Rationalising subjectivity: using learning technology to automate and flexibilise marker standardisation in higher education assessment

Practitioner Research
In Higher Education

University of Cumbria
Online First pages 60-68

Tom Mayer
Kaplan International Pathways UK

Abstract
When designing a marker standardisation training programme, institutions must respond to their context and choose the method appropriate to it. This article details an approach based on a self-access, automatic package of training materials that can be flexibly deployed via a learning management system. This responds to the needs of a large, multi-college teaching network marking shared assessments, providing an authentic, flexible, and inclusive experience that is worth the up-front resource cost and has been broadly positively received. Lessons have been learned centring on the implementation of this method of training, communicating the intention behind the change, and the approach to adapting it to suit an academic community of practice. The training methodology and evaluation may serve as informative to other institutions seeking an approach to standardisation that meets their needs.

Keywords
Marking; assessment; standardisation; reliability; calibration; moderation.

Introduction

Marker standardisation training is central to successful human-marked summative assessment. It helps to improve rater reliability and helps academic staff who are both marking and teaching prepare their students for the assessments to come. There is, however, no single best practice for delivering standardisation training to realise its benefits. This is owing to the varying circumstances of higher education institutions.

This article details the approach to marker standardisation training developed at our network of nine international pathways colleges feeding into UK higher education. The method is asynchronous, assisted by technology that allows the training process to be automated and also flexible. This saves time of academic managers who no longer need to run synchronous training for all incoming markers and has dividends for workplace inclusivity. The benefits of this approach to us – authenticity, flexibility and inclusivity – are conditioned by our institutional environment, but it is hoped that some, if not all, of its features may help other institutions develop practices that cater to their own needs. As well as these positive outcomes, there have been some lessons learned, leaving room for improvements that we expect to implement in future iterations of the training. The use of benchmarked samples to test standards has also been an exploration in the socio-material nature of assessment standards, highlighting the interplay between marking tools, student work, and markers' expertise.

This approach to standardisation training was used to standardise marking on English for Academic Purposes achievement tests (both written and spoken) and written and spoken assessments on study skills/research modules. Therefore, the insights of the method's successes and points for improvement may be of use to practitioners working on a number of assessment types in a range of different contexts.

Citation
Mayer, T. (2024) 'Rationalising subjectivity: using learning technology to automate and flexibilise marker standardisation in higher education assessment', *PRHE Journal Online First*, pp. 60-68.

What is marker standardisation?

Marker standardisation, as discussed here, refers to the training undertaken prior to the subjective marking of an assessed piece of student work to support the reliability of that marking.

Standardisation is necessary when human markers judge the assessed work of different students against performance expectations described in marking tools. It is particularly important when these markers make decisions that lead to accreditation of learning or other significant life-impacting decisions. Standardisation training is needed regardless of whether an assessment is being marked by one marker or a team of many. In the former scenario, the one marker must reliably judge different pieces of work relative to one another, ensuring that they are able to differentiate comparable performances against appropriate award or marking outcomes. Where there are multiple markers, it is also necessary to eliminate the impression or reality that more or less lenient markers will have unjustifiable impacts on students' outcomes. There is therefore a role for standardisation in ensuring both inter-rater (consistency between markers) and intra-rater (consistency of a given marker between marking instances) reliability.

Standardisation training therefore seeks to achieve Sadler's (2013, 2016) goal of 'calibration'. A calibrated academic is one who is able to apply the expertise by which they qualify to mark a given assessment in order to reach a grade to comparable that given by a similarly calibrated colleague or counterpart. Sadler pointedly and purposefully distinguishes between calibration and moderation, a process sometimes confused with 'standardisation', where markers review one another's grading decisions after the fact (often with some recourse to revision or scaling of the originally awarded mark).

While marker moderation is a vital element of monitoring the reliability and quality of marking, when it leads to a review of marks it has the potential to increase the resource cost of marking, such as in a scenario where all assessed work is double marked. As Sadler (2016) details, moderation also tends to have less of the narrowing effect that successful calibration has on marker judgements. Furthermore, uncalibrated academics may struggle to agree on equitable moderation outcomes when they lack an agreed standard to work toward. There is therefore a need for marker standardisation both in terms of its own value but also with respect to how it further assists the good practice of moderation.

