Studies in Higher Education

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/cshe20

Social moderation and calibration versus codification: a way forward for academic standards in higher education?

Berry O'Donovan, Ian Sadler & Nicola Reimann

To cite this article: Berry O'Donovan, Ian Sadler & Nicola Reimann (26 Feb 2024): Social moderation and calibration versus codification: a way forward for academic standards in higher education?, Studies in Higher Education, DOI: 10.1080/03075079.2024.2321504

To link to this article: https://doi.org/10.1080/03075079.2024.2321504

Published online: 26 Feb 2024.

STUDIES IN HIGHER EDUCATION
https://doi.org/10.1080/03075079.2024.2321504

Routledge
Taylor & Francis Group

Social moderation and calibration versus codification: a way forward for academic standards in higher education?

Berry O'Donovan [a], Ian Sadler [b] and Nicola Reimann [c]

[a]Business School, Oxford Brookes University, Oxford, UK; [b]School of Sport and Exercise Sciences, John Moores University, Liverpool, UK; [c]School of Education, Durham University, Durham, UK

ABSTRACT

A key responsibility of higher education providers is the accurate certification of the knowledge and skills attained by their students. However, despite an intense focus on developing relevant quality assurance regulations, academic standards in higher education have remained resistant to explication and consistent application. In this paper, we initially deconstruct and evaluate academic standards and dominant practitioner perspectives on their nature and use, including techno-rational, sociocultural and sociomaterial approaches. The limited prior research on the effectiveness of calibration and social moderation processes is reviewed, highlighting the significant challenges in sharing tacitly held understandings of assessment criteria (attributes of quality) and standards (levels of achievement). Further complications are considered that arise from the varying expertise and power relationships of assessors and the complexities inherent in the development and use of codified artefacts for capturing and sharing standards. We opine that because of the difficulties in clearly representing and agreeing standards, it is unsurprising that there is little evidence of marking consistency to be found in the literature even in contexts where carefully crafted artefacts, such as rubrics, are in use. We conclude that effectiveness would be enhanced through sharing understandings more widely and refocusing the use of assessment codifications towards how they may catalyse effective social moderation and calibration dialogues. Dialogues that foreground individuals' positions of consensus and dissensus at significant points of interpretation in the assessment process are identified within the paper.

ARTICLE HISTORY
Received 16 August 2023
Accepted 16 February 2024

KEYWORDS
Academic standards; assessment; social moderation; calibration; higher education; assessment criteria

Introduction

Trust in the accurate certification of the knowledge and skills attained by their students is centrally important to higher education institutions across the world (Biesta 2008; Naidoo and Williams 2015). Globalisation has since the early 1990s accentuated the interest of nation states in higher education as a basis for international competitiveness arising from knowledge, innovation and professional development (Mok 2016). This more market-orientated and competitive higher education context (O'Byrne and Bond 2014) has provoked 'several decades of increasing regulation and accountability regarding academic standards' (Bloxham and Boyd 2012, 615), which has been further fuelled by

CONTACT Berry O'Donovan bodonovan@brookes.ac.uk Business School, Business School, Oxford Brookes University, Headington Campus, Room CLC. 1.27, Oxford OX3 0BP, UK

ongoing concerns in the media about 'dumbing down' and grade inflation (Lambert 2019). However, the academic standards used in higher education assessment practices remain difficult to pin down and resistant to processes seeking their transparency, consistent application and portability across contexts. This paper aims to conceptualise a complex and ill-defined area and challenge a range of assumptions that are taken for granted in contemporary higher education assessment and marking policy and practice. We initially respond to calls for clarity on the nature of academic standards and key terms and concepts. Subsequently, we examine the challenges involved in, and suggest ways of improving, the accurate and consistent representation and application of academic standards through calibration and social moderation assessment processes.

