW/ACT Promoting Excellence Network Event, The University of New South Wales, 16 May 2012

**Workshops**

**Collaborative Learning, Assessment and Feedback**

3. UTS Faculty of Engineering Information Technology, 16 November 2011
4. UTS Faculty of Engineering Information Technology, 28 November 2011
5. Western Australia Workshop held at The University of Western Australia, 30 November 2011
6. South Australia Workshop, University of South Australia, 15 February 2012
7. New South Wales Workshop, University of Technology, Sydney, 22 February 2012
8. Victorian Workshop, RMIT University, 29 November 2012

**SPARK**<sup>PLUS</sup> **Power to Learn**

1. UTS: Institute for Interactive Media and Learning, 2011
2. Western Australia Workshop, The University of Western Australia, 1 December 2011
3. South Australia Workshop, University of South Australia, 16 February 2012
4. New South Wales Workshop, University of Technology, Sydney, 22 February 2012
5. Victorian Workshop, RMIT University, 30 November 2012

**Benchmarking Judgment and Standards**

1. UTS: Institute for Interactive Media and Learning, 2011
2. ALTC Fellows Forum, Australian Catholic University, 23 April 2012
3. NSW/ACT Promoting Excellence Network Event, The University of New South Wales, 16 May 2012
4. ALTC Assessment Event, University of Technology, Sydney, November 2010
5. Benchmarking to Facilitate Collaborative Conversations to Inform Standards, University of Technology, Sydney, 26 April 2012
6. First Year Experience (FYE) Workshop, University of Technology, Sydney, 24 February 2012
Reference Groups

1. ‘Achievement Matters: External peer review of accounting learning standards’, funded by Australian Business Deans Council (ABDC), CPA Aust, the Institute of Chartered Accountants in Australia (ICAA) and Australian Learning and Teaching Council (ALTC).

2. ‘Hunters & Gatherers: Strategies for Curriculum Mapping and Data Collection for Assurance of Learning’ (Australian Learning and Teaching Council (ALTC) Project)
Chapter 4: Impact

In my opinion it is somewhat difficult to determine the real impact of any fellowship. I believe this can only truly be done well after your fellowship has finished, and you are able to look for evidence of outcomes that have been sustained over a period of time. Having said that, I do believe that I have provided the foundation on which sustainable outcomes can be built. I’ve been particularly pleased to have observed instances of what I refer to as secondary dissemination. This is where during your fellowship you provided support, training and resources to an individual to develop certain teaching and learning activities, and then, this individual has used, modified and/or developed these resources to provide support and training to a new cohort of potential early adopters.

In the last year I’ve had the pleasure of sitting in two teaching and learning forums where I heard somebody presenting on their collaborative learning and/or self and peer assessment activities and acknowledging the support of someone I had supported during my fellowship. That is, the early adopters I had supported during my fellowship had subsequently passed these activities onto a new cohort who were now disseminating them to a different cohort through their presentations. I regard this as arguably being the best indicator of the impact of my Fellowship.

However, given the difficulty in evaluating such outcomes I now provide a brief discussion on the impacts that I was able to observe during my fellowship.

In all more than 1000 people attended one of the 41 presentations and/or workshops that I conducted during my fellowship. Another indicator of the impact of my fellowship was that I only initiated 18 of these activities. The remaining 23 were all invited presentations where I was requested to speak on some aspect of my fellowship activities. These invitations range from keynotes at international conferences and national forums to speaking to special interest groups within a faculty at a university.

A summary of these activities is as follows:

Keynotes: 3 (1 international, in excess of 400 attendees)

Presentations: 19 (approximately 350 attendees)
  Including 8 open-invitation state-based presentations (150 attendees from 15 different universities and 4 other education providers)

Workshops: 19 (approximately 300 attendees)
  Including 8 open-invitation state-based workshops (132 attendees from 15 different universities and 4 other education providers)

During my fellowship it became apparent that to achieve sustained impact provision needed to be made for ongoing support. I approached this in two ways.
Firstly I identified early adopters at individual institutions. I then provided additional support to these early adopters including assisting them in planning and developing their individual assessment tasks and activities. Furthermore, these early adopters became the key people for further dissemination within their own institution. That is, any questions from their university about the resources I provided or the activities I promoted were directed to them and then passed on to me. After discussing these issues with the early adopter they then passed this information back to their university.

This allows the development of high-level local expertise within an institution, enabling these early adopters to provide ongoing support for their institutions after my fellowship was completed. In fact this strategy was so successful that three of these early adopters have subsequently indicated to me that they are now looking to their university to take over this support role as they have been overwhelmed by the requests for assistance.

Secondly I decided that a series of short video presentations posted on the Web would be more successful in supporting ongoing dissemination than printed resources and follow-up workshops. I subsequently streamlined my written resources to only contain material that may be needed in electronic form. For example, information to be included in course outlines. The rest of the information I put in the series of video presentations.

A summary of the access to these resources is as follows:

Webpage Hits (individual IP addresses only, counted once per day)
Video resources (8 videos) plus written resources: Titled “SPARKPLUS Supporting Resources” available at <sparkplus.com.au/factors>. More than 1500 views in the first four months, increasing to over 2000 views, seven months after posting. The Figure below shows that while substantially these resources were downloaded by users in Australia, they also have been viewed internationally in 21 countries (the six most frequent viewing countries are indicated in the Figure 13).

![Figure 13: Webpage hits (1500) and major country totals to <spark.uts.edu.au/factors> in the first four months of posting.](image)

Further evidence of the impact of my Fellowship activities included the invitations I received to speak at and/or join forums and interest groups. While most of these activities consisted
of one-off presentations, I accepted invitations to join two reference groups on other ALTC projects including the:

- Achievement Matters: External peer review of accounting learning standards funded by Australian Business Deans Council (ABDC), CPA Aust, the Institute of Chartered Accountants in Australia (ICAA) and Australian Learning and Teaching Council.
- Hunters & Gatherers: Strategies for Curriculum Mapping and Data Collection for Assurance of Learning (Australian Learning and Teaching Council Project)

While it is not possible to determine the impact and more specifically the uptake of activities/resources not related to SPARKPLUS disseminated throughout my fellowship, anecdotal evidence and follow-up e-mail enquiries suggests that it has been significant and continues to grow. These activities/resources include the design of collaborative learning activities, the innovative use of IF-AT cards, the adoption of collaborative design frameworks and scaffolding and the design of learning-orientated assessments.

In contrast the usage statistics of SPARKPLUS provide a reliable indication of its increased use to support learning activities.

In the two and a half years prior to commencing my fellowship SPARKPLUS was used as a tool to facilitate learning by academics at 11 Australian universities and Mining Education Australia across approximately 150 classes involving 12,000 students. In the two and a half years since commencing my fellowship SPARKPLUS is being used at 23 Australian universities and Mining Education Australia across approximately 900 classes involving 80,000+ students. Furthermore, international users have increased to 9 universities (Canada 1, USA 3, South America 1 and European Union 4).

In addition, while not developed as part of my fellowship, the SPARKPLUS webpage received more than 21,000 hits during my fellowship from 115 countries (see map and major country usage statistics below).

![Map showing webpage hits](image-url)

**Figure 14:** Webpage hits (21,000) and major country totals to <spark.uts.edu.au> during Fellowship.
Chapter 5: My Evaluation

Unfortunately, I was usually so engaged in delivering the workshops and presentations in my Fellowship I often forgot to hand out the evaluation forms. The following four tables report the results of evaluation conducted for the open-invitation Western Australia, South Australia and New South Wales workshops. The average response rate was approximately 60 per cent.

While one respondent reported the Self and Peer Assessment using SPARKPLUS presentation to be a waste of time, the vast majority of respondents reported that the activities were either useful or very useful and reported they had obtained a number of ideas and/or strategies that they intended to use in their teaching.

I realise that evaluation is important, but I found that it was not uncommon for participants to be less than enthusiastic about filling out an evaluation at the end of a workshop or presentation. Most participants were more interested in asking questions or discussing how the presented activities would impact on or could be integrated with their teaching.

To address this in some workshops I tried using online evaluations to be completed on the days following these sessions. The feedback from these responses was probably more beneficial and insightful than that which I obtained by handing out evaluation forms at the end of the session, however, not unexpectedly the response rate was lower.

I intentionally did not distribute evaluation forms for the many invited activities that I conducted. I took the view that being invited to present and/or hold a workshop by these organisations was sufficient evidence of impact.

