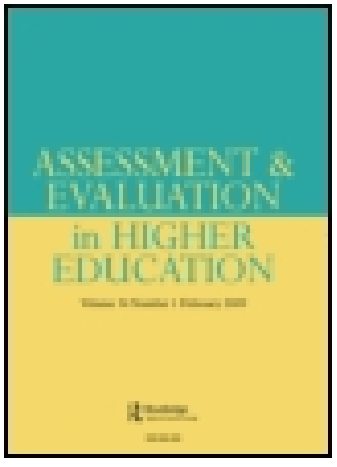

Assessment & Evaluation in Higher Education
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/caeh20

Does calibration reduce variability in the assessment of accounting learning outcomes?

Brendan O'Connell[a], Paul De Lange[b], Mark Freeman[c], Phil Hancock[d], Anne Abraham[e], Bryan Howieson[f] & Kim Watty[g]

[a] School of Accounting, RMIT University, Melbourne, Australia

[b] Curtin Business School, Curtin University, Perth, Australia

[c] The University of Sydney Business School, The University of Sydney, Sydney, Australia

[d] University of Western Australia Business School, University of Western Australia, Perth, Australia

[e] School of Accounting, University of Western Sydney, Sydney, Australia

[f] School of Accounting and Finance, University of Adelaide, Adelaide, Australia

[g] School of Accounting Economics & Finance, Deakin University, Melbourne, Australia

Published online: 24 Feb 2015.

To cite this article: Brendan O'Connell, Paul De Lange, Mark Freeman, Phil Hancock, Anne Abraham, Bryan Howieson & Kim Watty (2015): Does calibration reduce variability in the assessment of accounting learning outcomes?, Assessment & Evaluation in Higher Education, DOI: 10.1080/02602938.2015.1008398

To link to this article: http://dx.doi.org/10.1080/02602938.2015.1008398

Assessment & Evaluation in Higher Education, 2015
http://dx.doi.org/10.1080/02602938.2015.1008398

Routledge
Taylor & Francis Group

Does calibration reduce variability in the assessment of accounting learning outcomes?

Brendan O'Connell[a]*, Paul De Lange[b], Mark Freeman[c], Phil Hancock[d], Anne Abraham[e], Bryan Howieson[f] and Kim Watty[g]

[a]*School of Accounting, RMIT University, Melbourne, Australia;* [b]*Curtin Business School, Curtin University, Perth, Australia;* [c]*The University of Sydney Business School, The University of Sydney, Sydney, Australia;* [d]*University of Western Australia Business School, University of Western Australia, Perth, Australia;* [e]*School of Accounting, University of Western Sydney, Sydney, Australia;* [f]*School of Accounting and Finance, University of Adelaide, Adelaide, Australia;* [g]*School of Accounting Economics & Finance, Deakin University, Melbourne, Australia*

Reliable, consistent assessment process that produces comparable assessment grades between assessors and institutions is a core activity and an ongoing challenge with which universities have failed to come to terms. In this paper, we report results from an experiment that tests the impact of an intervention designed to reduce grader variability and develop a shared understanding of national threshold learning standards by a cohort of reviewers. The intervention involved consensus moderation of samples of accounting students' work, with a focus on three research questions. First, what is the quantifiable difference in grader variability on the assessment of learning outcomes in 'application skills' and 'judgement'? Second, does participation in the workshops lead to reduced disparity in the assessment of the students' learning outcomes in 'application skills' and 'judgement'? Third, does participation in the workshops lead to greater confidence by reviewers in their ability to assess students' skills in application skills and judgement? Our findings suggest consensus moderation does reduce variability across graders and also builds grader confidence.

Keywords: accounting education; standards; assessment; calibration

Introduction

Major worldwide developments on higher education learning standards are attracting widespread attention (Bloxham and Price 2013). Examples of the developments are: subject benchmark statements in the UK, the Tuning Educational Structures in Europe Project (2014), the Organisation for Economic Cooperation and Development's feasibility study for the Assessment of Higher Education Learning Outcomes, Lumina's Degree Qualifications Profile in the US and the Australian Quality Framework (AQF) along with threshold learning standards in Australia. It should be noted that a learning standard is the fixed level of achievement expected of a student to be awarded a recognised grade, while a learning outcome is the learning demonstrated by actual student work. The AQF initiative is the focus of this paper.

In addition to the imperatives of national interest in assurance of learning, another force is the desire to achieve comparability of cross-institutional learning

*Corresponding author. Email: brendan.oconnell@rmit.edu.au

outcomes (Bloxham and Price, forthcoming). In addition, there is considerable interest in how best to assess learning outcomes, in particular for stakeholders wanting assurance that graduate capabilities are being achieved. Typically, business schools wishing to attain external assurance of their ability to deliver capable graduates voluntarily submit to accreditation systems. For example, nearly 700 business schools in 45 countries have achieved coveted accreditation with the Association to Advance Collegiate Schools of Business International (AACSB 2014a), with reaccreditation appraisals normally occurring every five years. So-called 'assurance of learning' or 'outcomes assessment' is a key element of these appraisals (AACSB 2014b), whereby learning outcomes – and any changes arising from expectation gaps – rather than simply inputs into the learning environment are assessed (Baker et al. 1994; Apostolou 1999; Shaftel and Shaftel 2007). The reliability and validity of both internal assurance of learning processes and subsequent external appraisal rely on internal and external assessors making similar judgements. Similar judgements, though desirable, are unlikely as evidence questions any assumption that assessors' judgements, especially of non-quantitative learning outcomes, are consistent as they can be 'poorly based, erratic and unreliable' (Royce Sadler 2014, 10).

While the AACSB accreditation system assumes that programme learning objectives reflect internally derived standards, external referenced standards have been an increasing focus in many countries, including in Europe and Australia. Moreover, in the US, the Lumina degree qualification profiles, while not externally derived standards, seek to create a shared framework across different degree levels – an initiative strongly supported by the Association of American Colleges and Universities (AAC&U 2011).

Within the accounting discipline in Australia, five threshold learning standards have been developed: judgement, knowledge, application skills, communication and teamwork, and self-management (Hancock and Freeman 2010; Freeman and Hancock 2011). Table 1 provides an overview of each of these standards. These threshold learning standards were developed for the accounting discipline in 2010 and represent a consensus on the threshold that any bachelor or coursework master graduate of any Australian higher education provider should be able to achieve by graduation. The learning standards targeted in this study are students' skills in application skills and judgement.