For the purposes of this article, reflecting our own practice and that found elsewhere in higher education, 'standardisation' refers solely to the pre-marking calibration training process. Some institutions use the terms 'calibration' or 'moderation' to refer to this training, and this article also speaks to those contexts.

Marker standardisation in practice in higher education

While marker standardisation can be defined, there is no single method of marker standardisation that an institution or manager can draw upon. The typology detailed by pedagogists at King's College London (King's Academy, 2018:3-4) illustrates a range of options available when designing standardisation training. All involve markers working with sampled student work, albeit to different ends. King's Academy's (2018) 'ideal' model is a synchronous standardisation meeting, where all markers gather to agree marks for a set of samples, contrasted with asynchronous comparison of remotely submitted marks and a third method based on markers having access to samples marked by an expert panel to refer to. These models each have their benefits, though it is noteworthy that the authors single out the synchronous meeting as 'ideal'.

This reflects a long history of using synchronous – once necessarily face-to-face – standardisation meetings for the purpose of marker standardisation, which has nevertheless been under scrutiny as

a necessary part of the standardisation process for some time (Knoch, Read and von Randow, 2007; Raikes, Fidler and Gill, 2009). A study by O'Connell et al. (2016) does admirably demonstrate the efficacy of this approach, but subsequent literature on the subject has affirmed that the synchronous-asynchronous divide between different approaches to standardisation does not have a significant impact on the quality of calibration outcomes; this depends instead on the more fundamental rigour of the process (Zahra et al., 2017).

In chasing this aim of 'rigour', it is perhaps necessary to follow writers (Smith, 2012; Beutel, Adie and Lloyd, 2017) who frame standardisation – a time- and resource-costly effort to undertake in an academic environment – not (perhaps begrudgingly) as a necessary condition for proceeding to get underway with a marking effort, but rather as an integrated part of good academic practice and constructive alignment. With this principled mindset in place, it is more important to establish a contextually appropriate method than it is to chase an orthodoxy regarding best practice.

Use of technology to assist marker standardisation

While the above-discussed literature does address standardisation training that occurs asynchronously – often mediated by technology as a means of communication, e.g., a spreadsheet that records different markers' responses to sampled work – there is less consideration of technology as a means of enhancing the process. A notable exception, work by Willey and Gardner (2011), highlights how software can be used to facilitate automatic inter-rater comparison of outcomes of standardisation training. While the technology they use highlights specific affordances on applications that may not be available to every marking institution, their focus on building a 'community of practice' is a valuable frame for the role that technology can play in ensuring that standardisation training fits within a holistic pedagogical process, as well as measuring its outcomes and success.

Our solution for standardisation training

The solution developed to standardise marking on our assessments involves a training workflow of four key stages. These stages are delivered via our virtual learning environment (VLE), taking advantage of self-access interactive learning media which includes activities that assist training and scaffold the calibration of marking. This is broadly as a replacement for standardisation meetings resembling the King's Academy (2018) ideal.

The stages of the workflow emulate standard practice in the marking of English language tests, such as that set out by the British Council Language Policy Division (2009). Given this origin, we first implemented the solution in training our English for Academic Purposes test markers before expanding the roll-out to the summative assessments for two extended project modules which are respectively taken by (the vast majority of) our pre-undergraduate and pre-master's students.

The human-marked English assessments are both a written essay and a spoken conversation, while the project assessments include academic skills assessments, extended written reports and assessed presentations. Therefore, this solution has been applied to different types of assessment, though all of them require interpretive human marking of qualitative outputs.

Acclimatisation

In an initial stage, prospective markers are introduced to the relevant assessment and its marking tool using a slideshow embedded in interactive media. Most of our assessments use analytical marking criteria, with different aspects of student performance measured by differently weighted sets of descriptors that correspond to a numerical scale within each criterion. A key aim of acclimatisation is therefore to highlight the different elements of work covered by different assessment criteria, helping markers to look past a title, which may have limited coverage. For

example, a criterion entitled 'Critical Appraisal' may assess students' synthetic writing skills in addition to their successful critical appraisal of literature, as the title suggests. This exercise prepares markers – especially those newer to the assessments – for applying marking tools to students' work. Our implementation encourages self-reflection, at this stage, however an interactive test could be added here.

