CAS2 Calibrating Academic Standards Workshop Report

Date: Sunday 4 & Monday 5 February 2018
Location: University of Newcastle, NSW

Contents

Tables

Introduction

Following the first successful Calibration of Academic Standards (CAS1) workshop held in August 2017, participants agreed that calibration workshops should be held twice yearly at the CAUTHE annual conference and mid-year meeting. It was proposed that future calibration workshops might examine multiple domains. Participants were encouraged to 'volunteer' assessment tasks and samples of student work that might be suitable for this purpose. The project team developed a plan and timeline to implement this process so that all five domains at AQF levels 7 and 9 can be covered over the next five years. This program of activity is being held in conjunction with ongoing peer-review of assessment. The body of material collected from calibration workshops will form the basis of the review of the 2015 Tourism, Hospitality and Events (TH&E) Learning and Teaching Academic Standards (LTAS) after five years.

Aims

Calibration is a multistage intervention in which external peer reviewers examine an assessment task and samples of student work against one or more of the threshold learning outcome (TLO) domains before, during and after a calibration workshop (O'Connell et al 2015). For the CAS2 workshop, the assessment task and student work samples were selected from a third year **Event Project** subject in a Bachelor of Business (Event Management) program. The task was a **Strategic Plan** group project worth 40% of the assessment regime.

The workshop aimed to calibrate two of the learning standards for *Interdisciplinary Inquiry* and *Professional Responsibility* for Bachelor level (AQF 7) tourism, hospitality and events graduates.

- *Interdisciplinary Inquiry,* for Bachelor level TH&E graduates, is defined as the ability to 'Integrate a broad and coherent elementary, theoretical and operational knowledge of tourism, hospitality or events as an interdisciplinary field of research and practice.'

- *Professional Responsibility* for Bachelor level Events graduates is defined as the ability to 'Reflect on their own conduct and the performance of others to improve their own interpersonal skills and independent learning capabilities in routine events settings.'

The workshop and calibration process aimed to determine:

1. What is the quantifiable difference in grader variability on the assessment of learning outcomes in *Interdisciplinary Inquiry* and *Professional Responsibility*?

2. Does participation in the workshops lead to reduced disparity in the assessment of the students' learning outcomes in *Interdisciplinary Inquiry* and *Professional Responsibility*?

3. Does participation in the workshops lead to greater confidence by reviewers in their ability to assess students' skills in *Interdisciplinary Inquiry* and *Professional Responsibility*?

Method

Participant selection

The CAS2 workshop was held at the University of Newcastle, *NewSpace* campus on the afternoon of Sunday 4 February and morning of Monday 5 February, immediately prior to the annual CAUTHE 2018 conference. An invitation to CAUTHE's 34 Chapter directors and Affiliate members to participate in the calibration of assessment process yielded 20 participants from 13 institutions, including seven universities and six technical and further education (TAFE) institutions and private providers (Appendix 1). This represented an increase from CAS1, which included 13 participants from 11 institutions. An informal briefing session was conducted with new participants immediately prior to the workshop.

Review process

Prior to the workshop, participants were divided into teams and asked to review and rate the validity of a de-identified assessment task, in terms of its ability to assess *Interdisciplinary Inquiry* and *Professional Responsibility* at AQF 7. They were then asked to review two de-identified samples of student work for the assessment task and assess whether the samples met (or did not meet) the threshold for that domain. The participants provided feedback and justification for their ratings using the *Qualtrics* online survey system. Once all reviewers had entered their evaluations, the results of all reviewers and their comments were

shared with all team members. The tool reported key metrics (highest, lowest, mean, median, mode, and standard deviation) as well as all verbatim comments.

There were eleven (11) respondents for the pre-workshop survey. The anonymised results were published and circulated to participants prior to the workshop as a key resource so that each participant could compare their assessments with that of their peers. Participants reported that the pre-workshop review took between one to fifteen hours, with new participants reporting that it took additional time to understand the background information.

The workshop involved participants meeting together to discuss and reflect upon their judgments and the justification for their assessment of the validity of the assessment task and the samples of student work. The aim was for this respectful sharing of results and the discussion about reasons for giving the grades would help the assessors achieve consensus in applying the *Interdisciplinary Inquiry* and *Professional Responsibility* domains.