Defining terms: academic standards

Before considering the issues involved in achieving the consistent representation and application of academic standards, we first seek to define terms. Often when the word 'standards' is used in higher education, the mind quickly moves to quality assurance processes. Based on Boyd and Bloxham's (2014) work, we define such standards as *quality standards*, i.e. processes put in place to ensure fair and consistent practice in assessment. Quality standards refer to adherence to regulations and policies as well as inputs such as staff qualifications or library resources. These differ from academic standards, which reside within student performances and work itself. Sadler (2014) suggests that what is meant by the term 'academic standards' is rarely explained or questioned by key stakeholders in higher education and commonly conflated with assessment criteria. This is understandable, as both assessment criteria and academic standards come together in codified form in, for example, rubrics, grade descriptors, marking guidelines and learning outcomes. However, whilst we acknowledge that assessment criteria and academic standards are inextricably entangled in practice, we conceptualise assessment criteria, as propounded by Sadler (2014), to be the *attributes of quality* that are being sought in the assessment of learning outcomes[1] as distinct from academic standards as the *level of achievement* at which such attributes of quality are enacted. We make this distinction because we consider assessment criteria and standards not only conceptually different, but also that the practical difficulties experienced in their representation and performance as distinct. To defend this position, we first define two distinct types of standards: sharp standards which refer to precise, measurable accomplishments; and relative standards that refer to matters of degree (Sadler 1987). *Sharp standards* tend to have precise boundaries and represent declarative knowledge involving specific facts or propositions which can be assessed consistently as correct or incorrect. Examples include the assessment of errors in a multiple-choice language test, a computer programme coding task or the use of referencing conventions. In many disciplines at university, sharp standards are often most relevant to early-stage learning outcomes. *Relative standards* are more challenging to represent and consistently assess. However, these are habitually used in higher education assessment contexts in which complex and open-ended tasks predominate. These include conventional as well as more contemporary assignments such as essays, dissertations, portfolios or podcasts where students often have considerable latitude in how they respond to a task. Precise verbal description of such relative standards is usually unachievable as their explication rests on relative terms, such as '*highly* analytical'. Terms such as these are meaningless unless they are illustrated by examples of student work at different levels (Sadler 1987). Relative standards are socially constructed and largely tacit, dependent on context and interpretation. Again, this can be seen in the example 'highly analytical', which could be applied to the work of a doctoral student or undergraduate but represents diverse levels of achievement (O'Donovan, Price, and Rust 2004).

Whilst the literature on academic standards does not generally distinguish between types of standards, it does offer two conflicting perspectives that have been summarised by Ajjawi, Bearman, and Boud (2021) and Bloxham and Boyd (2012). The first is a techno-rational or representational view in which standards are considered as unaffected by values, culture or power (Bloxham and Boyd 2012). Here, academic standards are deemed able to be explicitly codified into stable and portable free-

standing artefacts, such as level descriptors in assessment rubrics. Such codifications are considered to be able to transfer a uniform meaning across separate audiences and the standards they represent are seen as separate from the knower. This seems realistic for assessments that embody sharp standards in which the level of performance can be applied [reasonably] consistently across different contexts, but problematic for relative standards that are both difficult to represent and subject to individual and contextual interpretation.

The second perspective is sociocultural whereby standards are understood as socially situated and interpreted. Standards are viewed as residing in and developed by the practices of academic and professional communities and are underpinned by, difficult to articulate, tacit knowledge (O'Donovan, Price, and Rust 2004). In this perspective, standards as levels of performance, are dynamic and interpreted and, over time, reinterpreted through the personal frameworks of educators operating in local contexts involving different assessment tasks, courses, and disciplines. Standards are seen as 'shifting, subject and practice-based' (Ajjawi, Bearman, and Boud 2021, 733) and consequently resistant to uniform representation. Unsurprisingly inter-rater reliability is higher for sharp standards and knowledge that requires recall but notably lower in relation to relative standards embedded in essay-style assignments and problem-style examinations (Watty et al. 2014).

Ajjawi, Bearman, and Boud (2021, 733) whilst accepting the contextual and interpreted nature of written standards argue that there *is* a need to render academic standards in 'tight form to span time, space and people' in today's global higher education sector. They propose a third perspective, a sociomaterial framing of standards. Considering standards as embodied in both attributes of quality and levels of performance, they propose that standards are routinely reified into concrete materials such as rubrics or marking guides, but these are then interpreted and enacted by people in complex social contexts (Ajjawi, Bearman, and Boud 2021). An inseparable entanglement between the individual, social and the material is suggested that 'takes the standard out of the internal cognitive processes of the individual and even beyond the "way things are done around here" of the social' (Ajjawi and Bearman 2018, 43). This differs from the techno-rational approach as a uniform meaning is not assumed. They caution that artefacts such as rubrics do not carry a single, agreed meaning, but require performances of interpretation by individuals and communities over varied contexts that inevitably lead to some 'tinkering' or micro-adjustment to the standard, but without it becoming unrecognisable. However, how far a standard can be stretched and adapted before rupturing remains debateable.