Illustration

Markers then review a set of authentic samples of work, again embedded in interactive media. These samples have previously been benchmarked by a panel of experienced markers who agree on a mark and produce a commentary that evaluates the work against each criterion in the marking tool. This stage will include a low-performing, a passing, and a high-performing sample. The use of interactive media allows us to include samples of both written work (as embedded PDF files) and audio-visual material, such as recordings of sampled presentation assessments.

After reviewing each sample and its commentary, markers are prompted to complete a formative drag-and-drop exercise linking the marks assigned by the benchmarking panel to each section of the commentary, along with the weighted average overall mark. They will be able to see the correct answer to the exercise upon completion.

The purpose of this stage is to cultivate initial mental expectations of student performance at different levels. By forming a link between a numerical mark, an example of student work and the benchmarking panel's rationale for their decided mark, markers begin to calibrate their perception of the relationship between concrete performance, marking descriptors and the relative measurement of success.

Practice

Following Illustration, markers are offered an opportunity to test their emerging marking expectations, continuing in the interactive media package. This stage includes additional benchmarked samples of work, this time accompanied by partial commentaries that display the comments and marks for some criteria but not others. The interactive interface leaves space for markers to write their own marking notes for those criteria without commentaries, leading toward a further formative test where they assign marks for the missing criteria (using a free-entry textbox) and thus the resultant overall score for the sample. As with the activity in Illustration, the benchmarking panel's judgement will be revealed to the markers on completion, along with the full commentary.

By removing some sections of the commentary, this stage is a freer practice exercise requiring markers to make their own judgements as to the best measurement for a student's performance. The outcome of the formative stage can prompt an entrenchment or adjustment of their marking expectations.

Marks Collection (and Follow-up)

The final stage, which forms the summative assessment of markers, is normally delivered as a quiz within the VLE. Markers are presented with two samples of student work without any commentary or benchmark material. They are required to review each sample and assign marks for each of the respective criteria and then an overall mark based on their allocations.

If the two overall marks they allocate are within the usual tolerances of our marking process (within 5 marks either side of the benchmark), the marker is automatically considered to be standardised and they may proceed to mark ongoing summative assessments.

If one or both of the marks given are not within these tolerances, trainees are shown the commentaries and marks assigned by the benchmarking panel to further calibrate their marking expectations. They then are prompted to complete the Follow-up quiz, which repeats the Marks Collection exercise with two further samples. If both of these are completed within the expected tolerances, the marker will again be considered automatically standardised. Otherwise, they will need further, human-led support to reach standardised status.

Why did we develop this approach to standardisation training?

This approach – like any institution's approach to marker standardisation training – arises from needs that reflect our context. Key attributes of our teaching context include:

- **The scale of our operation**: our network of nine colleges (as well as our online education offering) caters to a large number of students, all working toward common accreditation. In the 2022-23 academic year, the number of students was close to 7,000. Students on the same or similar programmes at different colleges work toward the same assessments on the same modules, with the same standard expected.
- **Multiple enrolment points**: across an academic year, there are multiple points at which students can join us to study on our programmes. All of our colleges have major Autumn and Spring intakes (normally in September and January) while many have additional joining points in the Summer and more intensive offerings in-between.
- **Our employment environment**: the scale of our operation, its dispersal across multiple colleges, and the variability of our student cohort size across an academic year necessitates new staff being brought on to teach at different points. Our academic teams include many fractional staff employed on a part-time basis alongside other employment, educational and life commitments, with no single joining point to the year. This compounds the ordinary pattern of changing personnel within an organisation, meaning that staff join the marking teams of our modules at different points.

These three broad contextual details drew us to three key benefits of this approach to marker standardisation training.