In three teams of five to eight people, participants firstly established that the assessment task was valid for demonstrating achievement of the *Interdisciplinary Inquiry* domain. They then repeated the consensus reaching process through open discussion, benchmarking the two pieces of student work against *Interdisciplinary Inquiry*. Finally, participants repeated the consensus-reaching activity with a new anonymised sample of student work; exploring and discussing their results and reasons until there was agreement on the threshold standard for a graduate. This process was conducted over three hours on Sunday afternoon. The workshop process was then repeated for the *Professional Responsibility* domain over two hours on Monday morning.

Timeline

The tasks, responsibilities and timeline for CAS2 is shown in Table 1.

Table 1: Calibration of academic standards timeline

Task	Resp	Date
Send workshop invitation to CAUTHE Chapter directors & project participants	Penny	29 August 2017
Choose TLO(s) to calibrate (and the AQF level) & identify an assessment item and a range of (3) marked examples	Paul, Pierre	18 December 2017
Set up a survey space (Qualtrics) where colleagues provide feedback on assessment items prior to the workshop	Penny	8 January 2018
Open pre-workshop survey in Qualtrics & share the items (Dropbox)	Penny	20 January 2018
Deadline for reviews	All	2 February 2018
Publish and circulate aggregated results	Penny	3 February 2018
Calibration of Academic Standards (CAS2) workshop	All	4 February 2018

The workshop agenda is attached in Appendix 2.

Results

Assessment task

Interdisciplinary Inquiry

General feedback was that the assessment task required students to draw upon and integrate a broad range of disciplinary knowledge. When graded on a scale of 1 to 100, there was a **range** of 41 to 100 for the assessment item, with a **mean** of 81 and **standard deviation (SD)** of 16.5 (Table 2). This range of views was reflected in the written comments and during the workshop discussion. Participants were divided about whether the groupwork required for the assessment adequately allowed students to 'independently and rigorously demonstrate' that they could meet the national learning standard. Some participants felt that each student could potentially focus on just one or two disciplinary areas, rather than being exposed to the full range of areas required to bring an event together. Further, there was no weighting assigned for 'interdisciplinarity' in the marking rubric and there was no requirement for theoretical references.

Others felt that the inclusion of meetings and minutes provided evidence that students were required to work together to integrate the range of concepts to effectively develop the strategic plan. Further, the mandatory word count would require students to provide a depth of research and theory, thus enabling them to demonstrate a solid level of interdisciplinary understanding.

Participants agreed that the task was valid for demonstrating achievement of the national standard for interdisciplinary inquiry in tourism, hospitality and events at AQF level 7. However, the 'interdisciplinarity' is implied and the need to 'integrate' knowledge is not necessarily established. The marking rubric could benefit from an item that addresses integration across different disciplinary areas and/or a checklist (requirement) that a student undertakes a certain number of disciplinary categories.

Table 2: Comparisons of reviewers' scores on interdisciplinary inquiry pre- and post-calibration

Activity	Task	Min	Max	Mean	SD	Differ	Count
Pre-calibration	Assessment item	41	100	81	**16.5**	59	11
	Student work 1	31	70	53	**12.0**	39	11
	Student work 2	53	81	72	**9.0**	28	11
	Student work 1 & 2	31	81	62	**14.4**	48	11
Post-calibration	Student work 3	20	70	48	**13.5**	50	15

0-50 NA = Not Acceptable / 50-100 A = Acceptable

Professional Responsibility

In general, participants did not feel that the assessment task was valid for demonstrating achievement of the national standard for professional responsibility. There again appeared to be some confusion around the purpose of the rating scale, which is intended to rate the degree to which the task met the standard, rather than to grade the task. Despite a high **SD** of 21.9 and **mean** of 50 (Table 3), there was a great deal of consistency in the qualitative comments about the validity of the task. These comments were confirmed in the workshop discussions, during which participants agreed that while the task required students to include meeting agendas and minutes which provided some understanding of professional responsibility, there was no requirement for critical reflection on one's own conduct – which is paramount for this domain. The inclusion of an individual reflective piece and a peer and self-review component would address this standard more explicitly.