The dominant perspective on academic standards in current higher education practice is the techno-rational view with an ever-increasing focus on the development of standardised rubrics, marking guidelines, level descriptors across departments or even whole institutions. The assumption is that standards are stable, portable, and capable of being objectively codified in written form. We posit that this perspective has gained traction not because of empirical evidence, but in tandem with the increasingly competitive nature of higher education in which institutions have become progressively commercial and managerial. An intensification of regulation and control has erupted as governments and institutions attempt to monitor, enhance, and make transparent the quality of their educational provision in their struggle to 'measure the unmeasurable' (Hill, Lewis, and Maisuria 2015, 607) and compete in a more hostile environment (Harvey 2005; O'Byrne and Bond 2014). In today's neoliberal higher education sector, many argue that trust is increasingly placed in transparent processes that can be monitored and far less in the professional judgement of skilled employees such as academics (Hill, Lewis, and Maisuria 2015; Tsoukas 2003). However, we would argue that the increase in the documentation of academic standards, in itself, is unlikely to improve consistency. It could even be considered to mask the issues as there are multiple 'points of interpretation' in the application of explicit standards where variability is likely to occur.

Calibration, social moderation and consensus moderation

Alarmingly, high levels of variation between markers have been repeatedly demonstrated, e.g. for dissertations (Stolpe et al. 2021), essays (O'Hagan and Wigglesworth 2015), portfolios (Pitts et al.

2002) and open-ended exam questions (Herridge, Tashiro, and Talanquer 2021). Such studies demonstrate that whilst the evaluative judgements of individual markers are often assumed to be understood and agreed upon, they are in fact rarely consistent or even 'open for scrutiny or discussion' (Delandshere 2001, 121). Such variation is borne not only from the non-standardised assessment tasks and associated relative standards common in higher education, but also the personal marking frameworks of individual markers (Bloxham, Boyd, and Orr 2011). Such frameworks are constructed, and reconstructed, through the act of marking and are influenced by assessors' previous experience of assessment and beliefs about higher education (Bloxham and Price 2015) as well as their current set of assignments (Bloxham, Boyd, and Orr 2011). Indeed, even when criterion-referencing is espoused, norm referencing most often plays a significant role in marking practice (Bloxham, Boyd, and Orr 2011).

Within this complex context the higher education sector looks to moderation practices to assure standards. Moderation, more broadly, has been described as an 'approach to agreeing, assuring and checking standards' (Bloxham et al. 2015, x). It is essentially a quality assurance process, already widely used in contemporary higher education, aiming to ensure the validity and reliability of assessment decisions and grades (Beutel, Adie, and Lloyd 2017). Whilst moderation practices commonly involve scrutinising and finalising assessment decisions that have already been made, more recent conceptualisations include a wider range of approaches that can take place pre-marking, intra-marking or post-marking (Bloxham et al. 2015). These include blind second marking, sample moderation, expert moderation, consensus or social moderation, or combinations of these approaches (Beutel, Adie, and Lloyd 2017). Here we focus on social or consensus moderation, terms that are used interchangeably in the literature. Social moderation is considered a dialogic process of peer review, carried out by members of a disciplinary and/or professional community, to discuss and compare judgements on exemplars of student work to reach a shared understanding of the academic standard (Bloxham 2009). However, social moderation has been used in slightly different ways by different authors. We would argue that in practice it subsumes a range of activities situated on a continuum. Consensus-seeking dialogue is core to all variants, while the exact purpose of this dialogue can vary. If responses to more than one assessment task type are considered and the dialogue is totally detached from the award of grades, the attention shifts from marking towards developing a shared understanding of the standards that underpin judgements about the quality of student work. This kind of social moderation, situated at one end of the continuum, is not about grading decisions. It is about the professional development of assessors, focused on clarifying standards and reaching an agreement on what performances that meet these standards might look like, beyond the concrete piece of work and the task under consideration. Such shared and portable understanding of standards is the key characteristic of Sadler's notion of *calibration* (2013). At the other end of the continuum are dialogic processes to negotiate and agree on the judgements about student work. The ultimate purpose being to agree a grade, albeit in a social, consensus-focused manner. In another variant, academics are brought together prior to marking student work, to discuss a sample of actual submissions or exemplars from prior cohorts. In this variant, the attention is on standards but with a view to supporting grading decisions which are coming up in the near future.