Several researchers have raised doubt as to whether examiners can assess such standards consistently (e.g. Sadler 1987; Elander and Hardman 2002; Bloxham 2009; UK Higher Education Academy 2013). Bloxham and Price (2013, 3–4), in reference to the mandatory external examiner system in the UK, argue:

> Reports and inquiries fail to investigate the more fundamental question of whether [external] examiners understand and can consistently apply academic standards in the way required by their, albeit possibly confused, role as defenders of academic standards. This is even though … most concerns about standards can be traced back to the judgement processes of markers and [external] examiners.

To redress the situation, Bloxham and Price (2013), Rust (2009) and Sadler (2013) recommend assessors participate in consensus moderation forums, which will, over time, result in a reduction in grader variability. In response to this observation, the present study seeks to provide evidence as to the extent of the problem of grader variability and whether a consensus moderation process can reduce the variability of assessors' judgements.

There is a paucity of prior research into the consensus moderation process itself and its effectiveness (Orr 2007; Yorke 2008; Bloxham 2009). Moreover, very little

Table 1. Threshold learning outcomes for accounting.

	Bachelor graduates in accounting will be able to:	Master (entry) graduates in accounting will be able to:	Master (advanced) graduates in accounting will be able to:
Judgement	Exercise judgement under supervision to solve routine accounting problems in straightforward contexts using social, ethical, economic, regulatory and global perspectives	Exercise judgement under supervision to solve routine accounting problems in diverse contexts using social, ethical, economic, regulatory and global perspectives	Exercise judgement under minimal supervision to solve emerging and/or advanced accounting problems in complex contexts using social, ethical, economic, regulatory and global perspectives
Knowledge	Integrate theoretical and technical accounting knowledge which includes a selection of auditing and assurance, finance, economics, quantitative methods, information systems, commercial law, corporation law and taxation law	Integrate theoretical and technical accounting knowledge which includes a selection of auditing and assurance, finance, economics, quantitative methods, information systems, commercial law, corporation law and taxation law	Integrate advanced theoretical and technical accounting knowledge which includes a selection of auditing and assurance, finance, economics, quantitative methods, information systems, commercial law, corporation law and taxation law
Application skills	Critically apply theoretical and technical accounting knowledge and skills to solve routine accounting problems	Critically apply theoretical and technical accounting knowledge and skills to solve routine accounting problems	Critically apply advanced theoretical and technical accounting knowledge and skills to solve emerging and/or advanced accounting problems
Communication and teamwork	Justify and communicate accounting advice and ideas in straightforward collaborative contexts involving both accountants and non-accountants	Justify and communicate accounting advice and ideas in diverse collaborative contexts involving both accountants and non-accountants	Justify and communicate accounting advice and ideas in complex collaborative contexts involving both accountants and non-accountants
Self-management	Reflect on performance feedback to identify and action learning opportunities and self-improvements	Seek and reflect on performance feedback to identify and action learning opportunities and self-improvements and initiate this process for others	Seek and reflect on performance feedback to identify and action learning opportunities and self-improvements and initiate this process for others

Source: Hancock and Freeman (2010, 10).

research has been conducted on assessor reliability using a rigorous experimental design. This is despite acknowledgement from bodies, such as the American Accounting Association Teaching and Curriculum Section Outcomes Assessment

Committee, that more reliable and accurate assessment of learning outcomes is needed (Baker et al. 1994; Shaftel and Shaftel 2007). Whereas researchers have identified the deficiencies of grading processes typically employed in education settings (e.g. Smith and Coombe 2006; Sadler 2011), empirical examination of this area is lacking and only beginning to emerge (Bloxham and Price 2013). Given these concerns, we believe it is critical that higher education institutions seek out processes that ensure judgements around learning standards are applied consistently by assessors. One potential benefit of this study is that it describes a process that may be applicable to assessment in various higher education situations.

The focus of our study is on whether a calibration workshop where assessors come together to debate and apply standards to student work can achieve consensus in applying threshold learning standards to students' work. The workshops comprise assessors drawn from 17 participating accounting departments from Australian higher education providers, including public universities and other public and private providers. The *Achievement Matters: External Peer Review of Accounting Learning Standards* (hereafter *Achievement Matters*) project's primary objective has been first to understand the extent of the problem of grader variability, and then develop a sustainable model for assuring achievement of accounting threshold learning standards using external peer assessors. For a full description of the *Achievement Matters* project, see Watty et al. (2014). The threshold learning standards for the accounting discipline are a useful 'test case' as accounting is the first business discipline to develop its own threshold learning standards, with economics, marketing and finance subsequently following suit. To be effective, the model has to produce reliable and valid judgements of students' learning outcomes. Informed by the literature on consensus moderation, the project team decided that the way to gain greater consistency in assessors' judgements was to develop a rigorous calibration process for judgements of learning benchmarked to threshold learning standards. The calibration process is a multistage intervention in which external peer assessors examine both the assessment task and the samples of student work against the threshold learning standards before, during and after a calibration workshop. To date, eight calibration workshops have been administered – the results of one are reported in this study.

In our study, we seek to answer three key research questions:

(1) What is the quantifiable difference in grader variability on the assessment of learning outcomes in 'application skills' and 'judgement'?
(2) Does participation in the workshops lead to reduced disparity in the assessment of the students' learning outcomes in 'application skills' and 'judgement'?
(3) Does participation in the workshops lead to greater confidence by reviewers in their ability to assess students' skills in application skills and judgement?

The third research question is based on the key issue raised by Bloxham (2009), who stressed the importance of ensuring that all key stakeholders in higher education, including academics, students and employers, are able maintain their confidence in the reliability of marking.

The context of the learning and teaching environment

In the second decade of the twenty-first century, the higher education landscape in many countries is characterised by: increasing student-to-staff ratios; greater

international student mobility; and greater international competition for student enrolments and uncertain funding arrangements (Altbach, Reisberg, and Rumbley 2009; Gu and Schweisfurth 2011). The above exigencies, especially the increasingly competitive environment, have resulted in calls for greater transparency and accountability of educational systems and outcomes, including a concerted push for graduate learning outcomes (also known as competencies, attributes or capabilities) to meet standards of achievement, without resorting to standardised delivery or testing. These developments, together with concerns in some countries from industry and government about the capabilities of graduates, have led to an increased focus on assurance of learning for academic standards in general, including the integrity and robustness of higher education providers' assessment practices.