Table 3: Comparisons of reviewers' scores on professional responsibility pre- and post-calibration

Activity	Task	Min	Max	Mean	SD	Differ	Count
Pre-calibration	Assessment item	31	80	50	**21.9**	49	11
	Student work 1	10	66	35	**21.5**	50	11
	Student work 2	20	85	55	**22.1**	65	11
	Student work 1 & 2	10	85	45	**23.6**	75	11
Post-calibration	Student work 3	5	60	27	**19.5**	55	13

0-50 NA = Not Acceptable / 50-100 A = Acceptable

Specific feedback about the assessment task included:

- The value of building reflection into assessment requirements to close the loop i.e. include a component for individual contribution, peer-assessment or peer-reflection.

- Use language from the standard/s in the marking rubric e.g. 'theoretical', 'interdisciplinarity', 'operational', 'critical reflection'.

- Include explicit reference to the standard/s being addressed so that students clearly understand the expected outcomes.

Student work

Interdisciplinary Inquiry

There was an overall **range** from a low of 31 to a high of 70 (difference 39) in the ratings for **student work 1**, with a **mean** of 53 and **SD** of 12. Participants generally agreed that the work lacked the depth of research expected of a capstone, final year unit. There were no academic references or evidence to demonstrate 'broad and coherent theoretical knowledge'. After discussion, the participant whose rating was significantly higher than other team members (70), acknowledged that a lower rating would have been more appropriate.

Student work 2 rated as a better piece of work, with a **mean** of 72, a lower **SD** of 9 and a difference between the minimum and maximum ratings of 28. This was consistent with the comments that it was a comprehensive and well researched plan, which demonstrated greater depth of research and analysis and a higher level of elementary, theoretical and operational knowledge. This result is consistent with previous workshop results which found that better quality work declares itself, while poorer work is more difficult to judge.

Professional Responsibility

Participants felt that the assessment design prevented students from fully demonstrating the professional responsibility standard due to the absence of the requirement for reflective learning about the individuals' learning experience.

The ratings for **Student work 1 ranged** from 10 to 66 with a **mean** of 35 and **SD** 21.5. Comments were that there was very little evidence that students had reflected on their own conduct and the performance of others, nor was this requested within the assessment criteria. There was no critical reflection on their individual contributions to the project, nor on their future roles within industry. The meeting agendas and minutes lacked detail and provided no evidence of being used to monitor and ensure progress and participation of all group members.

Student work 2 rated better with a **mean** of 55, with a large **range** in scores from 20 to 85 with a **SD** of 22.1. The large difference (65) in scores was reflected in the comments and discussion. Whilst participants recognised that this work was not a reflective piece, there was evidence of engagement with topics such as sustainability, ethics and stakeholder engagement. The meeting minutes included team evaluation of each members' contribution. There was evidence of reflection about the potential impacts of the event and on performance objectives such as 'food miles'.

Others felt that the work did not meet the standard because the domain was not requested within the assessment criteria there was no evidence of reflection or discussion about conduct. After discussion, one participant who had rated the work highly (85), acknowledged that a lower rating was more appropriate.

Post-calibration

Following discussion, participants were asked to review and rate assessment item student work 3. In relation to interdisciplinary inquiry, the work was rated just below the middle of the range with a **mean** of 48. The **SD** 13.5 was lower than the **SD** 14.4 for the combined student work prior to calibration. By team, the SD increased for all teams, with the greatest increase for the red team (Table 4).

There was a small increase in the gap between minimum and maximum ratings of 50, up from 48. There were 15 (rather than 11) respondents for this survey. Participants agreed that the assessment was poor. While there was some integration of theory and practice, it was heavily operational with significant gaps in the resulting work.

Table 4: Comparisons of reviewers' scores for interdisciplinary inquiry by team

Team	Student Work	Minimum	Maximum	Mean	SD	Count
Blue	Sample 1 (Pre)	41	60	54	8.9	3
	Sample 2 (Pre)	61	80	70	7.8	3
	Sample 3 (Post)	30	60	42	9.8	5
Red	Sample 1 (Pre)	53	81	69	11.2	5
	Sample 2 (Pre)	31	70	51	13.9	5
	Sample 3 (Post)	20	65	51	14.5	7
Green	Sample 1 (Pre)	45	68	56	9.3	3
	Sample 2 (Pre)	75	81	77	2.6	3
	Sample 3 (Post)	44	70	57	10.6	3
Total	Sample 1 (Pre)	41	81	53	12.0	11
	Sample 2 (Pre)	31	81	72	9.0	11
	Sample 3 (Post)	20	70	48	13.5	15

For professional responsibility, the work was rated with a lower **mean** of 27 and a **SD** 19.5. This result represented a significant (17%) reduction from 23.6 in SD for the combined student work. There was also a substantial reduction in the gap between minimum and maximum ratings of 55, down from 75. There were 13 respondents for this survey. Interestingly, the reduction in SD by team was greatest for the red team (Table 5).