Social moderation dialogues can therefore have different primary purposes and adopt varied approaches in terms of timing and design. The term 'calibration' is, however, trickier to deconstruct. Whilst often used synonymously with social moderation, some authors following Sadler's (2013) example distinguish calibration in terms of its focus on the professional development of assessors. The 'calibrated academic', as described by Sadler (2013, 5), can make future grading decisions that align with the community standard. It is the assessor who is the object of calibration, not a particular batch of assessed work. Whilst in this article we use the term calibration in this way, we question if the term is misleading as the scientific connotations of calibration infer an objective accuracy of measurement that is unachievable in the context of relative standards.

Prior research on social moderation and calibration

Based on the nature of academic standards, logic would suggest that activities to promote the shared social construction of standards are beneficial. However, there are several issues with this. Firstly, there is only limited empirical evidence to support the assumption that social moderation and calibration activities do help to reduce variation in the application of academic standards in higher education. Secondly, based on the limited work that has been done, there are a range of practical and conceptual challenges to consider and overcome to ensure any activities are likely to be effective. Ultimately, the realisation of uniformity in the application of relative academic standards is tricky, possibly even unachievable, across different contexts. Next, we consider prior research on the challenges that undermine the effectiveness of social moderation and calibration in practice.

The benefits of social moderation and calibration have been repeatedly highlighted in the literature. Perhaps the most important argument that has been made is that social moderation and calibration enhance consistency between markers and reduce variation. However, there are surprisingly few studies that have investigated this empirically. Why might this be the case? Firstly, marking is an intuitive, private and individual practice in higher education (Watty et al. 2014) with little professional development for assessors (Beutel, Adie, and Lloyd 2017). Social moderation and calibration have not been embedded systematically into policy and are therefore not widely practised. Consequently, few naturally occurring data that could be used to evidence impact on marking consistency are available. In addition, studies which convincingly demonstrate impact on judgements made over longer periods of time require complex longitudinal designs that are neither easy to develop nor implement on the ground. However, there have been some advances in Australia based on the Tertiary Education Quality and Standards Agency (TEQSA), which require institutions to demonstrate approaches for assuring quality and standards. In response, a range of projects to explore peer review and moderation have emerged (e.g. Krause et al. 2014; Hancock et al. 2015). These projects have broadly reported positive outcomes in terms of the extent to which consensus can be reached through social moderation processes. Several published studies, based on the 'Achievement Matters' project (Hancock et al. 2015), have found reduced variation between assessors following calibration activities. O'Connell et al. (2016) reported a reduction in standard deviation of the scores awarded and in the gap between minimum and maximum ratings for the experimental group. This was based on two groups of research participants who individually graded the same three exemplars twice, with the experimental group undertaking calibration activities between the two rounds of grading. There were limitations, such as the small sample size that affected the reliability of the significance tests and the fact that the same exemplars were graded again in the second round, rather than testing whether the shared understanding gained through calibration affected future judgments. Despite these limitations, this study is noteworthy since it provides empirical evidence of enhanced consistency between assessors. Additional self-reported data, arising from the same project, showed that participating academics also believed that engaging in calibration facilitated consistency (Watty et al. 2014). However, the conclusion was that calibration reduces but does not remove variation between assessors as some differences remain (Hancock et al. 2015). Similar findings are reported by Palermo et al. (2018), Crimmins et al. (2016) and in earlier investigations such as the small-scale study in medical education conducted by Pitts et al. (2002) and unpublished practitioner research on UK-based calibration in Law by Hanlon et al. (2004); they are also echoed in school-based research (Wyatt-Smith, Klenowski, and Gunn 2010).

