Calibration Standards: What, Why and How?

Introduction

Calibration within the higher education sector is an approach that aims to ensure consistent standards for judging the quality of student work. A 'calibrated' academic is able to make grading judgments consistent with those of calibrated academics in other institutions across the UK. The aim of calibration is to achieve comparability of academic standards across institutions and stability of standards over time.

This concise report provides senior staff and policy makers, working in or with the higher education sector, with a summary of Advance HE's Degree Standards Project work on academic calibration over the last two years. It provides a brief rationale for calibration of external examiners, a description of the range of calibration activities undertaken, a synopsis of the key learning that has emerged from the project and recommendations for taking the work forward. The report recommends that the UK takes the next steps forward in developing a sustainable system of formal calibration for external examiners.

The case for calibration

The rationale for calibration has been building. It has arrived in response to pressures to safeguard academic standards across an expanding, diverse, marketized, higher education system both in the UK and internationally. In 2007 the Quality Assurance Agency responsible for UK higher education concluded that:

> *'It cannot be assumed students graduating with the same classified degree from different institutions, having studied the same subject, will have achieved similar standards'*
>
> *QAA 2007*[1]

In the UK, calibration has its roots in the formal requirement and public expectation of comparability of awards at least at the level of threshold academic standards[2] combined with a view that a diverse system should not mean a dilution of standards – minimum standards need to be maintained[3]. To this end, the sector has witnessed 30 years of determined efforts to assure the standards of UK higher education. This has included substantial texts designed to describe standards for the sector such as Subject Benchmark Statements and the Framework for Higher Education Qualifications (FHEQ) at the national level and programme specifications, learning outcomes and assessment criteria at the local level. Such work is still continuing with efforts to

---

[1] Quality Assurance Agency for Higher Education (2006b) Backround Briefing Note: The classification of degree awards Gloucester: QAA.

[2] Higher Education Academy, 2015. *A review of external examining arrangements across the UK.* Available at: https://www.heacademy.ac.uk/project-section/review-external-examining-arrangements

[3] Brown, R (2014), *Comparability of degree standards?* HEPI https://www.hepi.ac.uk/wp-content/uploads/2014/02/47-Comparability-of-degree-standards-summary.pdf

describe first class, upper second and other performances, partially in response to the charge of grade inflation[4] and anxieties about lack of comparability of standards.

Within our quality assurance system external examiners have a unique role as typically the only external check on academic standards as demonstrated through student achievements in exams, coursework and performances. External examiners are therefore seen as essential in helping secure both threshold standards and a level of comparability of standards across programmes and higher education providers (HEPs). In order for external examiners to advise the HEPs to which they are appointed, they need to have an appropriate and consistent sense of standards in line with documented national expectations.

However, although the sector has produced vast quantities of agreed written standards, achieving a shared interpretation of their meaning is a very different matter. Repeated studies over many years demonstrate considerable inconsistency in academics' judgements about student performance and variation in the meaning they accord to written standards. Studies of external examiners have found similar inconsistency[5]. This is not a criticism of examiners but a recognition that the language of standards always needs a level of interpretation and individuals differ in the meaning they accord to them. Such fluidity in standards leaves the sector open to charges of grade inflation as institutions reference sector norms (for example, proportion of firsts) rather than agreed national standards. Indeed, over twenty years ago, the forerunner of the QAA both noted this problem and suggested the makings of the possible solution:

> *'Consistent assessment decisions among assessors are the product of interactions over time, the internalisation of exemplars, and of inclusive networks. Written instructions, mark schemes and criteria, even when used with scrupulous care, cannot substitute for these.'*
>
> *Higher Education Quality Council (HEQC), 1997*[6]

The HEQC was describing informal calibration; a process where academic standards were sustained over time by an oral tradition through contact between Universities and subject communities. In the 1990s, this approach was considered no longer 'sufficient as a basis for standards in a mass system of higher education'[7]. And, whilst documented standards have been an important step in replacing this informal calibration, there is nothing formal in our QA systems which provides for the interaction, exemplars and networks needed to complement written standards. Thus, there is a need for organised calibration.