The above concerns are exacerbated by recent research on higher education assessment practices. For example, Bloxham and Boyd (2011) identify issues about grading in the UK, such as the dearth of external examination of students' work, the lack of uniformity of assessment standards across higher education providers and the presence of grade inflation. Furthermore, Bloxham and Boyd (2011, 1) contend that debate is largely absent on 'the sovereignty of professional judgement that embodies our sense of academic standards'.

In 2011, after a lengthy period in which the tertiary education sector in Australia had considerable autonomy in managing and maintaining academic standards, the Federal Government introduced a national regulator, the Tertiary Education Quality and Standards Agency (TEQSA). TEQSA is responsible for higher education provider quality assurance and enhancement, and for ensuring higher education providers meet appropriate teaching and learning standards, as exemplified in paragraphs 1.2 and 5.5 of the Provider Course Accreditation Standards contained in the *Higher Education Standards Framework (Threshold Standards) Act (2011)*:

> There are robust internal processes for design and approval of the course of study which take account of external standards and requirements, e.g. published discipline standards, professional accreditation, input from relevant external stakeholders, and comparable standards at other higher education providers.
>
> The academic standards intended to be achieved by students and the standards actually achieved by students in the course of study are benchmarked against similar accredited courses of study offered by other higher education providers.

While a Higher Education Standards Panel (HESP) will ultimately determine how the teaching and learning standards are to be revised for Australian higher education providers, neither TEQSA nor HESP will be developing learning standards at the discipline level. In anticipation, the accounting discipline has taken the initiative to develop five threshold learning standards (listed and defined in Table 1). For information on how these threshold learning outcomes were developed, see Freeman and Hancock (2011).

Although the threshold learning standards for accounting have been nationally endorsed, there is no shared understanding to date as to how to assess them fairly, consistently and cost effectively. A standard as defined by Sadler (2013, 13) is a 'definite degree of academic achievement established by authority, custom, or consensus and used as a fixed reference point for reporting a student's level of attainment'. Consistent with Sadler's definition of a standard, the learning outcome achieved by any student may vary, but the threshold learning standards should be a stable benchmark over time and across institutions. However, this ideal is

problematic because higher order learning standards, like application skills and judgement, are not directly measureable and open to individual interpretation. Attempts within Australia to assess threshold learning standards include an exercise by eight of the leading Australian research universities to verify each other's standards (Group of Eight 2011). Another exercise involves 12 diverse universities assessing sample learning outcomes for final year units of study (Deane and Krause 2012). Only one exercise has focused on benchmarking student achievement against agreed threshold learning standards: the *Achievement Matters* project that is the focus of this paper (Watty et al. 2014).

Review of literature on developing a shared understanding of academic standards

This section outlines research to date pointing to a critical problem for educators and students alike, namely, that the grading process of student work is neither consistent nor fair (Shaftel and Shaftel 2007; Price et al. 2008; Higher Education Academy and Quality Assurance Agency 2013). Much grading entails non-quantitative assessments of students' achievements against expected standards, based on markers' professional expertise and assumptions about how they should judge the work. Sadler (2013) notes that, regardless of which assessor assesses a student's work, the student is entitled to expect the mark to be comparable to that given by other assessors (within an acceptable tolerance range). In essence, a reliable grading process produces consistent assessment outcomes.

However, researchers have pointed to considerable variability between assessors' assessment criteria and grades assigned (Shaftel and Shaftel 2007; Price et al. 2008). The concern of consistency was foregrounded recently in a study commissioned in the UK by the Higher Education Academy with the cooperation of the Quality Assurance Agency (2013, 6). Their examination found that when six experienced external examiners from chemistry, history, psychology and nursing marked five pieces of borderline student work from a number of institutions (i.e. 4 disciplines × 5 pieces of student work = 20), there was little examiner consistency, specifically:

> Of the 20 assignments only one was assigned the same rank (highest or joint highest) by all six examiners and other assignments were given grades that 'ranked' them against the other assignments in at least three different positions (that is best, second best, and so on). Nine of the 20 assignments were ranked both best (or joint best) and worst (or joint worst) by different examiners. Even where the overall judgements about an assignment were similar, examiners frequently made different judgements about the strengths and weaknesses of particular aspects of the work, for example, the quality of argument.

This finding echoes the results of a summit of 40 experts (Price et al. 2008), and together these findings led to the UK Quality Assurance Agency (2013, 9) making the following recommendation:

> The Higher Education Academy should promote and facilitate disciplinary community processes beyond local practices to include inter-institutional disciplinary benchmarking processes for sharing, developing and assuring standards. Higher education institutions should actively support external examiners in participating in these processes.

Academics attach qualities to students' work according to their varying professional knowledge and experience (Read, Francis, and Robson 2005; Smith and Coombe

2006; Bloxham 2009). The grading process in higher education has not been investigated exhaustively; however, the following tentative findings are noted: grades differ in their distribution shapes (Heywood 2000); marker reliability is lower for essay and problem-style examinations, and conversely higher for knowledge that requires recall (Elton and Johnston 2002). It follows that the learning outcomes in some disciplines lend themselves more to reliable grading due to the object-like nature of the outcomes – for example, a calculation in the natural sciences compared with an essay in existential philosophy (Knight 2006). Given that accounting educators in the past two decades have increasingly acted on calls by employers to incorporate a greater focus on generic skills development into the curriculum (see AAA 1986; Cook et al. 2011), the problematic nature of assessing non-technical skills has come to prominence, as highlighted by Royce Sadler (2014). Moreover, as complexity of learning outcomes increases, reliance on 'connoisseurship' increases; assessors rely upon 'expert and reliable' judgements supported by their education and common sense understanding of standards within their relevant discipline (Knight and Yorke 2003).

To address the inherent problems of variability in assessment, efforts have been made to increase the reliability and validity of marking via the use of assessment criteria, grade descriptors and marking rubrics. However, Sadler (1987, 205) argues that 'fuzzy standards cannot be transformed into sharp standards by simply using more detailed language'. Furthermore, researchers have found that these efforts are often thwarted, because markers tend to develop fixed marking habits (Wolf 1995), and may not value the outcomes they are meant to be judging (Baume, Yorke, and Coffey 2004). In addition, assessors tend to rely on heuristics and ignore (or decide to not adopt) written criteria in rubrics (Price and Rust 1999; Ecclestone 2001; Smith and Coombe 2006; Sadler 2009). Indeed, as noted by Moss and Schutz (2001, 8), 'judgement becomes an "interpretation" of standards, not a "reflection" of them'. Another problem identified in the literature is that markers sometimes view generic institutional standards as lacking in rigour and, therefore, feel the need to apply higher standards to their specific modules (Price 2005). Other factors considered to contribute to the lack of reliability in marking are the propensity for assessors to act alone in marking, and (some) academics' lack of scholarship (Bloxham 2009).