Authenticity

When large volumes of submitted work need to be marked within tight turnarounds, such as those that allow for post-marking moderation, it is necessary for markers to make and stand by their academic judgements, relying on colleagues' support only in cases where they are truly unable to reach a confident grading decision. Therefore, our marking is normally carried out individually. The move to the new approach is thus more authentic to the marking experience that our colleagues can expect on our assessments. There is therefore a case to be made that the training will be more valid, and that it can act as a better introduction to the assessment more broadly.

Flexibility

As the training packages are self-access, curriculum leaders can direct prospective markers to complete the package at any point in the academic year. This responds to the changeability of our marking teams and the competing demands on many of our team members' time. Given the fact that a marker can, provided that they mark both or three of four of the test samples satisfactorily, complete the training without further input from their managers, who would previously have had to organise standardisation meetings, we can also say that we benefit from the automation of the process.

Inclusivity

Our college staff and thus our marking teams come to standardisation training with a variety of levels of experience in both education and our network. When standardisation training is conducted as a meeting, it is possible that the less experienced teachers and markers present will feel less able to offer their judgements or contradict the contributions of more experienced colleagues, preferring instead to minimise engagement. Conversely, more experienced attendees may consciously or unconsciously experience pressure to be 'correct' in their judgements. Personality and other group dynamics may further exacerbate this situation. Moving to an approach where the outcomes of standardisation training are measured in an individual, private test taken by all markers ensures a fair and equal coverage regardless of these differences in disposition.

The flexibility of the training also has inclusive outcomes, including accommodating markers who work around caregiving responsibilities and allowing them to complete the training at their own pace.

Implementation

In order to deliver the standardisation training package, a benchmarking panel of experienced markers of the relevant assessments is convened, chaired by an academic developer who acts as project manager. Ahead of scheduled benchmarking meetings, the project manager provides the panel with samples. The panel mark these samples ahead of the meeting, sending their marks to the project manager.

Considerations for the approach to benchmarking meetings are:

- Whether the entire benchmarking panel will meet to agree the benchmarks for all samples or whether they will work in smaller sub-groups to agree a smaller number. The former ensures a greater consensus on the benchmarks assigned, while the latter is potentially more efficient. In events where the sub-groups cannot come to a complete consensus, they can refer the sample to the full panel.
- Whether the panel will mark the exact number of samples needed for the finished packs or whether they will instead mark a larger number, with the project manager selecting those that will be taken on in the pack, thus requiring full agreement and commentaries. The former means that the panel will need to mark fewer samples, however the latter can help filter out samples that the panel might struggle to agree on a benchmark for.

Regardless of approach, the panel will meet to agree a mark for each criterion and its rationale. While the time taken varies by the length of submission and the nature of discussion, we generally found that samples marked against four or five marking criteria could be agreed in thirty minutes. Following the meetings, drafting and reviewing commentaries for each sample is assigned to panel members. These can be built into the training package by the project manager.

Oversight of the training package is the responsibility of each college's module coordinator, who assigns the training to new markers and monitors results. When a marker is still not considered to have been successfully standardised after the Follow-up quiz, it is the responsibility of their module coordinator to offer extra support and ultimately judge the standardisation status of the marker.

Evaluation of practice

The roll-out of this approach to standardisation training over two academic years has been successful but it has also offered lessons to learn for future improvements.

Staff feedback was sought on the training materials via workplace surveys and, among 53 respondents, there was broad support for many aspects of the training. A majority had favourable attitudes towards the format of standardisation while substantial majorities appreciated the ability to carry out the Marks Collection stage in their own time. The training packages were also widely referred to as easy to use, and module coordinators were generally of the view that the packages were easy to administer.

Preference for meetings

The feedback was not, however, wholly positive. Specifically, those who reported having previously undertaken the meeting-style training were overwhelmingly of the view that this had been a positive approach and there was correspondingly some feedback that regretted the move from such meetings. Among these responses were a number that expressed confusion as to why this method had been chosen.