Calibration is important. The inability to safeguard appropriate and comparable standards has implications for the reputation of UK HE. We can no longer reasonably ask students, employers, parents and others to trust that our standards are correct because we don't have the evidence that they are.

---

[4] THE (28th Nov. 2018) 'The tripling of fees could be reason for rise in firsts'. THE p6-7

[5] Quality Assurance Agency, Higher Education Academy (2013) *External examiners understanding and use of standards*. https://www.heacademy.ac.uk/system/files/downloads/external-examiners-report.pdf

[6] Higher Education Quality Council (1997a) Assessment in higher education and the role of 'graduateness' London: HEQC. paragraph 4.7

[7] Brennan, J. 1996, "Introduction: The standards debate" in Changing Conceptions of Academic Standards, ed. J. Brennan, Quality Support Centre, Open University, London, pp. 9-26.

> *The public purse supports higher education to the tune of £15 billion and it is essential those studying at higher education institutions are awarded degrees that measure accurately and consistently the intellectual development and skills that students have achieved.*
>
> *Select Committee, 2009 p147*[8]

It also has implications for fairness to students if similar outcomes represent different levels of achievement. Assumptions of different standards at different HEPs undermine widening participation initiatives when high grades at some institutions are not believed, and hence valued, for example by employers. We already have evidence of Professional Statutory and Regulatory Bodies (PSRBs) no longer fully trusting degree outcomes and changing entry requirements to their profession accordingly. So far, higher education has rightly resisted national curricula and common examinations but pressure for these might grow unless we can more clearly demonstrate how we safeguard standards across multiple, autonomous HEPs.

What is calibration?

Calibration is a process of peer review carried out by members of a disciplinary and/or professional community who discuss, review and compare student work in order to reach a shared understanding of the academic standard which such work needs to meet. For example, these social moderation processes can involve sharing and agreeing examples of student work which meet the standard for 1st class, upper second, and so on. National and international projects and research[9] experimenting with calibration methods in conjunction with documented standards have served as a starting point for the Degree Standards Project. These experiments have been shown to help individuals and institutions to calibrate and demonstrate appropriate standards.

An Australian researcher, Royce Sadler, has been involved in sustained work on identifying and safeguarding standards including the importance of academics calibrating their standards within subject disciplines. He describes a calibrated academic as someone

> *'able to make grading judgements consistent with those which similarly calibrated colleagues would make, but without constant engagement in moderation. The overall aims are to achieve comparability of standards across institutions and stability of standards over time'* [10]

Central to Sadler's model is the idea that an academic standard cannot be determined simply by a written description but requires examples (for example, coursework or exam scripts) combined with dialogue leading to a description of why the examples meets the relevant standard. These can be produced by groups of subject specialists and then used in processes of social moderation at individual, programme, regional or national level.

The evidence from attempts at calibration in higher education looks promising. Prior calibration interventions have generated positive responses from academics who have found them to be worthwhile with some evidence that they have helped them to calibrate their standards. The

---

8 Select Committee, The Innovation, Universities, Science & Skills Committee *Students and Universities* https://publications.parliament.uk/pa/cm200809/cmselect/cmdius/170/170i.pdf

9 (Accounting) Watty, K., et al., 2013. Social moderation, assessment and assuring standards for accounting graduates. *Assessment and Evaluation in Higher Education*, 39(4), pp.461–478. (Law) Hanlon, J., Jefferson, M., Molan, M. & Mitchell, B (2004) *An examination of the incidence of error variation in the grading of law assessments*. Warwick: UKCLE, http://www.letr.org.uk/references/item/2047.html (Hospitality, leisure, sport & tourism) HSLT Subject Centre

10 Sadler, D. R., 2012. Assuring academic achievement standards: from moderation to calibration. *Assessment in Education: Principles, Policy and Practice*, 20(1), pp.5–19.

calibration strand of the Degree Standards project is committed to generating additional evidence needed to underpin calibration in UK higher education and develop its potential for contrasting disciplines. This is because current evidence is limited to a few projects with relatively low numbers of participants from a small range of disciplines. It is not known yet whether the enhanced grasp of standards achieved by participating academics has influenced their own assessment practice, how lasting the effect has been and whether it has made a difference to the standards used in their local contexts and the wider disciplinary community. The calibration strand of the Degree Standards project is committed to generating additional evidence needed to underpin calibration in UK higher education and develop its potential for contrasting disciplines.