Table Collaborative Learning Assessment Feedback Presentations
Location: Perth, Adelaide and Sydney
Total Respondents = 47

<table>
<thead>
<tr>
<th>Collaborative Learning Assessment Feedback Presentations Location: Perth, Adelaide and Sydney Total Participants = 47</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>The presentation gave me a number of ideas / strategies that I could use to improve collaborative learning, assessment and/or student feedback in my subject/s</td>
<td>0%</td>
<td>0%</td>
<td>59%</td>
<td>41%</td>
</tr>
<tr>
<td>I intend to implement some of the ideas / strategies discussed in the presentation</td>
<td>0%</td>
<td>9%</td>
<td>54%</td>
<td>37%</td>
</tr>
<tr>
<td>As a result of this presentation I will consider using / changing the design of collaborative learning activities in my subject/s</td>
<td>2%</td>
<td>11%</td>
<td>59%</td>
<td>28%</td>
</tr>
<tr>
<td>Overall I found the presentation to be</td>
<td>0%</td>
<td>2%</td>
<td>49%</td>
<td>49%</td>
</tr>
</tbody>
</table>
Table - Collaborative Learning Assessment Feedback Workshops
Location: Adelaide and Sydney
Total Respondent = 39

<table>
<thead>
<tr>
<th>Collaborative Learning Assessment Feedback Workshops</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location: ALTC Assessment Forum, Adelaide and Sydney</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I obtained a number of ideas/strategies that I could use to introduce/improve collaborative learning, assessment and/or student feedback in my subject/s</td>
<td>3%</td>
<td>3%</td>
<td>54%</td>
<td>41%</td>
</tr>
<tr>
<td>I feel confident that I could implement ideas/strategies learned in the workshop</td>
<td>3%</td>
<td>3%</td>
<td>69%</td>
<td>26%</td>
</tr>
<tr>
<td>As a result of this workshop I will consider using/changing the design of collaborative learning activities in my subject</td>
<td>0%</td>
<td>8%</td>
<td>50%</td>
<td>42%</td>
</tr>
<tr>
<td>Overall I found the workshop to be</td>
<td>0%</td>
<td>10%</td>
<td>33%</td>
<td>56%</td>
</tr>
</tbody>
</table>

Waste of time | Somewhat useful | Useful | Very useful
--- | --- | --- | --- |
Overall I found the workshop to be | 0% | 10% | 33% | 56%

Self and Peer Assessment using SPARKPLUS Presentations
Location: Adelaide and Sydney
Total Respondents = 26

<table>
<thead>
<tr>
<th>Self and Peer Assessment using SPARKPLUS Presentation</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location: Adelaide and Sydney</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The presentation gave me a number of ideas of how I could use SPARKPLUS to introduce/improve collaborative learning, group work, assessment and/or student feedback in my subject/s</td>
<td>0%</td>
<td>8%</td>
<td>65%</td>
<td>27%</td>
</tr>
<tr>
<td>I understand how SPARKPLUS can be used to provide students with feedback and insights regarding their team contribution and development.</td>
<td>0%</td>
<td>8%</td>
<td>63%</td>
<td>29%</td>
</tr>
<tr>
<td>As a result of this presentation I will consider using /or extending the use of SPARKPLUS in my subject/s</td>
<td>0%</td>
<td>27%</td>
<td>31%</td>
<td>42%</td>
</tr>
<tr>
<td>Overall I found the presentation to be</td>
<td>4%</td>
<td>22%</td>
<td>44%</td>
<td>30%</td>
</tr>
</tbody>
</table>

Waste of time | Somewhat useful | Useful | Very useful
--- | --- | --- | --- |
Overall I found the presentation to be | 4% | 22% | 44% | 30%
Self and Peer Assessment using SPARKPLUS Workshops
Location: Perth, Adelaide and Sydney  Total Participants = 31

<table>
<thead>
<tr>
<th></th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I obtained a number of ideas of how I could use SPARKPLUS to introduce/improve collaborative learning, group work, assessment and/or student feedback in my subject/s.</td>
<td>0%</td>
<td>3%</td>
<td>52%</td>
<td>45%</td>
</tr>
<tr>
<td>I appreciate how the SPARKPLUS factors can be used to provide students with feedback and insights regarding their team contribution and development.</td>
<td>0%</td>
<td>3%</td>
<td>56%</td>
<td>41%</td>
</tr>
<tr>
<td>As a result of this workshop I will consider using/changing the design of my collaborative learning/group work activities in my subject/s.</td>
<td>0%</td>
<td>6%</td>
<td>45%</td>
<td>48%</td>
</tr>
<tr>
<td>As as a result of this workshop I am considering using SPARKPLUS or extending the use of SPARKPLUS in my teaching.</td>
<td>3%</td>
<td>10%</td>
<td>42%</td>
<td>45%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Overall I found the presentation to be</th>
<th>Waste of time</th>
<th>Somewhat useful</th>
<th>Useful</th>
<th>Very useful</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
<td>7%</td>
<td>36%</td>
<td>57%</td>
</tr>
</tbody>
</table>

Improving learning and developing professional judgment in large classes through collaboration and self and peer assessment
Chapter 6: Fellowship Reflections

My Fellowship was a very rewarding experience. It enabled me to disseminate tools, resources and framework/methodologies developed through many years of research. Given that most of my educational research has been published and presented in discipline related engineering forums, conferences and journals, my fellowship enabled me to reach a broader audience across many disciplines.

While the engineering education community is well established and vibrant my fellowship allowed me to extend my national academic network across other disciplines. However, arguably the greatest personal benefit of my fellowship was the opportunities to meet with other fellows and participate in the fellow’s network. This cross disciplinary network has members who are excellent researchers, practitioners, policymakers and administrators, providing a rich source of contacts, discussion and the exchange and challenge of ideas.

Developmental benefits relating to my fellowship included the constructive feedback I received from workshops and presentations I gave. In some instances these presentations included reporting statistics, algorithms and data to an audience less numerate than my engineering colleagues. The feedback received helped me appreciate the need to focus on the most important themes and how best to clearly articulate them. For example in one workshop I was explaining a series of graphs when a member of the audience from the social sciences said they were unable to interpret them. As a result of this feedback I had to think of different ways of presenting mathematical material, for example using images and analogies instead of equations or tables. In addition, I further developed my skills to engage an audience, not waste their time and resist the temptation to cover too much but rather allow time for audience engagement, conversations and interactions.

In retrospect I found embarking upon a fellowship that focused on dissemination to be more demanding and time-consuming than I anticipated. This in part related to the fact that as a result of the feedback I received when conducting presentations and workshops I changed the type, format and content of the resources I had planned to produce. Additionally, I quickly become aware of how difficult it was to fit multiple events into the windows of opportunity within an academic year (typically at either end of the academic semester) when academics have the time to engage in developing new activities.

My fellowship activities mostly related to dissemination beyond both my own university and academic discipline (I had already disseminated most of the resources within my university through various teaching and learning forums and within the engineering community in particular through the Australasia Association for Engineering Education). This together with the fact I bought out my teaching to undertake fellowship activities left me somewhat invisible to and isolated and disconnected from my school, faculty and discipline cohort. On returning to my school there was no opportunity to return to teaching my previous subjects. This meant substantial unexpected work had to be undertaken to develop and teach new subjects. This feeling of solitude was not helped by the fact that I chose not to engage an assistant to help with organisation, administration and/or other fellowship activities, but rather undertook the work myself to enable more to be achieved with the funds provided.
If I was to undertake a fellowship again I would consider sharing it with another colleague and would definitely use some of the funds to engage significant administrative support. Would I apply for a fellowship again? Yes probably, but not straight away. Would I recommend colleagues to apply for a fellowship, definitely!
Chapter 7: External Evaluation

External evaluation report by Professor Robin W King

Context

I accepted the invitation to evaluate this project in late 2010 and watched its progress since, including as a participant in two workshops at RMIT University on 29-20 November 2012. I was also in the audience for Keith Willey’s invited keynote address to the Australasian Association for Engineering Education conference in Sydney in December 2010. This evaluation is based primarily on my observations of those events and relevant documentation, including Keith’s Fellowship Proposal and his draft of the Final Report.

I also declare my interests in the success of the work undertaken in this fellowship, as it addresses a key area of higher education practice that I have been supporting for some time, specifically in the field of engineering. Keith Willey’s work takes forward some actions recommended in the 2007-8 Review of Engineering Education led by Australian Council of Engineering Deans (ACED) and conducted as an ALTC-supported discipline scoping study, and for which I was project manager and report author. In subsequent work as Chair of Engineers Australia’s Accreditation Board, I have been advocating for increasing students’ capacity for self-reflection to improve educational outcomes and other projects, and for using Keith Willey’s work and insights. I have specifically referred to his work in international presentations in China, Europe, Japan, and the Philippines on improving outcomes-based education. I have also nominated and supported Keith and his colleagues for institutional, state and national awards, all of which, from memory, have been successful.

This evaluation report is, however, an assessment of the extent to which the fellowship goals have been met by the planned activities, and their impact on university educators. In the spirit of the Fellowship, I also hope that this assessment is of value to Keith himself.

Overall Impressions

The fellowship has been clearly focussed on its stated intent to improve assessment to improve students’ learning. Its core methodology was to engage teaching academics in doing assessment better. From reading the Executive Summary in the Final Report the impression gained is one of activity (keynotes, presentations, workshops) and product development (video, software and written resources). The increased level of usage of SPARKPLUS over the fellowship period and beyond, and its national and international reach is creditable evidence of value and impact.

This level of interest indicates that the fellowship has been an effective vehicle for Keith to address a sequence of propositions on current shortcomings of assessment practices and academics in higher education that (in my words) are:

- good student assessment should be focussed on learning (not only for grading);
- students can learn many desired skills for employment, including judgment, through active (and collaborative) participation in well-designed assessments,
individual and collaborative;

- academics are not generally good at either design of ‘assessment for learning’, or providing genuinely constructive feedback, especially in large classes;
- academics need support to design effective and administratively efficient collaborative (and assessed) learning tasks, and these can be supported by operational frameworks and support by suitable software tools for self- and peer-assessment.

Recognising each of these propositions may be quite challenging for academics (and students) to face. Indeed some aspects of them may be truly radical for academics and students from ‘didactic’ cultures and in quantitative, content-rich disciplines such as engineering and accounting. Nevertheless, there is plenty of evidence to support the need for work on these propositions in these disciplines, both from within and outside their academic communities.

Progress in these areas can be made by encouraging academics to engage in their own learning journey, as Keith has successfully done during his Fellowship. The following sections of this report elaborate my observations on Keith’s fellowship proposal, the activities conducted and resources developed and the Final Report.