Social constructivism

The literature cited above supports the proposition that variation in academic outcomes and assessment of standards is problematic. More recently, attention has been given to the important matter of epistemology (way of viewing knowledge) in the 'doing' of assessment. Bloxham and Boyd (2011) argue that there are two paradigms that apply to the assessment of learning outcomes: techno-rationalist and sociocultural. The techno-rationalist epistemology aligns with a positivist view that standards can be made explicit via a criterion-referenced approach, and thereby interpreted in an objective manner (Orr 2007; Bloxham 2009; Bloxham and Boyd 2011). On the other hand, the sociocultural paradigm takes the position that assessment is a 'context-dependent, socially situated, interpretive activity' (Bloxham and Boyd 2011, 3). Suffice to say that there has been little investigation into how academics arrive at their epistemology, but it has been assumed that learning about standards is linked with research practices and exchanges with colleagues via discussion, debate and moderation (Reimann et al. 2010). As noted by Bloxham and Price (2013, 10):

> The interpretive [sociocultural] perspective would tend to emphasise the notion of examiners learning, not through reference to documentation and explicit standards, but through being part of an assessment community.

These observations by Reimann et al. (2010) and Bloxham and Price (2013, forthcoming) lend support to the notion that assessment design and process are contextually and socioculturally constructed. This argument is consistent with Bloxham (2009), who asserts that moderation of student work is a socially constructed process that goes beyond the narrow concept that implies a resolution of differences. Her position on the best way to develop assessors' knowledge of the required assessment standard is summed up as follows (Bloxham 2009, 218):

> [it should be] created through a social process involving dialogue and experience and using artifacts such as assignment guidance and assessment criteria but, in essence, it remains essentially an individual construct, heavily influenced by traditions in the subject discipline.

Bloxham (2009) advocates 'calibration' principles that enable academics to develop sufficiently shared understandings of academic standards in order to make consistent grading decisions without continuous checking and recalibration. Sadler (2013, 3) makes a similar recommendation, using the term consensus moderation as the means to 'regularise' the assessment of academic standards:

> consensus moderation is generally accepted as both appropriate and necessary whenever multiple assessors are involved in marking student responses to a single assessment task. The proposed strategy is to regularize and refine the existing practice of consensus moderation, and extend its scope.

Sadler (2013) proposes that, in adopting a social constructivist approach, assessors take the following steps in the consensus moderation strategy:

- trial mark the same sample of student work;
- compare with each other provisionally allocated marks;
- engage in focused discussion about how marks should be allocated;
- reach agreement on an appropriate academic achievement standard; and
- mark the remaining student assessments independently.

Reflecting on Price et al.'s (2008) manifesto for change in assessment, Rust (2009) proposes that established discipline communities are best placed to organise and undertake calibration events involving external peers.

While we support the social constructivist approach to assessment and grading of non-quantitative learning outcomes, we wish to note the following challenges to its implementation:

(1) This approach relies on some academics changing the way they conceive of the assessment task (Biggs 2001). For example, assessment is often developed by academics in isolation, whereas the constructivist perspective sees assessment and learning as necessarily reciprocal and interdependent.
(2) Politics (i.e. power relationships) plays a part in moderation decision-making; therefore, a degree of egalitarianism is needed if junior academics are to be partners with senior academics (Bloxham and Boyd 2011).

(3) Use of external assessors may aid the process of consensus moderation considerably, but there is the risk that insistence on the involvement of external people may have a dampening effect on innovation in assessment practices at the home institution (Biggs 2001).

Additionally, it might seem the process would involve high costs in terms of travel and accommodation and time for those involved, but these costs can be minimised by arranging calibration workshops around major disciplinary events such as conferences and workshops. This was the approach adopted in the project reported in this paper. The academic time commitment is not significant as participants were only asked to assess five pieces of student work.

In summary, we acknowledge that marking is a complex and inherently problematic activity, where it cannot be declared definitively that a grade awarded by one academic would be the same as that awarded for the same work by another academic. Nevertheless, except for the assurance of validity, we are attracted to Bloxham's (2009, 212) rather optimistic claim that consensus moderation is 'a process for assuring that an assessment outcome is valid, fair and reliable and that marking criteria have been applied consistently'.

Method

Sample selection and trialling

Our call to Australia's 40 publicly funded higher education providers for participants in the pilot review and consensus moderation process yielded 17 volunteer institutions, including one technical and further education (TAFE) institution (a community college) and one private provider. Though this was not a random sample, the research team's view is that it represents a good cross-section of the contemporary Australian accounting higher education environment.

Treatment and control group

Consistent with the rigour of experimental design, we sought comparability of participants in both the treatment and control groups. The treatment group for this study consisted of 30 assessors from the participating institutions. The control group consisted of 15 assessors, invited from a range of Australian universities in terms of geographic location and global ranking. A review of Table 2 reveals that the composition of group members is comparable in terms of demographic characteristics such as gender and seniority; there were slightly more men in the treatment group (63%) than the control group (60%), and both groups possessed a similar proportion of junior and senior academics. It should be noted that the treatment group also included four assessors who were employed as instructors/lecturers at TAFE and private providers and two from a major professional accounting body. We tested for sensitivity of our results to this latter group by excluding them and rerunning the tests: since we found the results were not significantly different, we kept the six in the sample.

The study followed the traditions of experimental research design, with three stages as follows:

Table 2. Demographic details of workshop (treatment) and control groups.

Demographic criteria	Treatment group (n = 30)	Control group (n = 15)
Gender		
Male	19 (63%)	9 (60%)
Female	11 (37%)	6 (40%)
Total	30 (100%)	15 (100%)
Academic seniority		
Level A/B (junior)	7 (23%)	4 (26%)
Level C	5 (17%)	6 (40%)
Level D/E	12 (40%)	5 (34%)
Other*	6 (20%)	0 (0%)
Total	30 (100%)	15 (100%)

*Includes four lecturers from TAFE (community college)/private providers and two from a major professional accounting body.

(1) participants in both groups were required to undertake a pre-assessment task in isolation (pre-test);
(2) the treatment group received their treatment; and
(3) all participants completed a post-assessment task (post-test) that was identical to the pre-test.

The participants and review process

The 30 assessors in the treatment group had to meet two criteria. First, they had to be selected by their participating institution. Second, they had to be willing to participate in a pre-workshop assessment exercise involving sample assessment tasks.