More recent studies have investigated the use of social moderation in different contexts such as vocational education (Gillis 2023) and teacher education (Brandenburg et al. 2023). In addition, there is research that has considered the perspectives on social moderation of different stakeholders including unit co-ordinators or module leaders (Mason, Roberts, and Flavell 2022) and expert academics (Mason and Roberts 2023). Finally, recent work has provided some initial empirical insights into the sources of inconsistency in judgements on the final thesis of primary school teacher education students (Rinne 2023). They found that variation in analytic judgements were based on whether the

assessor was focussing on the theoretical framework used or the academic language style. Holistic judgements varied due to the different weighting being given to particular aspects within the work. As becomes clear, the outcomes of the research on social moderation are inconclusive and there are a range of issues and contexts that require consideration. Clearly more research is needed.

Challenges of calibration and social moderation

At the heart of social moderation and calibration are the conversations with others to share, and challenge if necessary, implicit marking judgements regarding specific pieces of student assessments. The aim is that those involved develop a greater shared understanding of the factors influencing their judgements because of their discussions. Despite this being seemingly straightforward, there are significant challenges that have important implications for undertaking calibration. The first challenge is that as relative academic standards are resistant to precise articulation and usually only visible in their embodied form (Bloxham and Boyd 2012), social moderation and calibration activities need to be grounded in conversations in relation to specific pieces of student work. This means that moderation discussions are almost always based on a single assessment item or type (e.g. report, presentation or essay) and usually in a selected topic area within a particular subject or sub-discipline within the subject area. The paradox here is the desire to develop 'calibrated academics' (Sadler 2013) within contexts where continuous calibration for different topics and assessment tasks is unfeasible. Therefore, if and how academics translate calibrated standards from one context to others is a significant consideration. Another issue that is brought into sharp focus through calibration is the role and influence of disciplinary contexts and the expertise of individuals. This is particularly the case for multidisciplinary subjects, for example in sport and exercise science, where the ways of thinking and practicing (as defined by McCune and Hounsell 2005) between the sub-disciplines of psychology and biomechanics are starkly different, or in business studies between organisational studies and accounting. Those marking and moderating work will have variable levels of expertise in relation to the specific assessment in hand. Interestingly, research has shown that the more expert the marker the more likely they are to be critical and hold higher expectations of standards (Grainger, Adie, and Weir 2016). All these issues have significant implications for both the consistency of application and portability of standards.

What individuals take from moderation and calibration activities and the extent to which they influence future independent judgements is largely unknown. The idea of 'conceding ground' and 'moving toward' the standards of others is an important message during moderation events where the goal is consensual agreement. But is consensus possible even in tight knit communities? Habermas convincingly describes the conditions under which true consensus can be attained, namely a context in which consensus is free from power, strategic intentions and where all participants have the same opportunities to contribute and confer identical meanings on 'the expressions employed' (Habermas 1996, 19). Habermas himself suggests that this state is never fully met in the real world, and, we would argue, clearly unattainable in the moderation of relative standards. Power relationships, particularly with reference to academics on sessional or casual contracts, can further complicate and undermine the process of achieving consensus (Adie, Lloyd, and Beutel 2013; Mason, Roberts, and Flavell 2023). Academics on short-term casual contracts tend to defer to those with more experience and status and who may have influence on their future employment (Grainger, Adie, and Weir 2016; Mason, Roberts, and Flavell 2023). However, in time-poor higher education praxis, the goal of social moderation and calibration processes may not always be to achieve consensus on standards but to reduce marking variation. This is a subtle shift of focus that moves away from achieving agreement on standards per se to understanding and learning from different perspectives and enabling compromise. Moss and Schutz (2001) suggest that understanding the alternate perspectives that lead to *dissensus* is in itself valuable and more achievable than attaining consensus. They propose that useful discussions may not be predicated on arguing against others but rather questioning the premise of other perspectives with the aim of bringing

out and learning from their strengths, resulting in more textured and inclusive understandings (Moss and Schutz 2001). However, up to now research has not focused on the nature and finer grained details of the dialogues that social moderation and calibration activities in higher education generate. While there are some school-based studies (e.g. Smaill 2020; Wyatt-Smith, Klenowski, and Gunn 2010), which employ observational methodologies that include an analysis of teacher social moderation dialogue based on recordings, there is as yet insufficient research evidence that illuminates the processes at play in social moderation and calibration in higher education. Our own experience of facilitating calibration sessions in different disciplines has shown how difficult it is to identify appropriate exemplars, to instigate dialogue in which concrete exemplification is actively used to illustrate and agree on a shared understanding of standards, and to achieve genuine consensus.