The Fellowship Proposal

The original fellowship proposal was well-argued and clearly linked into other ALTC/OLT projects and activities. The central concept is that ‘assessment for learning’ is critical – perhaps especially in disciplines such as engineering where academics tend to focus on delivering ever more content, rather than understanding, by good assessment for example, what their students are learning.

The fellowship plan was extremely ambitious. It envisaged examination of the latest self- and peer assessment practices, and development of documented guidelines and instructional video resources for academics. No less than six dissemination modes (methods and audiences) were proposed, some involving other ALTC/OLT grant recipients. These activities (mostly state-based presentations linked to participant workshops) would contribute to the formation of development of an active user community of academics engaged with self- and peer assessment (using the SPARKPLUS software tool) that Keith and colleagues could support beyond the duration of the Fellowship. The plan was ambitious in terms of the rate at which the different activities could be scheduled and delivered. Rather than curtail them, the ALTC/OLT allowed a time extension and the ultimate number of presentations and workshops exceeded original plans.
Fellowship Activities and Deliverables

Presentations and Workshops

Keith reported delivering 41 presentations and workshops to audiences totalling more than 1,000 during the course of his Fellowship. Their titles and modes were adapted to suit their audiences and available time, but essentially were on two topics:

- Collaborative Learning, Assessment and Feedback: in summary, an evidence-based exposition of the collaborative learning cycle, assessment in constructive alignment, and giving constructive feedback;
- Using SPARKPLUS with an emphasis on supporting academics to understand how its use can change students’ behaviour, promote beneficial learning outcomes (including judgment), and incentivise student teamwork (as well as detecting ‘freeriders’, ‘saboteurs’, etc.).

For both topics, the workshop mode included an introductory and interactive presentation on the relevant topic. I observed one keynote presentation on collaborative learning and participated in one workshop on each of these topics.

Keith’s style of presentation and facilitation is (literally) highly engaging. He successfully demonstrates how to have a dialogue with large classes by building in mini-collaborative exercises into his presentations. He demonstrates the use of IF-AT (scratch cards) for self or group testing, as well as SPARKPLUS. Keith uses stories from his own industry and academic experiences effectively to demonstrate the principles he is aiming to get across and the good practices that he wishes to promote. In the classroom, his practical and engaging approach to teaching would be valued by students, likely to be the majority, who learn best from examples, rather than ‘theory’.

My only criticism is that Keith can try to cover too much material, tell too many stories, and, in his enthusiasm, can try to illustrate too many features of the software. However, he does provide his academic audiences with adequate background and links to enable them to pursue topics further, as discussed next.

Software Support and Video Resources

Keith has produced a 45-page handbook to support SPARKPLUS users and nine videos. The handbook is clearly written and well-illustrated, and is designed to lead users progressively through the SPARKPLUS software tool. The videos address each major topic, with an exposition of principles supported by PowerPoint slide images, followed by a demonstration of using the particular software tool. The approach in both these sets of resources is logical, painstaking, and user focussed. I am confident that a naive or occasional SPARKPLUS user would feel well supported.
Final Report

Keith’s’ draft Final Report broadly covers the OLT reporting requirements by discussing the background, activities and outcomes of the Fellowship. In the Executive Summary, Keith also proposes three sensible recommendations to assist future fellowship holders that may be useful for OLT to consider further, especially if they are echoed by other fellows’ experiences.

Here, I address the quality of the Final Report as a stand-alone document on the issues proposed and addressed, and as a scholarly contribution. On the negative side, the Report contains no detailed or explicit ‘examination of latest self and peer assessment practices, needs and issues at both national and international universities ...’ as proposed as a fellowship activity. However, to do justice to this goal the fellowship would have needed much more time and research effort than it had it at its disposal. In contrast, the principal Fellowship activities and goals that have been about disseminating state-of-the-art knowledge and techniques about self-and peer assessment for learning have been accomplished very well and are adequately reported.

Chapter 1 adequately outlines both the principles and value of collaborative learning and the use of scratch cards and SPARKPLUS. This is a useful and persuasive introduction to the field.

Chapter 2 develops the theme of collaborative learning, scaffolding it with two reasonably well-established ‘frameworks’ and with developing ideas and practices on assessment standards and benchmarking. Again, Keith’s detailed exposition of this area is very useful, and worth developing into a future journal paper in its own right. This is relevant to the whole area of outcomes assessment and benchmarking becoming increasingly significant in higher education. From the work he has done with accounting academics, Keith would be well placed to lead and contribute to further work in this area, perhaps with members of the (international) engineering community who are seeking to take forward the findings and experiences of participating in the OECD AHELO project.

The remaining material lists the activities, and their impact. Keith’s reflections on impact and its evaluation are insightful, and demonstrate a strategic approach to implementing change in higher education. The data that Keith has provided on the increases in web-site hits and institutional adoption of SPARKPLUS speak for themselves. The evaluations by workshop participants are strongly positive towards making changes in their practice. A follow up evaluation of this would be worth subsequent in depth study. Keith concludes the Final Report with three co-authored conference papers that are normally provided to workshop participants. These contribute to the Final Report, making it a more complete document on the subject, as well as the mechanics and outcomes of the Fellowship.
I would encourage Keith and his colleagues to make more of their work in one or more extended journal paper or even a short monograph. The former would also satisfy academic pressures to publish in peer reviewed journals.

Conclusion

Although Keith had more ambitious goals than could have reasonably been achieved, this fellowship has been effective in dissemination of previous and ongoing work on collaborative learning and supporting the further use of self and peer assessment. Keith has established a national and international reputation for theoretically sound, evidence-based work, and the development of a well-supported software tool that could become an international standard.

The fellowship has probably reached more academics and has had greater reach than might have been envisioned. The content of the Final Report and associated published work demonstrate that the Fellowship goals have been broadly achieved.

Robin W King
Emeritus Professor, University of South Australia
1 November 2013
References


Appendix A: Selected Research Papers

The work reported in the following research papers was not part of my fellowship activities however the content of these papers was presented in the presentations and workshops, and the papers were distributed as part of the dissemination activities.

Getting tutors on the same page

Citation:

Keith Willey
University of Technology, Sydney, Sydney, Australia
Keith.Willey@uts.edu.au

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

Abstract: In large engineering subjects, it is common to have multiple tutors where each tutor is responsible for grading the assessment tasks for students in their tutorial. An issue regularly faced by subject coordinators is how to achieve a consistent standard of marking and feedback quality amongst different tutors. To address this issue the authors initially used a number of methods including double blind marking to support consistent grading. However, with increasing demands on academics these time-consuming activities became an unrealistic option. This process was improved by using a software tool to compare both the marking and feedback provided by different tutors for a number of randomly selected project tasks. In this paper, we report using new software features developed as a result of this previous research to quickly establish and build a community of assessment practice amongst subject tutors. The reported process promotes inclusiveness by using a software tool to anonymously record and report tutor assessments allowing all opinions to be considered during a subsequent discussion activity. Even though this pilot exercise was undertaken by experienced tutors it significantly influenced their feedback skills and to a lesser extent their marking standards.
Introduction

As a result of changes in the last two decades, Australian and UK universities have seen a reduction in staff–student ratios often resulting in large classes. Furthermore, research funds are often used to buy permanent academic staff out of teaching, resulting in an increasing number of less experienced casual or sessional teaching staff being used to conduct core teaching activities such as tutorials and marking of student work (Price 2005, White 2006).

Grading is often an activity that results in anxiety for both teachers and students. This is especially so for less experienced staff when holistic marking is used in part due to the difficulties in justifying grading decisions to students. This issue is further complicated in large classes by the fact that often a number of staff are used to mark the same activity for different students. Even experienced staff differ in their understanding of academic standards. The fact that increasing marking is being undertaken by less experienced sessional teachers and tutors only compounds this problem.

For consistent marking amongst tutors, it is important for all assessors to share a common view of the value of a given grade. Tomkinson & Freeman (2007) suggest that some form of induction, for example, a small number of ‘yardstick’ assessments be used as a basis for discussion about standards.

Several researchers including Price report that “An assessment standards discourse is needed to support the functioning of assessment communities of practice…” (Price 2005, p. 226). That is, tutors develop their understanding of the assessment criteria and language of feedback by discussing marking with other academics. This aligns with a social constructivist view of learning, that is, learning requires “active engagement and participation” (White 2006, p.237) this being true for both tutors and students.

In large engineering subjects, it is common to have many tutors where each tutor is responsible for marking the assessment tasks for students in their tutorial. An issue regularly faced by subject coordinators is how to achieve a consistent standard of marking and feedback quality amongst different tutors.

To address this issue the authors initially used a number of methods including double blind marking and re-marking a random sample of assessments to support consistent grading. However, with increasing demands on academics these time-consuming activities became an unrealistic option. In our first effort to improve the process we used the benchmarking tool in SPARKPLUS to allow tutors to compare both the average mark and feedback provided by tutors for a number of selected project tasks (Willey & Gardner 2010, a & b).

This activity proved effective in reducing the variability in marking amongst different tutors. Furthermore, we found that using a software tool to record tutor assessments and feedback before exploring their understanding in a subsequent discussion activity promoted inclusiveness of less experienced and less confident tutors by allowing all points of view to be considered. As a result of this initial research, we developed a number of new SPARKPLUS features to promote more in depth conversations to improve both the standard and consistency of tutor marking.
In this paper, we report a pilot study conducted using the new SPARKPLUS features to examine the mechanisms by which tutors develop a shared understanding of assessment requirements through these collaborative conversations.