Table 5: Comparisons of reviewers' scores for professional responsibility by team

Team	Student Work	Minimum	Maximum	Mean	SD	Count
Blue	Sample 1 (Pre)	10	55	25	21.2	3
	Sample 2 (Pre)	20	85	45	28.4	3
	Sample 3 (Post)	20	45	33	10.3	4
Red	Sample 1 (Pre)	19	55	36	14.2	5
	Sample 2 (Pre)	20	70	55	18.3	5
	Sample 3 (Post)	5	14	9	2.8	6
Green	Sample 1 (Pre)	10	66	44	24.2	3
	Sample 2 (Pre)	51	75	64	9.8	3
	Sample 3 (Post)	52	60	56	3.3	3
Total	Sample 1 (Pre)	10	66	35	21.5	11
	Sample 2 (Pre)	20	85	55	22.1	11
	Sample 3 (Post)	5	60	27	19.5	13

A number of participants commented that there was no explicit requirement for reflection in the assessment and that in the absence of meeting minutes, there was no indication that reflection on self or others had occurred around performance or impact. In contrast, others pointed to implicit evidence of reflection e.g. the stated need to provide an event permit (professional responsibility); contact and communication with event sponsors and participants (interpersonal communication skills); proposed volunteer induction (professional responsibility) and evaluation of event success (reflection on own and other's performance and conduct). The assessment task would benefit from scaffolding within the rubric and/or assessment outline about what the term 'professional responsibility' means, and how this will be assessed for a consistent basis of marking/meeting the standard.

The results for the 'interdisciplinary inquiry' domain did not convincingly confirm the effectiveness of undertaking calibration to improve assessor consistency. In addition, there was only a small improvement in consistency for the 'professional development' domain. A factor which affected the results was the nature of the two domains addressed at this workshop. These domains were less 'straightforward' than problem solving, which was the focus of the CAS1 workshop. Participants' understanding of these dimensions varied significantly both prior to the workshop and had not moved significantly after the workshop discussions. The differences were exacerbated by the fact that there were more respondents for the post- than pre-calibration surveys.

Reviewer confidence and feedback

The calibration workshop activity, including reaching consensus on the pre-workshop rating exercise and reflecting on additional student work in the context of the agreed national learning standards, contributed to an enhanced understanding of standards that might apply locally. One workshop participant reported that 'the calibration workshop process is a valuable innovation for tourism, hospitality and events tertiary education, which has enabled participants to collectively improve their teaching and learning'.

Table 6 shows that the confidence rating pre- and post-calibration decreased slightly from 45% to 42% for 'strongly agree' and increased for 'somewhat agree' from 45% to 48% (interdisciplinary inquiry) and 52% (professional responsibility). Participants agreed that as a result of being involved with the process they were generally more confident in rating:

- the capacity of assessment task requirements to allow students to demonstrate the national learning standard in the two domains for in tourism, hospitality and events.

- students' ability in the two domains benchmarked against the national standard.

It was clear at the workshop that participants (on the red table in particular) found it difficult to come to a common understanding of the interpretation of the domains. This was reflected in the post-workshop survey feedback comments:

- At times, I became more confused after hearing other viewpoints, then kept doubting my own judgements! Perhaps I could have done with more background or examples of the process. Anyway, I will be better prepared next time having been through it once! Learnt a lot!

- More contextual information [is needed] for new members of group. A glossary of terms so members can be closer to definitions (not standard definitions) of the terms that are being used.

- [Need to] formulate a set of expectations and standards before meeting for the calibration exercise

- I would have liked the group to have consolidated its understanding of the TLOs. The group I worked within had very different interpretations of the TLOs. Some prescriptive guidance on the intended interpretation of each TLO would have assisted us greatly.