The assessment task consisted of three pieces of de-identified student work that addressed the threshold learning standards of application skills and judgement (see Table 1). Each piece of work entailed the grading of non-quantitative assessments of students' achievements against expected standards. Data collection matched the intended purpose of the threshold learning standards, where graduate outcomes were measured at the end of a course of study. Accordingly, student work in this study was drawn from an accounting unit of study taken toward the end of the course. Each piece of work was a discursive piece of approximately 750 words, and the assessments were based on markers' professional knowledge and preconceived assumptions about how they should judge the work (Sadler 2013).

The review process was conducted in three distinct stages. In *Stage one,* the 30 assessors in the treatment group and the 15 in the control group assessed the three students' non-qualitative work, and then entered their judgements and comments to support their ratings in an online repository known as the *Self and Peer Assessment Resource Kit*^{PLUS} (*SPARK*^{PLUS}). *SPARK*^{PLUS} is a web-based tool that allows reviewers to input their assessments and comments online. Once all reviewers have entered their assessments, the results of other reviewers and comments can then be shared with all team members. The tool reports all comments and the range of assessments showing a dispersion around the mean mark.

In *Stage two*, the 30 assessors in the treatment group met in person and shared their reviews. In groups of four or five, participants discuss and defend judgements

and justifications around the validity of the assessment task until a consensus is reached within the group. This also provides a further opportunity for each reviewer to reflect on their initial review. Following this, the facilitator leads a discussion between groups where participants defend key differences until a consensus is reached across all groups. The existence of discipline standards assists these calibration/tuning conversations. Assuming the assessment task is valid for demonstrating achievement of the standard(s) in focus, reviewers then repeat this consensus-reaching process through open dialogue, benchmarking each piece of student work against the relevant standard. Reviewers repeat the consensus-reaching process with new samples of student work until there is confirmation that the calibration process has been effective, and there is agreement on the discipline-specific standard for a graduate. This approach was consistent with the principles of consensus moderation advocated by Sadler (2013). These face-to-face workshops varied between four and six hours in length.

In *Stage three*, after the 'treatment', both the treatment and control groups reassessed the same student work they had assessed in stage one. This reassessment was undertaken within two weeks following the workshop. The reviewers were asked to reassess the same three samples of student work as if this was the first time, as all marks and comments from the first assessments were removed from $SPARK^{PLUS}$. This post-workshop re-examination of assessors' grades effectively constitutes a test of the impact of the workshops on assessors' judgements of the threshold learning standards. The treatment group's results were then compared to the control group's results.

Data integrity and confidentiality

Reliability and validity requires data integrity and confidentiality, as well as reviewer anonymity. The graded student sample work and assessment requirements used for the calibration workshop were provided by one of the project teams' institutions and were de-identified before distribution to all 45 assessors.

Results and analysis

Reviewer variability

Tables 3 and 4 show descriptive statistics (mean scores, range and standard deviations) for both the treatment and control groups in addressing the first two research questions.

Table 3 shows that the mean scores for the treatment group (possible range of 0–100) for the application skills standard across the three student assessments were approximately 66, 69 and 59, respectively. The post-workshop mean scores for this group were 62, 61 and 62, respectively. This table also shows that the mean scores for the control group for the three student assessments were approximately 77, 65 and 69, respectively (pre-workshop) and 73, 59 and 63, respectively (post-workshop).

Test results as measured by standard deviations are of particular interest for the present study, given that the aim of the process was to understand better grader variability by quantifying the problem, and then, via workshops, attempt to control the problem. In Table 4, we observe the size of the problem where grader variability varies considerably, with standard deviations from the mean pre-workshop ranging from 14 to 21 for both groups. Having quantified grader variability, controlling the problem through intervention is measured by the reduction in variability in

Table 3. Comparisons of assessors' scores on the application skills standard for each assessment pre- and post-workshop for the treatment and control groups.

Application standard	Skills					
	Student assessment 1		Student assessment 2		Student assessment 3	
Statistic	Pre-wshop score	Post-wshop score	Pre-wshop score	Post-wshop score	Pre-wshop score	Post-wshop score
Mean						
Treatment	65.97	61.63	69.40	60.93	59.37	62.27
Control	76.80	73.27	65.33	59.00	69.00	63.07
Minimum						
Treatment	40	44	47	45	24	47
Control	53	50	24	24	17	22
Maximum						
Treatment	87	78	94	73	94	84
Control	100	94	94	85	97	91
Std Dev.						
Treatment	14.86	8.07	14.35	6.76	18.03	10.30
Control	14.92	13.48	20.83	17.46	21.57	19.38

Table 4. Comparisons of assessors' scores on the judgement standard for each assessment pre- and post-workshop for the treatment and control groups.

Judgement standard	Skills					
	Student assessment 1		Student assessment 2		Student assessment 3	
Statistic	Pre-wshop score	Post-wshop score	Pre-wshop score	Post-wshop score	Pre-wshop score	Post-wshop score
Mean						
Treatment	61.13	62.13	63.50	60.63	57.63	60.47
Control	73.67	69.07	69.40	64.40	67.93	60.40
Minimum						
Treatment	39	46	37	47	21	43
Control	54	51	37	37	44	38
Maximum						
Treatment	89	81	92	78	96	83
Control	100	88	94	93	97	85
Std Dev.						
Treatment	14.43	7.94	13.57	7.04	20.09	9.48
Control	13.57	11.35	15.41	16.12	14.84	12.25

assessors' scores for each piece of work following assessors' participation in the workshop. The change in standard deviation from pre-workshop to post-workshop for the treatment group was significant for all three post-workshop assessed tasks. The standard deviations fell (pre- to post-workshop) from 14.86 to 8.07, 14.35 to 6.76 and 18.03 to 10.30 for students one to three, respectively. The change for the control group was much smaller: 14.92 to 13.48, 20.83 to 17.46 and 21.57 to 19.38, respectively. We applied both parametric (F test) and non-parametric (Kendall's *Tau*)

tests to examine the statistical significance of these group changes, and while both tests supported this interpretation, it remains that the small sample size could question the reliability of the tests.

Additionally, the gap between the minimum and maximum ratings fell across the two periods for the treatment group, declining from 47 to 34, 47 to 28 and 70 to 37 for each of the three assessments, respectively. Overall, these findings indicate that the workshops were effective in significantly reducing variability across different assessments of the application skills standard.