**New Software Features**

The new multiple assessor mode in SPARKPLUS (SPARKPLUS 2011) allows participants to rate work and provide written feedback on categories of criteria. After the activity, participants can compare their rating and feedback to those of other participants that is provided anonymously. Individual ratings are displayed by using colour-coded triangles superimposed on a rating slider (Figure 1).

![Figure 1: Benchmarking results screen in SPARKPLUS: Upper triangle shows coordinator’s marking, lower triangles show individual tutors’ marking of this report.](image)

Written feedback from different participants is displayed anonymously in viewing windows provided for each category of criteria. In the instance shown in Figure 1, the ratings of the subject coordinator are shown on the top of each slider while those of the participating tutors are shown on the bottom.

Participants also receive feedback via the rating and feedback summary screens (Figures 2 & 3 respectively). The rating summary screen shown in Figure 2 provides histograms (which expand when clicked) showing the distribution of ratings across the rating scale frequency bins. An associated slider also shows the minimum, maximum, average and standard deviation of participants’ ratings making it easy to identify the criteria where there is
Improving learning and developing professional judgment in large classes through collaboration and self and peer assessment

While the feedback summary screen allows participants’ comments for each category to be viewed for either all or one particular rating level (Figure 3).

Figure 2: SPARK PLUS Benchmarking rating summary screen (only the first two categories have been shown)

Figure 3: SPARK PLUS feedback summary screen
Method

A second year core (all disciplines) engineering degree subject, at the University of Technology, Sydney, typically has an enrolment of approximately 300+ students per semester. In addition to lectures, students are distributed amongst ten tutorials where individual tutors are responsible for grading assessment tasks. One of the assessment tasks for this subject is the Requirements Specification report which was used in the benchmarking activity described in this paper.

The reported investigation, conducted during Autumn semester 2011, consisted of a number of stages:

Prior to Tutor meeting
Stage 1: Tutors were provided with a copy of two Requirements Specification reports from the current semester. (These functioned as the ‘yardstick’ assessments as suggested by Tomkinson & Freeman (2007)). Tutors marked these reports against specified criteria and entered their assessment (marks and feedback comments) into the multiple assessor tool in SPARKPLUS (Figure 1). The subject co-ordinator also entered his assessment (marks and feedback comments) into SPARKPLUS to allow comparison with the tutors’ assessment.

During Tutor Meeting
Stage 2: Tutors logged on to SPARKPLUS and compared their marking and feedback to that of the other tutors (displayed anonymously) and the subject coordinator.
Stage 3: Tutors were formed into a group and asked to discuss their individual marking (previously recorded in SPARKPLUS) and subsequently to collaboratively re-mark one report i.e. they were required to reach a consensus about the appropriate mark for each assessment criterion and agree on an overall holistic grade for the submission. (For this part of the exercise, we chose the Requirements Specification report with the most variation in ratings and feedback comments)
Stage 4: The subject co-ordinator explained his marking of the reports which was compared to that of the tutors. Subsequent discussions explored the reasons for particular ratings and grading differences.

At various stages in the project (pre, during and post meeting) tutors were asked to complete a series of reflective questionnaires that consisted mainly of open-ended questions. Subsequently, tutors were interviewed to further explore the impact of the reported exercise. The authors also observed the interaction amongst tutors and kept notes during the tutor meeting.

Results and Discussion

In the reported semester there were 10 tutorials taught by six staff (five tutors and the subject coordinator) all of which agreed to participate in the pre-tutor meeting activities. However, only four tutors and the subject coordinator participated in the tutor meeting.

Prior to the tutor meeting, these four tutors were asked to assess their perceived level of expertise and confidence in the subject material, understanding of the assessment criteria
and their capacity to grade and give feedback to students on their reports. The results of these assessments are shown in Table 1.

Table 1: Results of pre-activity survey.

<table>
<thead>
<tr>
<th>Selected Questions From Survey 1</th>
<th>n = 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>My expertise in the subject material covered in these reports is:</td>
<td>Low</td>
</tr>
<tr>
<td>I am confident in my ability to grade these reports to the required standard.</td>
<td></td>
</tr>
<tr>
<td>I am confident that I understand / interpret the assessment criteria.</td>
<td>1</td>
</tr>
<tr>
<td>I am confident that I can clearly articulate and explain the strengths and weaknesses of these reports to students when I provide them with feedback.</td>
<td>2</td>
</tr>
</tbody>
</table>

All the tutors rated themselves as having high subject material expertise and confidence in their ability to mark the reports. This is not surprising as each was experienced, having tutored the subject for at least three semesters. Some of the tutors were less confident with their understanding of the assessment criteria and their ability to provide student feedback.

Each of the two reports was assessed against 20 criteria spread across four different categories. Despite the relative experience of the four tutors, they reported that during the conversations they questioned or changed their understanding of both the subject material and the assessment criteria for at least one criterion and up to three criteria as shown in Table 2.

Table 2: Impact of activity conversations for each tutor.

<table>
<thead>
<tr>
<th>UNDERSTANDING OF SUBJECT MATERIAL</th>
<th>Tutor 1</th>
<th>Tutor 2</th>
<th>Tutor 3</th>
<th>Tutor 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>For how many criteria did the conversations with others cause you to question or change your understanding of the subject material?</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>INTERPRETATION/UNDERSTANDING OF ASSESSMENT CRITERIA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>For how many criteria did the conversations with others cause you to question or change your interpretation/understanding of the assessment criteria?</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

Arguably more important, three tutors (one tutor did not respond to these questions) and the subject coordinator reported that the discussions with their peers would impact/alter the way they mark their reports, the issues on which they would give feedback and the language they will use to provide this feedback (Table 3). Interestingly, the subject coordinator who arguably knew more about the subject than the tutors, also reported that the discourse increased his understanding of the subject material “A Lot”.

Improving learning and developing professional judgment in large classes through collaboration and self and peer assessment
Three of the tutors reported that they learnt equally from explaining their opinions to others as they learnt from listening to the opinions of their peers. The remaining tutor reported his learning usually occurred when listening to other members of the group. Not surprisingly, we observed this tutor to be the least confident in their discussions with the others.

The tutors and subject coordinator also commented that while their ability to mark the reports and provide appropriate feedback to students increased through reading the ratings and feedback provided by others in SPARKPLUS, the most significant improvement resulted from the discussions. However, all participants agreed that without the availability of this data and the clear way in which SPARKPLUS presented it, they felt the discussions would have taken much longer and been less focussed and effective.

When asked to describe the best thing about having a focussed conversation with other tutors about grading and feedback typical responses were:

- “It validated my understanding of the subject and fine tuned a few concepts”
- “I realising that I marked too easily”
- “It was a fast way to get issues discussed and resolve differences”

After the collaborative re-mark, it was interesting to note that both the grade and mark given to the report reduced significantly to a 65% (mark) Credit (grade), from an average 77% (mark) Distinction (grade) (individual tutor marks pre-activity were 81%, 75%, 78% and 74%). On investigating this difference, we found that it was mostly a result of changes in the tutors’ expectation of what was required in the test plan and the report presentation including the standard of grammar, spelling etc. Through their discussion, the tutors raised their expected standard for these two aspects of the report. Furthermore, the tutors reported that the activity would have a significant impact on the topics, language and breadth of the feedback that they would give students. One tutor commented that through the discussions they “learnt strategies/tactics for dealing with student issues”, while another tutor remarked that “talking first-hand and listening to the other tutors thinking” helped him clarify what was required by the students to demonstrate the subject learning outcomes.

The participants agreed with our observations that having all points of view recorded and available anonymously promoted inclusivity as it ensured all differences in opinions were available to be discussed irrespective of the participant’s expertise. A number of the
participants commented that in the absence of this reporting they may not have been confident to report their opinion especially where it differed significantly from the rest of the group. Furthermore, having these differences so clearly reported discouraged group think or being distracted by and subsequently focusing discussions on only one or two issues.

In summary, although only a pilot trial the results support the findings of other researchers (Tomkinson & Freeman 2007, Price 2005, White 2006) that conversations between academics are an effective method of developing a shared understanding of assessment criteria and improve marker consistency. In addition, this study suggests that the addition of the software tool has potential to make the process more inclusive, whilst enabling conversations to focus quickly on the pertinent issues.

Conclusion

Our findings support the conclusions of other researchers who found that conversations with other academics about assessment standards and marking is an effective method of developing a shared understanding of assessment criteria and improving marker consistency. In addition, the results suggest that the software tool has the potential to make the process both more inclusive by providing the opportunity for all to present their opinion irrespective of their perceived expertise or confidence and efficient by clearly identifying differences in opinions, enabling conversations to focus quickly on the pertinent issues.

References


Copyright © 2011 Willey & Gardner: The authors assign to AaeE and educational non-profit institutions a non-exclusive licence to use this document for personal use and in courses of instruction provided that the article is used in full and this copyright statement is reproduced. The authors also grant a non-exclusive licence to AaeE to publish this document in full on the World Wide Web (prime sites and mirrors) on CD-ROM or USB, and in printed form within the AaeE 2011 conference proceedings. Any other usage is prohibited without the express permission of the authors.
Scratch that itch to learn: a comparative study

Citation:

Keith Willey
University of Technology, Sydney, Sydney, Australia
Keith.Willey@uts.edu.au

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

Abstract: Engineers today are required to make critical judgments involving decisions that often extend beyond traditional discipline boundaries. This requires professional engineers to undertake ongoing learning. Much of this learning is informal, learnt on the job from peers from different disciplines. To enable students to develop the skills required for professional practice they need opportunities to experience, practise, reflect and improve their ability to work in a collaborative environment. One method used at the University of Technology, Sydney to develop these skills is collaborative activities incorporating immediate feedback. Subject topics are tested through quizzes that are initially undertaken individually and then collaboratively using immediate feedback assessment technique (IF-AT) cards. These activities allow students to first identify and subsequently have gaps in their learning addressed initially by their peers within the one activity. This paper reports on a comparative evaluation of the collaborative use of IF-AT quizzes in four subjects taught by the authors. We found that these methods not only consistently improved student engagement, learning and developed skills required for lifelong learning, but also promoted changes in their learning culture by having them take more responsibility for their own learning.