While the standardisation package was designed to be accessed individually by markers, it is possible for the Acclimatisation, Illustration and Practice stages to be used to structure synchronous discussions and meetings. Where they are possible (which they would not be at all times in our context, as discussed above), conversations with colleagues that use the initial three stages of the package as a stimulus may enhance the training's pedagogical value. This is provided that such meetings are conducted in line with the principles that underpin the training; more experienced or dominant personalities should not dominate discussions. In this vein, individual engagement with the Marks Collection test should remain the ultimate end point of the training.

Change management

This feedback has also prompted reflection on how the new training method was rolled out across our network. As these materials were created centrally – by an academic developer from outside the colleges working with a benchmarking panel from across them – it is important to see this move, like others, through the lens of change management. When change is approved by organisational leadership, it is important to make effective contact and provide adequate communication about the justification for the change in order to achieve acceptance and prevent abandonment even after costly implementation (Connor and Patterson, 1982). This is especially true when, as appears evident in the feedback, the *status quo* was appreciated. This has driven us to be more considered in our communication approach in future.

Implementation costs

The drive to make use of the self-access nature of the training may also reflect the resource cost of assembling the training package. A package containing ten benchmarked samples of work would require at least five hours of panel meetings, as well as time to mark the samples and produce the commentaries followed by building time. This must then be repeated for all assessments on a module alongside the other commitments that academic staff may have. Such investment of human resources rationally anticipates the hope of a return that may come in the form of the time saved by not preparing and running a standardisation meeting. Note also that this resource demand adds an additional consideration for a manager implementing a training package of this type, that being that it is unlikely that an institution would think it worthwhile to make a considerable resource investment on a smaller module that affects fewer students and thus a smaller marking team.

Potential for disagreement

As well as this implication for resources, the use of a select benchmarking panel to develop a training package that will then be used to test the marking expectations of a larger group of markers presents a potential challenge. While such a panel has a role to play in setting the standard that will be expected in marking, it is possible – as is not unreasonable in dealing with subjective academic

judgements – that the benchmark set may be irreconcilable with a view held in the wider marking population, including those with comparable experience to the panel members. Continual monitoring of the results inputted at the Marks Collection stage, facilitated by the VLE, has proven to be an effective way of measuring any deviation. Where there is a low rate of trainees entering the 'correct' answer for a sample, we have been able to analyse underlying causes. Where there is a systematic error in one 'direction' (i.e., trainees are consistently rating a sample too high or too low), it has been grounds to re-convene benchmarkers to consider either an adjustment or to affirm the existing mark. If differences are more diverse, however, it would be grounds to address whether there is sufficient guidance at the earlier stages to set markers up for success in the test.

Some might point toward this adjustment as a failure of training package, however it perhaps instead points toward the socio-material nature of assessment standards described by Ajjawi, Bearman and Boud (2021). This is to say that for a student's work to be judged and graded, there must be a dynamic interplay between the work itself, the stated standards of the marking tool and the interpretive expertise of the marker. No pre-marking standardisation training can anticipate all the permutations of student performance that the marker is going to encounter on completion. Instead, the training should both condition markers' expectations and further develop their ability to use their expertise to apply written marking standards. In this light, testing for the grade assigned to benchmarked training samples could be considered secondary to providing the opportunity to practice marking and the interplay it necessitates.

Emotional impacts

This point of contention goes together with an affective dimension to the training. As the training requires each marker's judgement to be tested against a standard, this raises the possibility that markers will be told that they are 'wrong'. Staff feedback included details of feelings of demotivation and demoralisation that arose from being told that some colleagues had 'failed' standardisation, including among those staff with more experience on the assessments in question. While this does reflect that the new approach moves away from the possibility of a marker being considered standardised on account of simply attending a meeting, it is not necessary to engender negative feelings toward the standardisation training among the professionals taking it. In subsequent iterations, therefore, we have decided to move away from the language of failure, stressing that the alternative outcome to passing is further support toward standardised marking.

Being receptive to the need to adjust benchmarks and their associated commentaries, as well as a non-punitive approach to not successfully completing the training reflects the desire to implement the above-discussed holistic view of marker standardisation as part of academic communities of practice rather than an obligation and barrier to continued professional conduct (Smith, 2012; Beutel, Adie and Lloyd, 2017).