There are yet more challenges as it is not just the articulation of, and agreement on, levels of performance that are difficult to attain. Further issues arise when an assessor translates the level and quality of a piece of work into numerical marks. Shay (2004, 323) argues that assessment judgements are 'context-dependent, experience-based and situational', Bloxham, Boyd, and Orr (2011) concur, stating that they are strongly informed by, the often tacit, standards of their academic community. For example, a mark of 64% in many American college contexts may not indicate an above average standard of achievement, but in the UK in many subject contexts it would be considered indicative of an above average piece of work (O'Donovan, Price, and Rust 2004). Such variations in perception of what marks should be given for certain levels of performance can occur locally within institutions or academic departments. Indeed, many institutions put forward mark profiles for units of study that are considered acceptable, averages above or below which prompt investigation and discussion. Essentially, marks do not represent statistical certainties; they take on meaning depending on their use and how they fit into social and organisational praxis (Spender 1996).

In summary, the evidence case for calibration and social moderation in terms of improving consistency of standards is extremely limited. Even with the necessary expertise, time and resources social moderation and calibration are not predictable or precise. Both research and our practical experience of facilitating calibration sessions highlight the significant challenges in sharing tacitly held understandings of attributes of quality (assessment criteria) and levels of performance (standards) particularly across different subjects and assessment contexts. All of which is further thwarted by the varying expertise and power relationships of assessors and the complexities inherent in the facilitation and development of activities and artefacts able to capture and share standards.

Discussion and practical recommendations

The focus of this article is the calibration of relative standards in educational settings within higher education's classical assessment system. Within this system, the dominant practice is the summative assessment of student achievement involving percentage marking or grading within discrete modules or units of study. Many experts have suggested alternative systems to enhance assessment validity and reliability, and indeed, student learning. Suggestions include relinquishing percentage marking or multi-step grading systems in favour of less granular (see for instance Rust 2011) and/or progamme-based assessment regimes (see for instance van der Vleuten et al. 2012). We acknowledge the qualities of these alternative systems and their potential to reduce marking inconsistency. However, as van der Vleuten et al. (2012) acknowledge in the context of relative standards at some point professional or subjective judgements on the quality of student achievement must be made. What follows are recommendations for the calibration of such subjective judgements in relation to relative standards in classical assessment systems.

Reaching a common understanding of terms through professional development

Here, we concede that it is easier to lay out the significant issues involved in the uniform representation, measurement, and calibration of relative academic standards than finding solutions to the

issues! However, improved practice would be supported by assessors 'who really understand what they are doing and for which purpose' (van der Vleuten et al., 2012, 212). The lack of a common vocabulary and evidence-based understandings along with the private character of marking has given rise to common misconceptions and beliefs about assessment criteria and standards (Price et al. 2010) and the associated use of numbers (Rust 2011). The 5-year Degree Standards Project in the UK funded by the Office for Students (AdvanceHE n.d.) has made some headway in the professional development of UK external examiners whose role is to evaluate the standards applied in other institutions. Their recommendations based on expert opinion and experience include emphasising that it is a key responsibility of subject communities to gain a common understanding of assessment standards through assessment calibration. Calibration should not be considered a one-off event but as an ongoing aspect of staff development with a need for members to 'maintain standing' and ensure their standards are representative of the subject community. Ideally, they suggest calibration activities should be separate from marking processes enabling time to reflect, discuss and come to consensus. Building on research undertaken in Australia (Watty et al. 2014) rather than agreeing the mark for a particular piece of work, findings from the Degree Standards Project suggest it is more effective to seek consensus on the level of attainment of specific attributes or criteria as this can then be transferred or translated across multiple assessment tasks (AdvanceHE n.d.). However, further empirical research is needed on calibration processes in action, as well as professional development for other key stakeholders particularly policymakers, assessors, and indeed, students. A practical starting point for those involved in assessment would be to clarify terms and acknowledge that not all standards are the same. Relative standards are never easily codified nor measured and more consistent marking in these contexts arguably will always require social moderation exercises involving the ranking and discussion of the relative merits of student work.