Introduction

Engineers are often required to make critical judgments involving decisions that extend beyond traditional discipline boundaries. This requires professional engineers to undertake ongoing learning. Much of this learning is informal, learnt on the job often from peers from different disciplines (Trevelyan 2007). To develop the skills required for professional practice students need opportunities to experience, practise, reflect and improve their ability to work in a collaborative environment.
Furthermore, many students resist having to take responsibility for their own learning rather expecting this to be the responsibility of their teaching academics. In some cases, this resistance is in part due to students’ previous educational experience that combined didactic teaching and passive learning. This type of tuition does not afford students the opportunities to both develop and practice exercising their own judgment in-class.

One method used at the University of Technology, Sydney (UTS) to improve students’ collaborative learning skills and provide opportunities to develop and practice judgment is collaborative activities that incorporate immediate feedback. In this paper, we examine an activity where subject topics are tested through quizzes that are initially undertaken individually and then collaboratively using immediate feedback assessment techniques (IF-AT) cards. These activities provide an immediate feedback cycle allowing students to first identify and then afterward have gaps in their learning addressed initially by their peers and subsequently by their teaching academic.

This paper reports on a comparative evaluation of the collaborative use of IF-AT quizzes in four different engineering subjects to investigate their effectiveness in promoting collaborative peer learning.

Background

Professionals, in addition to being technically competent, require the skills of collaboration, communication and the ability to work in teams (Lang et al 1999, Scott & Yates 2002, Mills & Treagust 2003). However, there are reported competency gaps between the skills required by employers including communication, critical thinking, leadership, teamwork skills and life-long learning capabilities, and those developed by students during their undergraduate courses (Hargreaves 1997, Meier et al 2000, Jones 2003, Bryan et al 2005, Markes 2006, & Chung et al 2008). Workplace learning and professional practice is often collaborative (Littlejohn, Margaryan & Milligan (2009). It follows that students’ preparation for entering this environment should include opportunities to practise collaborative learning with their peers. Collaborative learning also provides opportunities to develop interpersonal and critical evaluation skills in addition to professional judgment. The ability to critically evaluate and clearly articulate your point of view are requisite skills for successful participation in most professional practice. Despite this, students often receive little training and infrequent opportunities to develop such skills during their academic studies.

Collaborative learning is also attractive from the perspective of the social constructivist model of learning (Jawitz and Case, 2009). The social constructivist view is that learning takes place when students construct their knowledge through individual engagement and social interactions with others (Wu, Beiber and Hiltz, 2008, Purzer, 2009). Hagstrom (2006) argues that “…contexts for new knowledge construction include a blending of people … and provides the occasion for the construction of new knowledge … If educators simply tell students what they need to know, they encourage reliance on memorization of facts. For students to make cognitive changes, the learning experience must begin with each student becoming aware of his or her own present understanding” (Hagstrom, 2006, p28).
While projects, assignments and laboratories are regularly considered as opportunities to incorporate collaborative learning activities it is less common to undertake collaborative examinations and quizzes. Stark (2006) describes conducting an individual exam and then giving the same exam to teams to complete. He observed that “... students engage each other in serious discussion of the material to the same end – that of understanding the material better than they did before.” Stark (2006) further reports students’ learning benefits from having to explain concepts to their peers and that “Team exams make post-exam feedback more of a student-directed and student-centred activity”.

Quizzes using IF-AT Cards

Immediate feedback assessment technique (IF-AT) cards (Figure 1) developed by Epstein allow students to immediately identify if they have answered multiple-choice questions correctly. These cards require the students to scratch off a covering over the response they think is correct (hence they are referred to in class as ‘scratch cards’). If they have selected the correct response a star is revealed (Figure 1). If they selected incorrectly, they consider the remaining options, and try again. In controlled trials the IF-AT method was shown to promote both retention of learned material (Epstein et al 2002, Dihoff et al 2004, Brosvic et al 2005, Brosvic & Epstein 2007) and higher levels of independent learning (Brosvic et al 2005; Persky & Pollack 2008).

Persky & Pollack (2008) report that use of IF-AT cards “allows a student to assess his or her own mastery of the material, indicates to the student areas of potential misconception, and allows the student to think about and rework problems. Each of these components potentially increases deep learning.”(p.5).

Method

Four subjects at different stages of a student’s educational journey were chosen to assess the effect of collaborative quizzes on student learning. The subjects chosen were a:

- second year core engineering subject Design Fundamentals taken by all engineering students
- final year Telecommunications Engineering subject Communications Theory
- final year Civil Engineering subject Concrete Design
- postgraduate Telecommunications subject Transmission Systems

Within each subject topics were tested through a series of quizzes (4 to 5). Students completed each quiz individually and then immediately afterwards completed the same quiz...
collaboratively in groups of four to five students using the immediate feedback assessment techniques (IF-AT) cards. When completing the quiz collaboratively if a group selected the correct response at their first attempt they are allocated full marks for that question. If they selected an incorrect response they continued to select another answer until the correct answer was obtained enabling them to receive partial credit (depending on the subject 60% to 33% second attempt, 40% to 16% third attempt, 0% fourth and fifth attempt). A student’s final quiz mark was calculated as 80% from what they scored individually and 20% from their group score.

A survey instrument based on previous IF-AT literature was developed. The survey questions were grouped into three broad categories (only two of which are reported fully due to the page limit) to investigate: the effect of the quizzes on student engagement (Category E), the effect on students’ perception of their understanding (Category C) of the subject material and the effect of immediate feedback. The instrument contained a combination of six point Likert scale (strongly disagree, disagree, slightly disagree, slightly agree, agree, strongly agree) open-ended and free response questions. The survey was paper-based and students completed it in class. To encourage students to provide honest responses without fearing anything they said could influence their subject results, the surveys were collected and collated by a third party with results only being provided to the authors after the posting of final grades.

In addition to the survey instrument, focus groups and observations were used to evaluate the impact of the collaborative quizzes on students.

Results

The percentage of responding cohort who agreed (slightly agree, agree, strongly agree) with the statements in the survey instrument are shown in Table 1. Figure 2 plots these results to enable easy comparison.

Table 1: Percentage of responding cohort who agreed with statements in the survey instrument

<table>
<thead>
<tr>
<th>Survey Statements</th>
<th>Category</th>
<th>Design Fundamentals (n=31)</th>
<th>Communications Theory (n=42)</th>
<th>Concrete Design (n=67)</th>
<th>Transmission Systems (n=81)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The use of the IF-AT cards (the scratch cards) made the group quizzes fun.</td>
<td>E1</td>
<td>97%</td>
<td>93%</td>
<td>73%</td>
<td>96%</td>
</tr>
<tr>
<td>Knowing that I would be expected to contribute to the group component of the quizzes increased my motivation to learn the relevant material.</td>
<td>E2</td>
<td>94%</td>
<td>93%</td>
<td>72%</td>
<td>84%</td>
</tr>
<tr>
<td>Having to collaborate to decide on answers during the quizzes has improved my understanding of the subject material.</td>
<td>C1</td>
<td>90%</td>
<td>88%</td>
<td>72%</td>
<td>89%</td>
</tr>
<tr>
<td>Having to collaborate to decide on answers during the quizzes has improved my ability to think through and resolve problems.</td>
<td>C2</td>
<td>90%</td>
<td>98%</td>
<td>72%</td>
<td>93%</td>
</tr>
<tr>
<td>Having to discuss the answers to the quiz with my group members helped me to understand material that I hadn’t previously fully understood by myself.</td>
<td>C3</td>
<td>94%</td>
<td>90%</td>
<td>79%</td>
<td>86%</td>
</tr>
</tbody>
</table>
Discussion

The results presented in Table 1 and Figure 2 show that the response to the collaborative activities using the IF-AT cards was overwhelmingly positive. We found that in all trials students reported the quizzes promoted high engagement, that the conversations they had with their group peers helped them identify and subsequently address gaps in their knowledge, that the IF-AT cards helped them clarify their understanding, caused them to reflect and increased their motivation to learn. Furthermore, these benefits were evident even when students got the correct answer on their first attempt.

“The collaboration components of the quizzes assisted me in my learning throughout the semester.” (Comms Theory student)

“The heavy use of group/teamwork was very effective in consolidating everyone’s knowledge, and kept us on the ball.” (Design Fundamentals student)

Encouragingly, students across all four subjects reported that the collaborative quizzes helped them to understand material that they had not previously fully understood on their own.

“Having an individual quiz followed by a group quiz forced discussion of the questions which was often productive in terms of learning” (Concrete Design student)

Our observations in all the subjects noted the high levels of student engagement in the collaborative quizzes as evidenced by the high noise levels in class as students animatedly discussed their answer to the questions. This is reflected in the student’s agreeing that using the IF-AT cards made the group quizzes fun (73% Concrete Design, >90% in the other courses).