Conclusion

The scale of our operation has made producing these training materials a viable and beneficial undertaking, but we have learned that they do not represent a static output but rather a medium for a more dispersed dialogue among our academic teams. Using technology means that staff performance in the training can be monitored, which not only aids its implementation but also the monitoring of its quality and flexible responses. These are affordances that neither standardisation meetings with no final individual check stage nor static training materials have, which is a benefit to this approach that may cut across all institutional contexts.

Acknowledgements

I am grateful to the colleagues who implemented and undertook the training under discussion and who fed back on it. I would like to also thank the journal editor and my anonymous reviewers for

their constructive comments. My further thanks go to Drs Margaret Bryndal, Panos Dendrinos and Kristine Sheets for their assistance in the ideation and realisation of this article.

References

Ajjawi, R., Bearman, M. and Boud, D. (2021) 'Performing standards: a critical perspective on the contemporary use of standards in assessment', *Teaching in Higher Education*, 26(5), pp.728–741.

Beutel, D., Adie, L. and Lloyd, M. (2017) 'Assessment moderation in an Australian context: processes, practices, and challenges', *Teaching in Higher Education*, 22(1), pp. 1-14. doi: https://doi.org/10.1080/13562517.2016.1213232.

British Council Language Policy Division (2009) Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR): A Manual. Strasbourg: British Council. Available at: https://rm.coe.int/1680667a2d (Accessed: 18 January 2023).

Connor, D.R. and Patterson, R.W. (1982) 'Building commitment to organizational change', *Training and Development Journal*, 36, pp. 18-32.

King's Academy (2018) *Standardisation of marking*. London: King's College London. Available at: https://blogs.kcl.ac.uk/aflkings/files/2018/09/Standardisation-FINAL.pdf (Accessed :17 May 2023).

Knoch, U., Read, J. and von Randow, J. (2007) 'Re-training writing raters online: How does it compare with face-to-face training?', *Assessing Writing*, 12(1), pp.26-43. doi: https://doi.org/10.1016/j.asw.2007.04.001.

O'Connell, B., De Lange, P., Freeman, M., Hancock, P., Abraham, A., Howieson, B. and Watty, K. (2016) 'Does calibration reduce variability in the assessment of accounting learning outcomes?', *Assessment & Evaluation in Higher Education*, 41(3), pp.331–349. doi: https://doi.org/10.1080/02602938.2015.1008398.

Raikes, N., Fidler, J. and Gill, T. (2009) Must examiners meet in order to standardise their marking? An experiment with new and experienced examiners of GCE AS Psychology. Available at: https://www.cambridgeassessment.org.uk/Images/109782-must-examiners-meet-in-order-to-standardise-their-marking-an-experiment-with-new-and-experienced-examiners-of-gce-as-psychology.pdf (Accessed: 09 May 2024).

Sadler, D.R. (2013) 'Assuring academic achievement standards: from moderation to calibration', *Assessment in Education: Principles, Policy & Practice*, 20(1), pp. 5-19. doi:: https://doi.org/10.1080/0969594X.2012.714742.

Sadler, D.R. (2016) 'Assuring academic achievement standards: from moderation to calibration', in Klenowski, V. (ed.) *International Teacher Judgement Practices*. Oxon: Routledge. pp.15-20.

Smith, C. (2012) 'Why should we bother with assessment moderation?', *Nurse Education Today*, 32(6), pp.e45-e48. doi: https://doi.org/10.1016/j.nedt.2011.10.010.

Willey, K. and Gardner, A. (2011) Building a community of practice to improve inter marker standardisation and consistency, Available at: https://opus.lib.uts.edu.au/bitstream/10453/19223/1/2010006640. (Accessed: 09 May 2024.

Zahra, D., Robinson, I., Roberts, M., Coombes, L., Cockerill, J. and Burr, S. (2017) 'Rigour in moderation processes is more important than the choice of method', *Assessment & Evaluation in Higher Education*, 42(7), pp. 1159-1167. doi: https://doi.org/10.1080/02602938.2016.1236183.