Recognising the limitations of codification

Secondly, those involved in the practice of, and policies on, assessment design would benefit from recognising the limitations of what can be achieved in terms of consistency in the marking of complex, open-ended university assessments without compromising task validity or overburdening resources. There are unavoidable trade-offs between reliable and consistent marking with the validity of assessment tasks and resource requirements (Stobart 2008). This may be particularly relevant in the assessment of professional competencies. For example, a well-written and researched essay on patient management authored by a student nurse may evidence 'know how' but not the achievement of a learning outcome that seeks evidence of clinical competence. Indeed, medicine and healthcare have long-used simulations, OSCEs, workplace-based assessments to demonstrate clinical competence as proposed by Miller in his pyramid of clinical competence (Miller 1990). The same argument can be made for competency-based assessments in engineering, architecture, and indeed, business and management, anthropology, sports science and so on. However, such simulation and practice-based assessments are rare and expensive to orchestrate and cannot always validly assess cognitive skills such as diagnostic reasoning (Witheridge, Ferns, and Scott-Smith 2019). The assessment task clearly has to be capable of validly demonstrating achievement of relevant learning outcome. But as van der Vleuten et al. (2012, 207) argue there is a tension between reliability and validity, whilst we may be able to reduce the subjectivity in assessment judgements 'if we try to achieve complete objectivation we will only trivialise the assessment process'.

The responsibility of higher education institutions to make transparent awarding decisions based on robust assessment of valid learning outcomes is of importance. Consequently, there is value in explicitly articulating academic standards and criteria, however, the limitations of this practice need to be understood. Within higher education, codification usually takes the form of an assessment rubric in which performance levels are broken down against individual pre-determined assessment criteria in the hope of increasing the objectivity of assessment decisions. However, even when

carefully constructed, in the context of relative standards rubrics are subject to interpretation by individual assessors and academics who ascribe meaning to terms based on their own individual experiences (Ajjawi, Bearman, and Boud 2021) and values (Shay 2008). Assessors can struggle to fully articulate their expectations for the work, and students therefore find it difficult to understand tutor expectations (Bloxham et al. 2016; O'Donovan, Price, and Rust 2004). Attempting to address this challenge through evermore detailed explanation that seeks to deconstruct academic judgements into more finely grained parts throws up yet more issues. Codifications become unwieldly, requiring more time and resources for their construction and are less transferable to other contexts. As Yorke (2002) states there is a tension between precision and utility in the creation of rubrics. There are also valid concerns that students' capacity for independent thought and self-regulation may be undermined as they become increasingly dependent on explicit guidance and seek to interrogate tutors more and more so that they can follow an explicit and precise 'paint-by-numbers' route map to higher marks (Torrance 2017). An increasingly frequent practice at many institutions has been to mitigate these issues by producing standardised assessment rubrics to be used across programmes, faculties or even institutions, in the hope that over time common interpretations will emerge. However, this ignores the issues of individual interpretation, the epistemic variation of assessment tasks and their associated academic standards, as well as possible unintended consequences including the erosion of task validity and student independence. Whilst laudable in the name of transparency, these efforts tend to focus on the development of rubrics as stand-alone artefacts capable of uniform interpretation rather than processes that support dialogue about their interpretation in relevant communities. This exemplifies the tension between what Broadfoot (2002, 157) describes as 'the scientific aspirations of assessment technologies to represent an objective reality and the unavoidable subjectivities injected by the human focus of these technologies'. There is benefit for all stakeholders in understanding and acknowledging the limitations of transparent codification. Indeed, if students understand the limitations their evaluation of assessment provision would be more valuable. We may do them a disfavour through propagating the myth that codified standards have an objective, fixed status. Indeed, we agree with Shay (2005, 677) that by exposing students to the contextually complex character of professional judgement, we 'prepare them for the kinds of rational thinking which their future professional contexts will require of them – decision-making which is relational, situational, pragmatic and value-based'. An important life-skill particularly in the brave new world of generative artificial intelligence.