Table 4 shows results for the judgement standard. There is a wide range in grader variability with standard deviations pre-workshop ranging from 13 to 20 for both groups. Having quantified grader variability, controlling the problem through intervention is measured by the reduction in variability of scores of assessors for each piece of work following assessors' participation in the workshop. Again, the changes from pre-workshop to post-workshop standard deviations are significant for all three post-workshop assessed tasks. The standard deviations fell (pre- to post-workshop) from 14.43 to 7.94, 13.57 to 7.04 and 20.09 to 9.48 for students one to three, respectively. This compared to a much smaller movement with the control group: 13.57 to 11.35, 15.41 to 16.12 and 14.84 to 12.25, respectively.

Table 5. Means scores for reviewers' confidence in the consensus moderation process 1 = strongly disagree; 5 = strongly agree.

Questions	Mean score prior to process commencing	Mean score following consensus moderation
I am confident rating the capacity of assessment requirements to allow students to demonstrate the national threshold learning outcome for judgement	3.4	4.1
I am confident rating the capacity of assessment requirements to allow students to demonstrate the national threshold learning outcome for application	3.6	4.3
I am confident rating students' judgement ability benchmarked against the national standard	3.3	4.0
I am confident rating students' application ability benchmarked against the national standard	3.5	4.2
I am confident that the my feedback, explaining my ratings and offering suggestions, will be useful to the assessor	3.6	4.3
The pre-workshop activity, requiring me to reflect on the students' work in the context of the agreed national academic standards, changed my understanding of academic standards for judgement that might apply locally	3.0	4.2
The pre-workshop activity, requiring me to reflect on the students' work in the context of the agreed national academic standards, changed my understanding of academic standards for application that might apply locally	3.1	4.1

Overall, these findings provide valuable data on the degree of grader variability (research question 1) and the effectiveness of the workshops in significantly reducing variability (research question 2) across assessors' evaluations of the application skills and judgement standards. When compared to the control group, the treatment group attained a significant reduction in reviewer variation.

Reviewer confidence in the consensus moderation process

Prior to attending each workshop, the treatment assessors ($n = 30$) responded to seven questions to ascertain their level of confidence in the consensus moderation process. They reanswered these questions after the workshop. Table 5 reports a significant rise in assessors' confidence in the process. For example, for the question 'I am confident rating the capacity of assessment requirements to allow students to demonstrate the national threshold learning outcome for application skills and judgement', the level of agreement increased from 3.4 to 4.1 for application skills and from 3.6 to 4.3 for judgement. Similar increases can be seen for the other questions. These results indicate that the moderation process reduced grader variability and improved confidence in assessors' ability to conduct grading of the threshold learning standards. This finding addresses research question 3: does participation in the workshops lead to greater confidence by the assessors in their ability to evaluate the application skills and judgement of students?

Summary and conclusions

We concur with researchers, such as Bloxham and Price (2013), that reliable, consistent assessment processes that produce comparable assessment grades between assessors and institutions are an ongoing challenge. Our study can be differentiated from its predecessors as we measure the degree of grader variability or the size of the problem. In practical terms, large standard deviations suggest that students from different institutions, who have their work assessed by different assessors, will not be awarded comparable achievement grades for comparable work. The likely implication of such inequality in grading is that employers and other stakeholders cannot rely on these as reliable assessments. In determining potential new recruits, prospective employers' often use grades as the first step in selecting candidates to interview. Moreover, scholarship committees often use similar processes.

Having demonstrated the size and nature of the problem, it follows that, if assessment variation can be reduced, students and other stakeholders can more confidently rely on assessors' judgements of students' capabilities. As proposed by Rust (2009), Sadler (2013) and Bloxham (2009), one solution to this issue is to develop processes that promote a shared understanding of how outcomes should be judged, and thereby reduce the incidence of assessment variation among assessors.

In this study, we addressed a gap in the literature in terms of the paucity of prior research into the assessment moderation process itself, and especially its level of effectiveness (Orr 2007; Yorke 2008). We are not aware of any study which looks at the reduction in grader variability as done in this study. We report results from the initiative to implement threshold learning standards in the accounting discipline in Australia, but also contend that our findings may have significance beyond the research setting. The findings of our study, in which 'treatment' resulted in significantly reduced assessor variability and increased confidence by assessors in their

ability to assess judgement and application skills of students, demonstrate the potential for investing in developing a shared understanding of standards through a workshop calibration process.

Once we quantified the problem, the calibration workshops resulted in a halving of the size of the problem. While a 50% reduction in variability indicates a large problem still remains, we contend that the process described in this paper delivers a significant improvement in the problem of grader variability. We further argue that the assessment calibration exercise is in its formative stages and, as we continue with additional cycles of calibration, assessor variability should further reduce. As Table 5 shows, academics improve in their confidence and ability to critically review with each cycle of calibration. If this confidence continues to develop over time, the variability between assessors will reduce further as assessors become more experienced and accustomed to the calibration process. Bloxham (2009) stresses the importance of ensuring that all key stakeholders in the higher education process, including academics, students and employers, are able to maintain their confidence in the reliability of marking. The evidence of enhanced confidence by assessors in applying the standards following the workshop is therefore welcome.

The findings of our study add quantitative support to the advocates of a consensus moderation process (e.g. Bloxham and Price 2013). Underpinning our study was recognition of the complexity and inherent difficulty of assessment of threshold learning standards, especially, in this case, on judgement and application skills. Assessing a national standard such as judgement without moderation would be open to diverse interpretations. We have demonstrated that access to a collaborative workshop can overcome major divergences in assessment. However, we do not propose this as a technical solution to the problem of variability. In recommending calibration or consensus moderation we concur with advocates of the sociocultural/constructivist paradigm, who argue that assessment is always an interactive task that is necessarily interpretive. Assessment of non-quantitative outcomes especially will always be open to interpretation. But a solution to this conundrum is to avoid trying to objectify learning outcomes or to reduce their assessment to a narrow measurement. Rather, the solution lies with realising that both subjective and objective knowledge are involved in this very human enterprise, and that through facilitation and collaboration this volatility can be significantly reduced.

While it might seem that the process outlined in this paper would involve high costs in terms of travel and accommodation and time for those involved, these costs can be minimised. Furthermore, participants have indicated in various surveys that they value the professional development associated with the project, and this was demonstrated by all 35 participants voting to continue with the two workshops each year beyond the life of the funded project.