We found that the collaborative quizzes not only improved student engagement and learning, but helped change their learning culture by having them take more responsibility for their own learning, improved their ability to provide critical evaluation and exposed them to the benefits of collaborative peer learning.

“Mentor groups allowed me to have others to ask when I was stuck both in and out of class. …. I learnt more.” (Comms Theory student)
“Improved my way of thinking.” (Comms Theory student)

Students’ also overwhelmingly approved of the fact that their learning did not stop when they got an incorrect answer.

“We really learnt more and thought more” (Design Fundamentals student – note underlining provided by student)

While the collaborative IF-AT activities were successful in all subjects the authors noted and the students reported, not surprisingly, that the physical environment had an impact on their ability to collaborate. Design Fundamentals and Communications Theory were held in rooms with flat floors and movable tables and chairs. In contrast, both Concrete Design and Transmission Systems being larger classes were held in traditional tiered lecture theatres. In Concrete Design (73%) and Transmission Systems (65%) of students agreed that “the physical environment of the lecture theatre made it difficult to participate adequately in the group quizzes”. Students found it easier to collaborate when the physical environment allowed them to sit comfortably facing each other around the shared workspace. The authors also observed that students seemed to be more animated and engaged in the flat desk environment. As more instructors move toward using collaborative activities, demand for rooms that allow flexible seating arrangements is likely to increase.

Future Research

It is unclear why the results for Concrete Design were not as positive as the other three subjects whose cohorts consistently reported 90% agreement with the statements in the survey instrument as opposed to approximately 75% agreement in Concrete Design. The authors intend to investigate this further however, the fact that this was the Concrete Design lecturer’s first experience in using the IF-AT cards probably contributed somewhat to these differences. Furthermore, the culture of the Civil Engineering students in Concrete Design was observed to be quite different compared to students in the other subjects (Design Fundamentals being a core subject has a multidisciplinary cohort, while Transmission Systems and Communications Theory are Telecommunication Engineering subjects).

Firstly, in Concrete Design a number of students expressed their resentment at being expected to actively participate in class, a complaint not evident in the other subjects. Secondly, because many Concrete Design students were not prepared to stay after the collaborative quiz to discuss their answers, the implementation of the quiz process differed from the other subjects. In Concrete Design the lecturer provided worked solutions to the quiz questions a week after the quiz. In the other subjects, a collaborative discussion led by the lecturer occurred immediately after the quiz was complete and no worked solutions were provided. This delay and the fact that worked solutions were available may have contributed to reduced engagement with the quiz process contributing to the lower positive response rate in Concrete Design.

Interestingly, unlike the other subjects, the quizzes in Design Fundamentals were formative (having no assessment marks associated with them) yet the students still enthusiastically participated. This may have been in part due to the small cohort (31 students) involved. In a
future study, we intend to investigate if participation in larger classes is significantly affected if the quizzes are formative. In other studies, the authors have found that with the correct scaffolding formative activities have the potential to improve learning outcomes by freeing students from the burden of strategically collecting marks, allowing them to focus on learning. Furthermore, the authors found that students often approach summative tasks, with some justification, strategically to achieve the best mark at the expense of learning (eg they may choose to divide up work, or move on without having their knowledge gaps addressed to save time) (Willey & Gardner, 2011).

Conclusion

We found that in all four subjects students reported the collaborative quizzes promoted high engagement, that the conversations they had with their group peers helped them identify and subsequently address gaps in their knowledge, that the IF-AT cards helped them clarify their understanding, caused them to reflect and increased their motivation to learn. We also recommend that to get the most out of collaborative activities students should be able to work in an environment where they can face each other around a shared workspace. While further investigations are needed to determine the impact of formative versus summative implementations, the activities were successful in improving student learning and engagement. In addition, by simulating the way professionals learn in industry such collaborative activities help students develop the skills required for continuous learning in the workplace.

References


Epstein www.epsteineducation.com/home


Copyright © 2011 Willey & Gardner: The authors assign to AaeE and educational non-profit institutions a non-exclusive licence to use this document for personal use and in courses of instruction provided that the article is used in full and this copyright statement is reproduced. The authors also grant a non-exclusive licence to AaeE to publish this document in full on the World Wide Web (prime sites and mirrors) on CD-ROM or USB, and in printed form within the AaeE 2011 conference proceedings. Any other usage is prohibited without the express permission of the authors.
Threshold exams to promote learning and assurance of learning


Keith Willey & Anne Gardner
Faculty of Engineering & IT, University of Technology, Sydney
Keith.Willey@uts.edu.au

BACKGROUND

Formal examinations are often used in engineering classes as the tool to evaluate student learning. These exams are often high stakes assessment tasks and provide no opportunity for feed-forward. Despite academic claims that all topics in their subject are requisite material, students are regularly able to pass these assessment tasks with unsatisfactory, and perhaps even no capacity to demonstrate learning in some topics. Furthermore, while undertaking the exam often highlights to students their learning deficiencies, it typically has no impact on their learning as they rarely receive feedback other than a mark or grade and there is no further opportunity to address these learning gaps. This paper reports on the impact of a two-staged examination process on both student learning and assurance of that learning.

PURPOSE

The aim of the staged examination process was to improve confidence that students had satisfactory knowledge in all requisite subject topics and to test its capacity to be learning-oriented in that it provides improved opportunities for students to learn while simultaneously increasing the level of learning assurance.

DESIGN/METHOD

The first stage of the process was an exam that covered all requisite subject topics. This exam consisted of multiple choice questions set at or just above the level of threshold learning outcomes. Students were required to score 80% on this exam to qualify to undertake the second part of the assessment process at a later date. Students used IFAT (Immediate Feedback Assessment Technique) cards for this stage to facilitate immediate feedback as to their strengths and weaknesses. The time between exams allowed students to review identified areas of weakness before attempting the second stage of the exam. Note: while not contributing to their final grade students who failed the first exam were also permitted to undertake the second exam as an opportunity to learn and as a means of evaluating the impact of the process. The second exam consisted of open-ended questions requiring students to explain their critical thinking and judgment used to arrive at their
answer. Evaluation of the effectiveness of this process was based on a student survey, focus group discussions and an analysis of student examination scripts.

RESULTS

The threshold learning outcome exam was effective in improving assurance of learning in that students had to demonstrate satisfactory learning across topics to achieve the 80% required to ‘pass’ the exam. Furthermore, students reported that they used the opportunity between exams to address identified learning gaps, hence demonstrating the learning orientation and feed-forward capacity of the two stage process. However, the fact that two students who did not achieve the threshold level of 80% in the first exam were able to address their learning gaps and pass the second and harder exam suggests that an alternative to the 80% exclusion criteria should be considered.

CONCLUSIONS

The study demonstrated that a two staged examination process improved confidence in assurance of learning while providing students with an opportunity to first identify and subsequently reduce learning gaps. However, the fact that some students who failed the threshold exam demonstrated significant improvement in their understanding in the second exam suggests that more research is needed to both understand the impact of and improve the benefits from this activity.

KEYWORDS

Assurance of learning, Feed-forward, Assessment design

Introduction

Formal examinations are often used in engineering classes as the tool to ‘measure’ student learning. These exams are often high stakes assessment tasks that provide no opportunity for feed-forward (apply feedback to subsequent learning). Despite academic claims that all topics in their subject are requisite material, students are often able to pass these assessment tasks with unsatisfactory, and perhaps even no capacity to demonstrate learning in some topics. Furthermore, while undertaking the exam often highlights to students their learning deficiencies, it typically has no impact on their learning as they rarely receive feedback other than a mark or grade and there is no further opportunity to address these learning gaps.

To address these issues we investigated the effectiveness of a two-staged examination process on improving confidence that passing students had satisfactory knowledge in all requisite subject topics while simultaneously being learning-oriented through providing feed-forward to students on their understanding of the subject material. This paper reports on the impact of this two-staged examination process on both student learning and assurance of that learning.
Background

Much of the literature on assurance of learning focuses on the program, institutional or national level. One of the issues to be resolved in a national quality and standards framework is the “... tension between minimum or threshold standards on the one hand, and excellence ... on the other” (Krausse, Barrie and Scott, 2012). While recognising the value in conversations at the institutional level in regard to what we mean by learning standards, we argue, along with Sadler (2009, 2010) and Knight (2006) that the reliability of such learning standards depends on the quality of the assessment in individual subjects. Sadler (2009, 2010) discusses the concept of assessment fidelity, defining this as “... the extent to which components that contribute to a course grade are correctly identified as academic achievement” (p.728) Sadler (2010) discusses ‘effort’ and ‘attendance’ as examples of components of a subject grade that do not provide evidence of learning outcome achievement. The authors, in discussions with academics and students at a number of universities, found regular instances of marks being given for activities that did not provide evidence of a subject’s learning outcome achievement including:

- revision of pre-requisite material: for example a 20% assessment task being given in week 3 of the semester for students to revise the material in the two pre-requisite subjects,
- frequency of contributions to a discussion forum (without regard to the quality of these contributions),
- participation in an activity, for example self and peer assessment or peer review, without regard to the quality of their contributions.

Sadler (2010) and Price et al (2011) also challenge us on the practice of progressive accumulation of marks throughout a semester from tasks set at a lower level than the threshold level for the subject (eg simple quizzes). There are two aspects to this issue, one is that marks from these early assessment tasks reflect learning at a lower level than is expected at the end of the subject, the other is that the understanding that students have in earlier stages of the semester may be significantly less than their understanding at the end of the semester. In either case this mark accrual process can “… misrepresent the level of achievement reached at the end of the course” (Sadler, 2010, p.735).