What we've done

Calibration activities undertaken by the Degree Standards Project to date have followed one of two formats:

a) A free-standing calibration event
b) Calibration activities incorporated into the Professional Development Course for external examiners.

As part of these formats, the project has developed a range of calibration approaches which broadly use a social moderation process involving groups of subject academics from multiple HEPs. Three main approaches have been used:

Discussion of student work using documented standards

- The Australian 'Assessment Matters[11]' model where participants individually assess a set of student work, anonymously share their judgements and then use small group and large group discussion to agree what standard each piece has achieved. This involves use of relevant reference points such as subject benchmarks. The process is designed to provide feedback on individuals' standards and develop a shared view of appropriate standards for a range of exemplars.

Discussion of student work to agree on shared criteria

- Debate of exemplar student work has been used to agree shared criteria for judging standards including what characteristics should be used for determining judgements at key classification boundaries (e.g. pass/ fail, 2.ii/2.i and 1st/2.i.)

Independent assessment of student work from different institutions

- Institutional feedback on marking standards (peer review model[12]) where subject academics have anonymously reviewed student assessments from other institutions. This differs from external examining as there are multiple markers for each piece (3-4) and the assessors have no information about the original grade awarded or which institution provided them. In this way, academics gain feedback on whether colleagues from other institutions confirm or challenge the standards as expressed in marked student work. We have used this approach to particularly examine standards at the borderline of 2.ii/2.i and 2i/1st.

[11] Watty, K., et al., 2013. Social moderation, assessment and assuring standards for accounting graduates. *Assessment and Evaluation in Higher Education*, 39(4), pp.461–478.

[12] Krause, K.,et al(2013). Assuring final year subject and program achievement standards through inter-university peer review and moderation. Available online: www.uws.edu.au/latstandards.
http://www.uws.edu.au/__data/assets/pdf_file/0007/576916/External_Report_2014_Web_3.pdf

The exact nature of each calibration activity has been influenced by the subject discipline, disciplinary standards, interests of collaborating PSRBs and the assessment methods of different subjects. Some events have been very specific in their focus, for example recital performance standards in music or reflective writing in veterinary education. Other initiatives have looked more broadly at final year standards, for example in law and geography.

Overall, the calibration activities undertaken to date has shown that participation in calibration activities impacts on standards in the sector at several different levels:

> *National* – Developing an in-depth understanding and agreeing the meaning of national standards (FHEQ, Subject Benchmark Statements, statements of first class, etc) to improve consistency in their interpretation in use. Creating graded exemplars for use within disciplines.
>
> *Institutional / department* – Providing feedback on local standards from multiple colleagues across mission groups. Enabling departments to consider whether their standards are in line with sector **standards** rather than sector **norms** of degree classification.
>
> *Individual* - Providing feedback on how individuals' standards compare with those of others and, where necessary, helping recalibrate their judgements.

Who we have worked with

The project has consciously sought a diverse range of subject and professional fields which are taught in both mainstream and small, specialist HEPs. We have carried out the work in collaboration with relevant PSRBs and learned societies or through a regional consortium of universities. Our partners and subjects have been:

Geography: Royal Geographical Society (with Institute of British Geographers) and Consortium of NW Universities

Music: Conservatoires UK and Royal Northern College of Music

Law: Consortium of NW Universities

Chemistry: Royal Society of Chemistry

Veterinary education: Royal College of Veterinary Surgeons

We have found all these groups to be enthusiastic partners in developing calibration.