Table 6: Results for pre- and post-calibration reviewers' confidence (n=11)

Question	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree
Pre-calibration				
I am confident rating the capacity of assessment task requirements to allow students to demonstrate the national tourism, hospitality and events learning standard for interdisciplinary inquiry & professional responsibility.	4	7	0	0
I am confident that my feedback, explaining my ratings and offering suggestions, will be useful to the assessor.	3	7	1	0
I am confident rating tourism, hospitality and events student's interdisciplinary inquiry and professional responsibility ability as benchmarked against the national standard.	5	6	0	0
The pre-workshop activity, requiring me to reflect on the students work in the context of the agreed national learning standards, changed my understanding of standards for interdisciplinary inquiry & professional responsibility that might apply locally.	2	4	2	3
Having to enter my ratings as scale in the survey tool required me to formalise my judgement.	7	4	0	0
Having to write my feedback in the survey tool caused me to reflect on the reasons for my judgement.	9	2	0	0
Total	30	30	3	3
	45%	45%	5%	5%
Post-calibration – interdisciplinary inquiry				
Following the calibration workshop, I am confident rating the capacity of assessment task requirements to allow students to demonstrate the national learning standard for interdisciplinary inquiry in tourism, hospitality and events.	3	7	0	1
Following the calibration workshop, I am confident rating students' interdisciplinary inquiry ability benchmarked against the national standard.	3	7	1	0
The calibration workshop activity, including reaching consensus on the pre-workshop rating exercise and reflecting on a third sample in the context of the agreed national learning standards, changed my understanding of standards for interdisciplinary inquiry that might apply locally.	8	2	1	0
Total	14	16	2	1
	42%	48%	6%	3%
Post-calibration – professional responsibility				
Following the calibration workshop, I am confident rating the capacity of assessment task requirements to allow students to demonstrate the national learning standard for professional responsibility in tourism, hospitality and events.	3	7	0	1
Following the calibration workshop, I am confident rating students' professional responsibility ability benchmarked against the national standard.	3	7	0	1
The calibration workshop activity, including reaching consensus on the pre-workshop rating exercise and reflecting on a third sample in the context of the agreed national learning standards, changed my understanding of standards for professional responsibility that might apply locally.	8	3	0	0
Total	14	17	0	2
	42%	52%	0%	6%

Conclusions

The calibration activity resulted in wide variation in the ratings and possibly some confusion about the definitions of the domains. The workshop highlighted the need for further clarification on definitions for each threshold and discipline. This may include formulating a set of expectations and standards before the calibration workshop for example:

- **Interdisciplinary inquiry.** Is it the sum total of the disciplines involved, or a synthesis and application of the theories? What counts as a discipline? Is it necessary for individuals to integrate knowledge across disciplines or is it sufficient for students to demonstrate that they can work with students from other disciplines to achieve an outcome?

- **Professional responsibility.** The reasons for different descriptors for each field need to be clarified. This is affected by the opportunities to impact on the performance of others for example:

  - *Tourism:* multiple stakeholders, opportunity to impact the economy and tourism development space, policy and planning early in the cycle

  - *Hospitality:* limited factors such as energy, water, liquid and solid waste

  - *Events:* less constrained than hospitality, but still conducted in built environments, cyclical ebb and flow

As in the CAS1 workshop, there again appeared to be some confusion around the purpose of the rating scale, which is intended to rate the degree to which the samples of student work demonstrated achievement of the standard, rather than grading the work. There was also some debate about how much detail should be provided to students - e.g. for a third-year subject, such detailed guidance for reporting may deter critical thinking and self-directed learning. However, there were some rich comments and reflection in the comments in both the survey feedback and workshop discussions.

In summary, the calibration process determined:

1. There was a quantifiable difference in grader variability on the assessment of learning outcomes in *Interdisciplinary inquiry* and *Professional responsibility*

2. Participation in the workshop lead to a reduction of the disparity in the assessment of the students' learning outcomes in *Interdisciplinary inquiry* and *Professional responsibility* – but further work is required.

3. Participation in the workshop generally lead to increased confidence by reviewers in their ability to assess students' skills in *Interdisciplinary inquiry* and *Professional responsibility.*

These results suggest a need to develop additional resources (e.g. assessment samples, online videos), which can be used by participants to better understand the five TLO domains and the calibration and peer review processes undertaken by the CAUTHE *College of Peers*.