Rather than discarding all our ‘during semester’ assessment, Knight (2006) suggests that we design our tasks for learning oriented assessment. Carless (2007) describes learning oriented assessment as assessment designed to meet both certification and learning purposes. He characterizes learning oriented assessment as having three major components:

- Assessment tasks as learning tasks,
- Students involvement in the assessment process (self and peer assessment),
- Feedback as feed forward.

Willey and Gardner (2012) present a learning framework that suggests that learning is maximised when an assessment activity provides a well-designed learning opportunity and participants (students) approach the activity with a learning focus. Furthermore, for all
assessment activities (both summative and formative) they recommend that academics should explain to students:

- why they designed the assessment activity the way they did,
- what learning opportunities the activity provides the students,
- how students can evaluate their learning from the activity, and
- how it is going to impact on their reality (enable them to see the world differently).

Hence in designing a summative examination process our specific aim was to:

- make it learning oriented and include a feedforward component, allowing students an opportunity to respond to feedback and reassess their learning.
- only provide credit for demonstrated achievement against subject learning outcomes,
- increase assurance of learning in that students were able to demonstrate satisfactory learning in all subject topics, and
- move students to approach the exam with a learning focus.

In response to the above we developed a two-staged examination process. This paper reports on the impact of this process on both student learning and assurance of that learning.

Design

The two-staged examination process was implemented as follows:

The first stage of the process was an exam covering all requisite subject topics. This exam consisted of multiple choice questions set at or just above the level of threshold learning outcomes. This exam was followed by an exam review where the material was discussed and any common misconceptions were addressed. Students answered this first stage multiple choice exam using Immediate Feedback Assessment Technique (IF-AT) cards. IF-AT cards (Figure 1) developed by Epstein Educational Enterprises allow students to immediately identify if they have answered multiple-choice questions correctly. These cards require the students to scratch off a covering over the response they think is correct (hence they are often referred to as 'scratch cards'). If they have selected the correct response a star is revealed (Figure 1). If they selected incorrectly, they consider the remaining options, and try again.

Figure 1: IF-AT card showing a star under the correct answer and multiple attempts to find the answer for Questions 2, 3 & 5.
In controlled trials the IF-AT method was shown to promote both retention of learned material (Epstein et al. 2002, Dihoff et al. 2004, Brosvic et al. 2005, Brosvic & Epstein 2007) and higher levels of independent learning (Brosvic et al. 2005; Persky & Pollack 2008). IF-AT allows students to assess their mastery of the material being assessed. The combination of immediate feedback and the capacity to think about and rework problems that they got wrong at the first attempt assists students in discovering gaps in their knowledge and areas of misconception. Each of these components has the potential to increase deep learning (Persky & Pollack 2008).

Students were required to get 80% to ‘pass’ this exam, the aim being for students to have to demonstrate satisfactory understanding in each topic in order to achieve a pass. In reality since the exam covered five broad topic areas it was theoretically possible for a student to master four of the five topics knowing them extremely well (four topics representing 80% of the assessed material) and not know one topic (one topic representing 20% of the assessed material) at all. The four scaffolding recommendations previously discussed were clearly articulated to students. They were told that the intent of the first exam was for them to have to demonstrate satisfactory knowledge in all requisite subject material; that the exam was learning oriented in that it provided immediate feedback allowing them to identify their strengths and weaknesses and recognise both misconceptions and topics that they may need to revise; and that the requirement to achieve 80% to pass this exam increased our assurance of their capacity to satisfactorily demonstrate the subject learning outcomes.

In the review that immediately followed the exam students were able to ask questions to clarify any misunderstandings and/or address any gaps in their learning. The fact that the IF-AT cards, in effect, grade the examination automatically, allowed the instructor to quickly view the cards and pay particular attention in the review to the most common mistakes. These were identified as questions where a cohort of students took multiple attempts to achieve the correct answer. The instructor conducted the review by discussing each question, clarifying any misconceptions and subsequently varying the question to check students’ understanding.

Students had their first exam scores multiplied by 0.625 to calculate the contribution to their final examination grade. That is a score of 80% in the first stage exam scaled to 50% of a student’s final examination mark, while a score of 100% in the first stage exam scaled to 60% of a student’s final examination mark.

Students who received 80% or greater in the first stage exam were invited to sit, at a later date, a second exam consisting of open-ended questions requiring students to explain their critical thinking and judgment used to arrive at their answer. The time between the exams provided students with the opportunity to revise material addressing gaps in their understanding identified through undertaking the first stage exam or during the exam review. This enabled students to feed forward the feedback they received from these activities. The questions in the second stage exam were set at a more ‘mastery’ level and hence considerably more difficult than those in the first stage exam. This is not unreasonable given that it is the result of the second exam that determines whether students received higher than a passing grade. Again the second exam was followed by an exam review providing students with an opportunity to address knowledge gaps highlighted
by the second exam and continue to learn even after the subject’s summative assessments had ceased.

Students who failed to achieve 80% in the first exam were given the opportunity to sit a supplementary exam at the end of the semester. Again, undertaking the first exam and participating in the exam review provided them with an opportunity to identify the areas of the subject that they needed to revise.

The open-ended format of the questions in the second exam meant that there was often several different ways for students to answer them. They were designed to require students to demonstrate and explain their critical thinking, reasoning and judgment. Grading these types of questions requires an experienced instructor and generally takes more time than marking mathematical or simple answer questions. However, the whole activity was designed to have a neutral impact on the marking effort required by the instructing academic. This was achieved as the second exam consisted of only four open-ended reasonably complex questions. Typically a traditional final exam would have more questions requiring more marking effort and would have to be graded for all students. The fact that in this case the first exam was marked automatically and that only students who received more than 80% were invited to do the second exam meant that grading these marking intensive open-ended questions resulted in no increase in overall workload for the instructing academic.

So in summary the examination process shown in Figure 2 consisted of:
- a multiple-choice exam covering all requisite material answered using IF-AT cards,
- an exam review discussing the examination, addressing common misconceptions and providing variation to students to check understanding,
- students who achieved at least 80% in the first exam were invited to sit a second examination consisting of more difficult open-ended questions at a later date, and,
- a further exam review discussing the examination, addressing common misconceptions and providing variation to students to check understanding.

Figure 2: Two stage examination process
Method

The subject Design Fundamentals, taught in English for the University of Technology, Sydney (UTS) in Hong Kong, was chosen as the vehicle to investigate the impact of the two-stage examination process on both student learning and assurance of that learning. The subject’s lectures are delivered in block mode over five consecutive days (four hour blocks on Friday, Saturday, Monday and Tuesday evenings and an all-day 10 hour block on Sunday). Students are provided with pre-lecture activities including readings supported by a series of formative multiple choice questions designed to help students identify gaps in their understanding.

The two-stage examination process was implemented as follows:

The first stage of the process, the multiple-choice exam covering all requisite topics, was held on the Monday evening of the block of lectures. Students calculated their own marks from the IF-AT cards scoring two points for answering correctly on the first attempt, one point on the second attempt, 0.5 points on the third attempt and 0 points on the fourth and fifth attempts. This exam was followed by a review where the instructor facilitated students discussing their answers to each question. The instructor then clarified any outstanding issues or misconceptions before varying an aspect of each problem for students to first attempt then discuss their answers to check their understanding.

The second examination was held on the following Tuesday evening enabling students to review identified areas of weakness before attempting the second stage exam.

As shown in Figure 2 students who received 80% in the first exam could accept a pass and not undertake the second exam. If the student wanted to try for a higher grade (credit, distinction or high distinction) then they were required to sit for the second exam. In addition, while not contributing to their final grade students who failed the first exam (received less than 80%) were encouraged to undertake the second exam as an opportunity to learn.

Evaluation of the effectiveness of this process was based on a student survey, focus group discussions, structured observations during the examination period and an analysis of student examination scripts. Students were asked to voluntarily identify themselves on the survey to allow the authors to investigate the effect of the two-stage process on students of varying ability, on the understanding that the surveys were sealed and not analysed until after students’ final grades had been posted.

Results

All students (n =28) chose to undertake both examinations and attend both review sessions. Furthermore, all students agreed to participate in this research and also chose to voluntarily identify themselves on the survey.

Of the 28 students who sat the first exam 15 (54%) achieved 80% or greater. Of these 5 achieved 100%, 5 achieved >= 90% and < 100% and 5 achieved >= 80% and < 90%.
Hence 13 (46%) students sat the first exam and achieved less than 80%. Of these 7 achieved >= 70% and < 80%, 3 achieved >= 60% and < 70% and 2 achieved >= 50% and < 60%. One student did not receive a grade due to academic misconduct.

The student survey amalgamating the Strongly Disagree (SD) and Disagree (D) and Strongly Agree (SA) and Agree (A) results respectively, are shown in Table 1.

Table 1: Results from the preliminary student survey amalgamating the Strongly Disagree (SD) and Disagree (D) and Strongly Agree (SA) and Agree (A) results respectively.