What we've learned

The calibration work undertaken so far is relatively small scale and at an early stage in the development of calibration in UK higher education. However, various patterns have been revealed which provide useful indicators for further development of the work.

Participant response

- Firstly, it is important to say that, in common with previous calibration initiatives, we found a strong positive reaction from participants across the subject disciplines and PSRBs.
- Calibration has proved an excellent means for intermixing and debate between representatives of the different university mission groups on the topic of assessment and

standards and we have found no sense of a hierarchy of institutions in terms of atmosphere. We do not think this cross-sector group discussion of standards happens in any other forum as external examiners are typically appointed from the same mission group[13].

- Whilst our focus has been external examiners and most participants were external examiners, the process has been strongly valued for its general contribution to professional development in relation to marking, standards and wider assessment literacy.
- Calibration events prompt participants to carry out similar activities in their departments or programmes to improve consistency of marking (e.g. Oxford Brookes University, Kings College London and the University of the West of England, Bristol).

Participants' initial standards

- In every case, marking activities demonstrated variation in judgements about the quality of student work which, if reflected in marking and external examining, can impact on standards. It is worrying to think about the implications if the popular press got hold of the wide range of marks awarded to the same pieces of work.
- In general, difference in markers' standards does not seem to be dependent on university mission group.
- We did not find standards in specialist institutions to be higher or lower than the same subject in university departments.

What to calibrate

- We found it was important to focus on calibrating achievement against specific learning outcomes or levels of achievement (e.g. 2.2. / 2.i borderline) rather than types of performances (e.g. essays, dissertations, recitals) which will, typically, combine different learning outcomes.
- Sub-disciplinary expertise (e.g. Equity and Trusts Law or Human Geography) was shown to be important in influencing decisions about standards. However, external examiners will typically have oversight of programmes involving multiple modules and therefore need to consider standards generally. For example, agreeing standards of knowledge, use of cases, application, analysis, drawing conclusions and communication that apply in all foundation subjects of law.

Creating shared written standards

- Calibration provided the opportunity for participants from different HEPs to develop shared, simple criteria for similar pieces of work, for example, music recitals, final year projects in chemistry, law essays.
- Exemplar marking appears to be crucial to surface what is important in making judgements (e.g. technical v expressive performance, grammar v content, ability to communicate experimental findings rather than understanding of the science involved, a synoptic view v particular aspects). Further calibration is needed to explore the balance of these characteristics in deciding 'cliff edge' judgements such as borderline 2.i./2.ii and 2.i./ 1st.
- Asking groups to highlight the three most important characteristics in deciding a grade worked well to determine key characteristics for judging standards.

---

[13] Higher Education Academy, 2015. *A review of external examining arrangements across the UK.* Available at: https://www.heacademy.ac.uk/project-section/review-external-examining-arrangements

- Producing good quality completed material as an outcome of calibration activities with the aim to share it more widely across a subject discipline can be a challenge. There is usually no dedicated staff time available in either PSRBs or university consortia.

Impact on consistency of standards

- It is too early to say whether the various initiatives improved consistency of standards beyond the local group assembled or over the longer term as these initiatives are just the first steps in a process; rigorously testing this requires well-funded, longitudinal evaluation.
- Institutional calibration, done well, appears to provide very powerful feedback, much more so than external examiner feedback, we would suggest. This involves examples of student work reviewed by multiple examiners. However, it involved enormous amounts of preparation in order to work smoothly in the workshop. It only works well where participants are marking topics/ assessments that they are reasonably familiar with.

Making it happen

- Calibration requires a collaborative approach. We have found that whilst PSRBs and learned societies are often interested, it is difficult to drive forward PSRB-led calibration in a timely way because they often lack sufficient staff or resources to lead such an initiative.
- Current success has rested on having a committed and influential individual to drive the process, but this does not bode well for a sustainable process. There is currently no mandated requirement for examiners or institutions to participate in calibration, and therefore delivering and sustaining activity in subjects has proved enormously difficult.
- There seems to be more appetite for calibration where a subject perceives a difficulty in obtaining consistency of marking, for example, marking final year projects or performances.