We believe the principle of calibration is already inherent to some degree in academic practices. For example, the much maligned committee meeting can be an effective vehicle for reaching academic decisions. And, to a lesser extent, calibration occurs in the peer reviewing of journal publications and examination of theses, albeit through the intervention of editors or committee chairs to resolve disparities in assessments. Needless to say, in the case of the latter, the calibration principle may be applied in a more concerted manner to good effect, which is surely an exciting possibility for future research.

There are two fairly obvious limitations to our study. First, though our experiment with treatment and control groups and associated pre- and post-tests is robust,

we acknowledge that the size of our groups is relatively small. Second, the moderation of assessment is such a complex set of activities; it is uncertain whether a workshop approach involving robust debate among participants, as described in this paper, would be viable in all education environments. Therefore, we strongly recommend that academics worldwide consider adopting cross-institutional assessor calibration exercises in an effort to tackle the problem of grader variability on a larger scale.

Acknowledgements

The authors would like to thank the Australian Business Deans' Council, the Australian Government Office for Teaching and Learning, CPA Australia and the Institute of Chartered Accountants in Australia for their financial support for this project. We also appreciate the cooperation of the numerous accounting programmes that have participated in this national initiative. We would also like to thank delegates and discussants at the European Accounting Association Annual Meeting in Paris in 2013, the International Association for Accounting Education and Research Conference (IAAER) in Amsterdam in 2012 and the Accounting and Finance Association of Australia and New Zealand (AFAANZ) Annual Meeting in Perth in 2013 for their helpful input into earlier drafts of this paper.

Notes on contributors

Brendan O'Connell is a professor in the School of Accounting at RMIT University. His research is in the areas of accounting education, corporate governance and accounting scandals. He is currently leading a study into the future of accounting education in Australian universities.

Paul De Lange is currently a professor and dean in Curtin Business School at Curtin University. He has published over 45 refereed publications and is a regular speaker at numerous international and national conferences. His research interests are in Accounting Education, Theory and the Accounting Profession. He is a CPA and since 2006, he has been an elected member of the AFAANZ Board of Directors and is the president for 2013–2015.

Mark Freeman is an associate professor at the University of Sydney and director accreditation at the University of Sydney Business School. His key research interests are in assessment and standards, educational technology and accounting education. His published research work can be found in Assessment and Evaluation in Higher Education, Journal of Higher Education Policy and Management, British Journal of Educational Technology, Australasian Journal of Educational Technology, International Journal of Management Education, Asian Social Science, Accounting Education: An International Journal, Accounting Journal of Education.

Phil Hancock is a professor of accounting and associate dean (Education) in the Business School at the University of Western Australia. His research work focuses on Accounting Education with an emphasis on accounting learning standards, Financial Reporting in both the For-profit and Not-for-profit sectors, and Corporate Governance. Phil is a fellow of the Accounting and Finance Association of Australia and New Zealand, Chartered Accountants Australia and New Zealand and CPA Australia.

Anne Abraham is an associate head of School (Engaged Research) in the School of Accounting at the University of Western Sydney. Her outstanding contribution to teaching and learning has been recognised by five teaching awards at international, national and university levels, and ongoing consultancies with educational publishers. Her major research interests are in the areas of accounting education, non-profit financial management and accountability, and social and environmental sustainability.

Bryan Howieson is an associate professor in the School of Accounting and Finance within the Business School at the University of Adelaide. His research interests include financial reporting, accounting theory, professional ethics and accounting education. Bryan is a fellow of the Accounting and Finance Association of Australia and New Zealand and CPA Australia.

Kim Watty is a professor of Accounting and associate dean, Quality Standards and Accreditation in the Faculty of Business and Law at Deakin University Melbourne, Australia. Her research area is accounting education in higher education and focuses on assessment, employability skills, quality and standards and more recently, technology and eportfolios. She has published widely in national and international refereed journals and has successfully led to completion several competitively funded national and international research projects focused on accounting education.

References

AAA (American Accounting Association). 1986. "The Committee on the Future Structure, Content and Scope of Accounting Education [The Bedford Committee], Future Accounting Education: Preparing for an Expanding Profession." *Issues in Accounting Education* 1 (1): 168–195.

AACSB (Association to Advance Collegiate Schools of Business). 2014a. "Accredited Institutions." Accessed July 10, 2014. http://www.aacsb.edu/accreditation/accreditedmembers.asp

AACSB (Association to Advance Collegiate Schools of Business). 2014b. "AACSB Assurance of Learning Standards: An interpretation, White Paper." Accessed April 5, 2014. http://www.aacsb.edu/en/publications/whitepapers/

Altbach, P., L. Reisberg, and L. Rumbley. 2009. *Trends in Global Higher Education: Tracking an Academic Revolution*. Paris: UNESCO.

American Association of Colleges and Universities. 2011. "AAC&U Statement on the Lumina Foundation for Education's Proposed Degree Qualifications Profile". Accessed July 25, 2014. http://www.aacu.org/about/statements/documents/lumina_dqs_2011.pdf

Apostolou, B. A. 1999. "Outcomes Assessment." *Issues in Accounting Education* 14 (1): 177–197.

Baker, R., F. A. Bayer, A. Gabbin, D. Izard, F. Jacobs, and S. Polejewski. 1994. "Summary of 'outcomes assessment'." *Journal of Accounting Education* 12 (2): 105–135.

Baume, D., M. Yorke, and M. Coffey. 2004. "What is Happening When we Assess, and How Can we Use our Understanding of this to Improve Assessment?" *Assessment & Evaluation in Higher Education* 29 (4): 451–477.

Biggs, J. 2001. "The Reflective Institution: Assuring and Enhancing the Quality of Teaching and Learning." *Higher Education* 41 (3): 221–238.

Bloxham, S. 2009. "Marking and Moderation in the UK: False Assumptions and Wasted Resources." *Assessment & Evaluation in Higher Education* 34 (2): 209–220.

Bloxham, S., and P. Boyd. 2011. "Accountability in Grading Student Work: Securing Academic Standards in a Twenty-first Century Quality Assurance Context." *British Educational Research Journal* 38 (4): 1–20.

Bloxham, S., and M. Price. 2013. "External Examining: Fit for Purpose?" *Studies in Higher Education* 1 (13): 1–17. doi:10.1080/03075079.2013.823931.

Bloxham, S., and M. Price. Forthcoming. "External Peer Review of Assessment: An Effective Approach to Verifying Standards." *Higher Education Research and Development*.

Cook, G. L., D. Bay, B. Visser, J. E. Myburgh, and J. Njoroge. 2011. "Emotional Intelligence: The Role of Accounting Education and Work Experience." *Issues in Accounting Education* 26 (2): 267–286.