Next steps

Participants were keen to continue the process and the current timing of the meetings. The next calibration workshop should be held in conjunction with the Mid-Year meeting with the focus on the final two domains at AQF 7 *Collaboration* and *Service and Experience Design*. It was proposed that future calibration workshops might focus on professional development e.g. assessment setting to meet the standards, rather than 'retrofitting' the TLOs to existing examples. Participants were encouraged to 'volunteer' assessment tasks and samples of student work that might be suitable for this purpose. The calibration program of activity will continue to be held in conjunction with ongoing peer-review of assessment.

Acknowledgements

References

O'Connell, B., De Lange, P., Freeman, M., Hancock, P., Abraham, A., Howieson, B., et al. (2015). Does calibration reduce variability in the assessment of accounting learning outcomes? Assessment & Evaluation in Higher Education. Retrieved from http://www.tandfonline.com/doi/full/10.1080/02602938.2015.1008398

Higher Education Standards Framework (Threshold Standards) (2015) TEQSA.

Appendix 1: Calibration of assessment participants

Participants were invited from CAUTHE Chapter member institutions, many of whom were involved with the previous [Setting the Standard: Establishing Threshold Learning Outcomes (TLOs) for TH&E](#) and [Meeting the Standard Peer-Review of Assessment and Calibration](#) projects.

Blue team

1. Associate Professor Pierre Benckendorff, The University of Queensland (Lead)
2. Dr David Beirman, University of Technology Sydney
3. Dr George Brown, International College of Hotel Management (ICHM)
4. Ms Marcela, Fang William Angliss Institute
5. Mr Warren Guest, Holmesglen Institute
6. Mr Alex Menge, International College of Management Sydney
7. Mr Sunny Vinnakota, Academies Australasia Polytechnic
8. Mr Brian Weir, University of Canberra

Red team

9. Associate Director Paul Whitelaw, William Angliss Institute (Lead)
10. Dr Naomi Dale, University of Canberra
11. Ms Sharon Hebdon, Holmesglen Institute
12. Ms Kellie Lumsden, International College of Hotel Management (ICHM)
13. Dr Niki Macionis, The University of Queensland
14. Associate Professor Laurie Murphy, James Cook University
15. Ms Karen McMurray, International College of Management Sydney

Green team

16. Dr Mieke Witsel, Southern Cross University (Lead)
17. Associate Professor Elspeth Frew, La Trobe University
18. Ms Esther Teo, Academies Australasia Polytechnic
19. Dr Aaron Tham, University of the Sunshine Coast
20. Dr Margee Hume, Kaplan Business School

Project Manager

21. Mrs Penny Jose ([CAUTHE Secretariat](#))

Appendix 2: Workshop agenda

Meeting the Standard: calibration and peer-review project

Calibrating Academic Standards Workshop 2:
Interdisciplinary Inquiry and Professional Responsibility

University of Newcastle
Room X502, NeW Space campus
Auckland St & King St, Newcastle
4-5 February 2018

Agenda

Sunday 4 February (Room X502)

3.00 – 3.15pm	Welcome (Pierre Benckendorff)	
3.15 – 4.00pm	Session 1	Interdisciplinary Inquiry – Assessment task validity
4.00 – 4.15pm	Break	
4.15 – 5.00pm	Session 2	Interdisciplinary Inquiry – Student work samples 1 & 2
5:00 – 6:00pm	Session 3	Interdisciplinary Inquiry – Student work sample 3 Post-calibration survey & results
6.00 – 7.30pm	Dinner	[The Dockyard](#) 13/1 Honeysuckle Drive, Newcastle

Monday 5 February (Room X205 – yes, this is a different room!)

9:00 – 9.45am	Discuss draft peer-review responses, issues and experiences to date (peer-review phase 2 participants only)	
9.45 – 10.30am	Session 1	Professional Responsibility – Assessment task validity
10.30 – 10.45am	Break	Morning tea
10.45 – 11.15am	Session 2	Professional Responsibility – Student work samples 1 & 2
11.15 – 12.00pm	Session 3	Professional Responsibility – Student work sample 3 Post-calibration survey & results
12.00pm	Close	