Running calibration events

- A critical mass of participants (and ideally institutions) is needed for an effective calibration process. We recommend at least 12 participants.
- Discussion is best facilitated by a skilful subject specialist because of the importance of familiarity with disciplinary terminology and ability to understand and reflect the impact of the nuances coming through the group discussions.
- Technology such as Google Docs, used to capture and display the delegates' marks and comments during the event, is very useful and time-saving. It also provides hard copy, for example of key characteristics, for later use in creating calibration materials.

What has worked best: Principles for running calibration events

The Project learning has generated the following principles for running calibration events:

- Focus on calibrating against a specific standard (e.g. reflective writing for vets, recital performance in music).
- Avoid narrow specialism (e.g. not just a violinist judging violin standards but all instruments).
- Focus on the learning outcomes to be demonstrated via the task, not the task itself.
- Focus on broad grades, not specific percentages.
- A critical mass of participants is required (we suggest a minimum of 12 participants).
- Allow enough time for quality dialogue.
- Ensure participants are aware of the pre-meeting work.
- Distribute pre-meeting work in good time.

- Keep judgements as anonymous as possible to allow for free discussion, to avoid defensive marking and to avoid problems of perceived differences in experience or status.
- Encourage participation through starting with tricky assessment tasks for a subject.
- Focus on agreeing what is important in making judgements – drawing out examples and shared descriptions of why the pieces were marked at that level – both are needed: exemplar and descriptor.
- Use technology to facilitate both pre-work and the calibration event.

Advance HE has created a set of toolkits and case studies that provide detailed guidance and resources for running a calibration event, go to https://www.heacademy.ac.uk/degree-standards

Moving Forward

Responses to our calibration initiatives have been extremely positive both for external examiners and for other academics in terms of attempting to establish consistent, appropriate, standards. However, calibration remains at an early stage in its development and needs further trial and evaluation. To have sector impact, it requires a sustained, structured, iterative cycle and the potential for this has been demonstrated elsewhere, for example by the Accountancy field in Australian universities. Calibration also has benefits for staff assessment literacy, fairness to students and evidence for TEF submissions of efforts to safeguard standards. There are other ways to safeguard and maintain standards such as national exams but these are largely unwelcome in the sector. Consequently, it is important to continue investigating the power of calibration to restore waning confidence in sector standards.

The Degree Standards Project seeks the opportunity to continue developing calibration with PSRBs, HEPs and other interested parties. In particular, it seeks to investigate the potential of the following recommendations to create a sustainable process.

1. If formal work on calibration across the UK is to succeed, it needs serious, sustained support at a national level from government, Office for Students, PSRBs, sector organisations and universities.
2. PSRBs and consortia should not rely on one individual to drive the work within subject disciplines, it needs committee support.
3. PSRBs and learned societies can support calibration by linking it to programme accreditation.
4. Universities can support calibration by linking it to the appointment of new external examiners as a desirable quality.
5. Universities can support calibration by including responsibility for local calibration activity in the job descriptions for heads of quality, heads of department and programme leaders.
6. Universities can support local calibration by building it into internal QA procedures.
7. Heads of Department can support national calibration by supporting subject initiatives.
8. HEPs should organise at the regional level (local HEP consortium) to facilitate participation for larger disciplines.
9. Smaller, specialist institutions can work together at the national level to provide fora for calibration.
10. Calibration can be used to create subject exemplars which can be located on PSRB/learned society websites for others to use at local and national levels.

Recommendation

The Advance HE Degree Standards project work on calibration has been encouraging and has demonstrated that calibration is a promising method to achieve comparability of standards when supported by documented national standards. External examiners are a crucial element of UK quality systems aimed at maintaining appropriate and comparable standards across the sector but only when their standards are calibrated. Therefore, we recommend that, at strategic and policy levels, the UK should take forward further development and evaluation of a sustainable system of formal calibration of academic standards for external examiners.

Sue Bloxham, Nicola Reimann, Chris Rust December 2018