Deane, L., and K. Krause. 2012. "Towards a Learning Standards Framework." Accessed September 6, 2014. www.uws.edu.au/__data/assets/pdf_file/0010/398620/Learning_Stds_Framewk_Final_Dec_2012.pdf

Ecclestone, K. 2001. "I Know a 2:1 When I See It: Understanding Criteria for Degree Classifications in Franchised University Programmes." *Journal of Further and Higher Education* 25 (3): 301–313.

Elander, J., and D. Hardman. 2002. "An Application of Judgment Analysis to Examination Marking in Psychology." *British Journal of Psychology* 93: 303–328.
Elton, L., and B. Johnston. 2002. "Assessment in Universities: A Critical Review of Research." Accessed September 10, 2014. http://eprints.soton.ac.uk/59244/1/59244.pdf
Freeman, M., and P. Hancock. 2011. "A Brave New World: Australian Learning Outcomes in Accounting Education." *Accounting Education* 20 (3): 265–275.
Group of Eight Australian Universities. 2011. "Go8 Pilots a New Quality Verification System." GO8 Update. Accessed July 6, 2014. http://www.go8.edu.au/__documents/newsletters/2011/go8_newsletter_july2011.pdf
Gu, Q., and M. Schweisfurth. 2011. "Rethinking University Internationalisation: Towards Transformative Change." *Teachers and Teaching* 17 (6): 611–617.
Hancock, P., and M. A. Freeman. 2010. *Learning and Teaching Academic Standards Statement for Accounting*. Learning and Teaching Academic Standards Project. Canberra: Australian Learning and Teaching Council.
Heywood, J. 2000. *Assessment in Higher Education*. London: Jessica Kingsley.
Higher Education Academy and Quality Assurance Agency. 2013. External Examiners' Understanding and use of Academic Standards. http://www.qaa.ac.uk/publications/information-and-guidance/publication?PubID=2686#.VM9cpPSmiSo.
Knight, P. 2006. "The Local Practices of Assessment." *Assessment & Evaluation in Higher Education* 31 (4): 435–452.
Knight, P., and M. Yorke. 2003. "Assessment, Learning and Employability." *Learning and Teaching in Higher Education* 1. Accessed July 6, 2014. http://www.york.ac.uk/media/staffhome/learningandteaching/documents/keyfactors/Conditions_under_which_assessment_support_students'_learning.pdf#page=124
Moss, P. A., and A. Schutz. 2001. "Educational Standards, Assessment and the Search for Consensus." *American Educational Research Journal* 38 (1): 37–70.
Orr, S. 2007. "Assessment Moderation: Constructing the Marks and Constructing the Students." *Assessment & Evaluation in Higher Education* 32 (6): 645–656.
Price, M. 2005. "Assessment Standards: The Role of Communities of Practice and the Scholarship of Assessment." *Assessment & Evaluation in Higher Education* 30 (3): 215–230.
Price, M., B. O'Donovan, C. Rust, and J. Carroll. 2008. "Assessment Standards: A Manifesto for Change." *Brookes E-Journal of Learning and Teaching* 2 (3). Accessed January 10, 2014. http://bejlt.brookes.ac.uk/paper/assessment_standards_a_manifesto_for_change-2/
Price, M., and C. Rust. 1999. "The Experience of Introducing a Common Criteria Assessment Grid Across an Academic Department." *Quality in Higher Education* 5 (2): 133–144.
Read, B., B. Francis, and J. Robson. 2005. "Gender, 'Bias', Assessment and Feedback: Analyzing the Written Assessment of Undergraduate History Essays." *Assessment and Evaluation in Higher Education* 30 (3): 241–260.
Reimann, N., K. Harman, A. Wilson, and L. McDowell. 2010. "Learning to Assess in Higher Education: A Collaborative Exploration of the Interplay of Formal and Informal Learning in the Academic Workplace." Paper presented at the Higher Education Close Up Conference, Lancaster, July 20–22.
Royce Sadler, D. 2014. "The Futility of Attempting to Codify Academic Achievement Standards." *Higher Education* 67 (3): 273–288.
Rust, C. 2009. "Assessment Standards: A Potential Role for Subject Networks." *Journal of Hospitality, Leisure, Sport & Tourism Education* 8 (1): 124–128.
Sadler, D. R. 1987. "Specifying and Promulgating Achievement Standards." *Oxford Review of Education* 13 (2): 191–209.
Sadler, D. R. 2009. "Indeterminacy in the Use of Preset Criteria for Assessment and Grading." *Assessment & Evaluation in Higher Education* 34 (2): 159–179.
Sadler, D. R. 2011. "Academic Freedom, Achievement Standards and Professional Identity." *Quality in Higher Education* 17 (1): 85–100.
Sadler, D. R. 2013. "Assuring Academic Achievement Standards: From Moderation to Calibration." *Assessment in Education: Principles, Policy & Practice* 20 (1): 5–19.
Shaftel, J., and T. L. Shaftel. 2007. "Educational Assessment and the AACSB." *Issues in Accounting Education* 22 (2): 215–232.

Smith, E., and K. Coombe. 2006. "Quality and Qualms in the Marking of University Assignments by Sessional Staff: An Exploratory Study." *Higher Education* 51 (1): 45–69.

TEQSA (Tertiary Education Quality and Assurance Agency). 2012. "About TEQSA: The Role and Functions of TEQSA." Accessed September 10, 2014. http://www.teqsa.gov.au/about

Tuning Educational Structures in Europe Project. 2014. "An Introduction to Tuning Educational Structures in Europe: Universities' contribution to the Bologna Process." *European Commission through the Socrates and Tempus programmes (of the Directorate Education and Culture)*. Accessed October 21, 2013. http://www.ihep.org/assets/files/gcfp-files/Tuning_Educational_Structures_in_Europe_Universities__contribution.pdf

UK Higher Education Academy. 2013. "External Examiners' Understanding and use of Academic Standards." Accessed December 2, 2014. http://www.qaa.ac.uk/publications/information-and-guidance/publication/?PubID=2686

Watty, K., M. Freeman, P. Hancock, B. Howieson, B. O'Connell, P. De Lange, and A. Abraham. 2014. "Social Moderation, Assessment and Assuring Standards for Accounting Graduates." *Assessment & Evaluation in Higher Education* 19 (4): 461–478.

Wolf, A. 1995. *Competence Based Assessment*. Buckingham: Open University Press.

Yorke, M. 2008. *Grading Student Achievement in Higher Education*. Abingdon: Routledge.