<table>
<thead>
<tr>
<th>QUESTION</th>
<th>SD/D</th>
<th>N</th>
<th>A/SA</th>
</tr>
</thead>
<tbody>
<tr>
<td>The requirement to achieve 80% in the first exam of the feedforward</td>
<td>0%</td>
<td>39%</td>
<td>61%</td>
</tr>
<tr>
<td>summative assessment meant that I tried to understand all topics in the</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>subject.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Immediate feedback provided by the IFAT cards in the multiple</td>
<td>4%</td>
<td>32%</td>
<td>64%</td>
</tr>
<tr>
<td>choice exam assisted me to identify areas of the subject material that</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I did not fully understand?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I revised the subject material to address these indentified gaps before</td>
<td>0%</td>
<td>26%</td>
<td>74%</td>
</tr>
<tr>
<td>the second exam.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The revision helped me in answering the questions in the second exam.</td>
<td>7%</td>
<td>14%</td>
<td>79%</td>
</tr>
<tr>
<td>Having a two staged feedforward summative assessment activity changed</td>
<td>4%</td>
<td>39%</td>
<td>57%</td>
</tr>
<tr>
<td>the way I approached this subject compared to other subjects.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Having a two staged feedforward summative assessment activity</td>
<td>4%</td>
<td>29%</td>
<td>68%</td>
</tr>
<tr>
<td>motivated me to learn more than I would typically in other subjects.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Discussion

Arguably, the most significant impact of the two-stage exam process was the improved confidence in the assurance of learning provided to both the instructor and students as a result of the requirement for students to demonstrate satisfactory learning in each topic to achieve the 80% required to ‘pass’ the first exam. This requirement also proved to be a significant motivator for student learning with 61% reporting that it motivated them to understand all the topics in the subject (see Table 1).

The immediate feedback provided by the IF-AT cards assisted students to identify gaps in their understanding. Students reported enjoying the opportunity to have a second attempt at a question they got wrong as in reconsidering their answer they had an opportunity to, address a silly mistake, reassess their thinking and/or assess how much they didn't understand. Students commented that taking more than two attempts to answer a question was a good indicator to themselves that they needed to thoroughly revise this topic.

The most frequently reported benefit by students of the two-stage process was that the combination of the first exam, exam review and time between exams enabled them to discover, identify and then review areas of the subject that they did not understand. Most students (74%) reported that they used the opportunity between exams to address identified learning gaps, with 79% reporting that this revision helped them to answer the questions in the second exam. Typical student comments were: “it was good that I found out what I didn't understand”; the “review after first exam was good for learning” and “it provided good revision before the second exam”; and the process “assisted me to identify areas of the subject I didn't understand”.

Improving learning and developing professional judgment in large classes through collaboration and self and peer assessment 72
These results demonstrate the learning orientation and feed-forward capacity of the two-stage process. Although one student who received 100% in the first exam commented that the exam was “too simple” so it didn’t help them “to identify errors in their understanding”. However, they did comment that the variation provided in the exam review enabled them to identify some gaps in understanding.

57% of students reported that the two-stage feed forward assessment activity changed the way they approach the subject compared to other subjects. Students most commonly reported that the activity gave them the opportunity to find out what they didn’t understand “improve my weaknesses after the first exam” and allowed them time to revise before sitting the second exam.

It is unclear why students felt this was better than for example a more traditional formative process where students are provided problems to revise before their final assessment tasks. However some comments in the focus group suggests that it was because both parts of the exam were worth marks improving motivation to revise, secondly, it is only under the exam conditions that they actually get to test their understanding and thirdly it was somewhat satisfying or a form of redemption to receive an opportunity and subsequent credit for addressing mistakes they made in the first exam.

68% of students reported that the two-stage feed forward assessment activity motivated them to learn more than they would typically in other subjects. The main motivator for this change appeared to be “because I needed to understand all the topics in the course to pass the exam” and that “I needed to understand more on this subject” to pass. Other students commented that the activities “motivated me to learn more on my weak areas of the subject” and “it reminds me what areas of topics I not understand so I’ll try harder to learn and get higher marks”. One student commented that “I have a chance to find out my weaknesses of the subject, to encourage me to study again to ensure I will fully understand all the subject in the course”.

The most common complaint in regard to the process reported by students in both the survey and the focus group was that it was too compressed leaving insufficient time for students to rigorously address all learning gaps they had identified. This was articulated in the following free response comments: “the time is too compressed no time to digest the material”; the learning is very compressed it effected my performance in the exam”; “not enough time after the first exam, after finding out what I misunderstood not enough time to revise this material”. Some students also raised the issue that having to get 80% in the first exam put them under too much pressure and that the requirement was unreasonable. A typical comment being “the exam is too much pressure” and “it is not easy to achieve 80%”.

In the described trial this matter was made worse with approximately 72% of the class enrolling late and subsequently not having access to the pre-work material before the block mode lectures. This provided a very compressed time frame for the students to learn the subject material. Not surprisingly, 93% of the students who failed the first exam were late enrolments.
As previously stated in the results section of this paper all students who failed the first threshold level exam chose to undertake the second and harder exam. This was promoted as a further opportunity to both learn and identify gaps in their learning. Subsequent marking of the second exam revealed that two students who failed the first exam actually passed the second exam demonstrating a significant improvement in their understanding. One student achieved more than 50% in each individual question in the second exam, while the other achieved greater than 50% in all but one question in which they achieved 41%. These students’ papers were awarded a grade without having to sit the supplementary exam as they had demonstrated satisfactory knowledge across all topics in the second exam (that was set at above threshold level). This anomaly was somewhat surprising especially given the fact that there was only one day between the exams to revise and that the second exam was significantly more difficult than the first. The authors requested a dialogue with these students to discuss this anomaly. One student agreed. This student described a preference for longer questions which were allocated more time to complete than the multiple-choice format. Arguably more importantly he reported that while he undertook significant study between the two exams focusing on the material he got wrong in the first exam his main reason for failing the first exam (he received 73%) “was careless reading” and problems with comprehension (recall the trial was conducted in Hong Kong where all students had English as a second language). This highlights the need when setting multiple-choice exams to pay particular attention to avoiding any ambiguity to reduce the chances of students misinterpreting the question. This is especially important when students are not being examined in their primary language.

To further analyse the effectiveness of the process for assurance of learning the authors examined the number of questions for which students scored at least 30% in the second exam. The 30% level was chosen to roughly correlate with what a student with satisfactory subject knowledge might be expected to achieve given that the second exam questions were set at a higher level with the intention of using it as discriminator for students to receive higher than a passing grade.

The second exam consisted of four questions. Three questions covered a single topic while the fourth question covered two topics. For comparison these results were arranged into two categories: students who passed the two stage exam process and those that were required to sit the supplementary exam even though their combined mark for exam one and two was greater than 50%. That is, the students who under normal circumstances (no requirement to achieve 80% in the requisite material) would have received a passing exam grade.

In interpreting the results shown in Figure 4 the following points should be considered: No student was required to undertake the second exam. Those that passed the first exam only had to sit the second exam if they wanted to attempt to achieve a credit, distinction or high distinction. Those that failed the first exam had to sit the supplementary exam but were encouraged to do the second exam as a means of revision and further indicating the areas in which they needed to revise. Also only 26% of students agreed that they revised gaps identified by the first exam before undertaking the second exam. Hence student performance in this exam could not be considered an absolute evaluation but is however indicative of a student’s knowledge across subject areas.
Improving learning and developing professional judgment in large classes through collaboration and self and peer assessment

The results plotted in Figure 4 show 82% of students (14 students of 17) who passed, achieved greater than 30% in at least three of the four exam questions compared to only 14% of the students (1 student of 7) who were required to sit the supplementary exam. Previously, when the authors used a single examination paper, it typically contains approximately 60% of questions aimed at the pass or satisfactory level, with the remaining 40% of questions used as the discriminators to award students a higher grade. Often this resulted in students who understand one or more topics well, passing the exam without satisfactory knowledge of the remaining requisite material. Indeed, Figure 4 shows that if the two exams were considered as one without the condition to achieve 80% in the requisite material then 6 of the 7 failing students who would have received greater than 50% overall, and hence passed, were unable to demonstrate 30% achievement in at least two topic questions.

This result clearly demonstrates the potential of a threshold exam component to improve confidence in assurance of learning for students who pass a subject. Furthermore, the two-stage exam with the exam reviews, is learning oriented, enabling students to continue to learn even after the summative stage of the examination is finished.

Recommendations

In consideration of keeping the process neutral in regard to academic workload we intend to change the process in the next trial as follows:

Stage 1: All students sit the first stage threshold exam held at the end of the semester and answered using IF-AT cards. This would be followed by an exam review.
Stage 2: One week later (allowing time for students to undertake further revision) students sit a second exam as follows:
i. Students who received greater than 80% in the first exam would be invited to sit a paper consisting of open-ended high-level questions that require critical evaluation and reasoning to answer for credit towards a higher grade (in the case of UTS credit, distinction or high distinction)
i. Students who failed to achieve 80% in the first exam, would sit another threshold level exam, again answered using IF-AT cards. Students would also be required to get 80% in this examination to achieve a pass in the subject.

These exams would again be followed by an exam review undertaken immediately before the start of the next semester.

Conclusions

The study demonstrated the potential of a two-stage examination process including exam reviews to improve confidence in assurance of learning while providing students with an opportunity to first identify and subsequently reduce learning gaps. However, the fact that two students who failed the threshold exam demonstrated satisfactory achievement in the mastery exam across topics suggests that more research is needed to improve the design, impact and benefits of this process.

References


Improving learning and developing professional judgment in large classes through collaboration and self and peer assessment


Copyright © 2012 Willey & Gardner: The authors assign to AAEE and educational non-profit institutions a non-exclusive licence to use this document for personal use and in courses of instruction provided that the article is used in full and this copyright statement is reproduced. The authors also grant a non-exclusive licence to AAEE to publish this document in full on the World Wide Web (prime sites and mirrors), on Memory Sticks, and in printed form within the AAEE 2012 conference proceedings. Any other usage is prohibited without the express permission of the authors.