Using codification to stimulate social moderation dialogue

Thirdly, despite all the issues, let us not 'throw the baby out with the bathwater'. A social material framing of standards as proposed by Ajjawi, Bearman, and Boud (2021) with a focus on both developing codifications where and as best we can, but without assuming uniform interpretation, and therefore incorporating social moderation processes, is a logical way forward and one that embraces legitimate variation. The problem is not the codifications themselves but as Shay (2005) suggests failing to use them as catalysts for dialogue about what we really value as assessors, individually and as communities of disciplinary practice, and how we negotiate and communicate differences and shifts in what we value. In the context of university assessments where relative standards predominate it is important to develop shared understandings through the deployment of systematic moderation and calibration policies and practices involving communal dialogue, particularly at key points of dissonant interpretation where inconsistencies commonly arise. These include: (i) interpretations of criteria and levels of performance; (ii) understandings on the nature of the academic standards at play (relative or sharp); (iii) the marking practices in actual use such as criterion and/or norm referencing; (iv) the numeric codification of levels of performance both within and across study levels. Calibration activities can be set up to foreground each of these points of interpretation in turn, although in practice they are often interrelated.

Identifying variation and dissensus

Fourthly, in terms of calibration activities centred on developing shared understanding of academic standards, as well as seeking consensus we would do well to attempt to identify points of variation and dissensus. In a university course there is likely to be quite stark epistemic differences between modules. At a simple level, many degrees focus on declarative knowledge and technique in their initial stages. For example, in an accounting degree students may initially learn accounting terms and techniques and only later focus on more contestable areas of knowledge where contextual interpretation and meaning takes centre stage. Answers are not so simple, not correct or incorrect, but better or worse, more or less convincing. Beyond this, there are inherent subject and epistemic differences between modules that make it difficult for academics affiliated to a course to agree on what counts as knowledge and why and how it is recognised (O'Donovan 2019). In conventional approaches the search for consensus still dominates. However, the acknowledgement and exploration of dissimilarities may help protect us from the false assurance of an articulated consensus and uniformity (Moss and Schutz 2001). Such a process would also help academics affiliated to individual modules to gain a more holistic knowledge of a course and its variations. Perhaps most importantly it would legitimise variation not just in terms of the nature of knowledge and its ambiguity and uncertainty but also in the expectations of students as they progress through a course of study. Dissensus and variation is normally viewed as a problem to be overcome (Moss and Schutz 2001) and as such even legitimate variations are usually viewed by key stakeholders, including students, as sources of dissatisfaction (O'Donovan 2019), not as valuable alternative perspectives.

Recognising the importance of relative standards in the light of artificial intelligence

Finally, whilst a deep dive on the influence of generative artificial intelligence (AI) tools on academic standards is beyond the scope of this article it would be remiss to completely ignore it. The profound change that AI is having, and will increasingly have, on tertiary learning outcomes and their assessment has provoked much needed and overdue discussion. AI may not only provoke the use of more personal and experiential assessments where students can be observed in the act of doing, but it is also likely to fundamentally change the landscape of employment and the skills that societies value. If sourcing, synthesising and summarising information is easy to automate, then not only will AI be able to produce such assignments for students (and also be able to assess what is produced), but many degrees will develop some skills that are redundant (Dickinson 2023). Skills that are likely to be increasingly valued are those which are particularly challenging to consistently assess. These include skills such as negotiation, persuasion as well as those involving collaborative application, original creation and critique. Consequently, the higher education sector may be compelled to refocus on what counts as 'good work' provoking valuable discussion on the validity of assessment tasks and the nature of their attributes of quality. Achievement levels are likely to be matters of degree and interpretation and the need for ongoing dialogue on relative standards at significant points of discordant interpretation even greater.

Conclusion

A key responsibility of higher education institutions is the accurate certification of knowledge and skills, yet empirical research suggests we do not do this reliably or consistently. This variation poses a serious threat to public expectations of comparability. To remedy this situation, contemporary higher education focuses on quality processes, codification of academic standards and the standardisation of practices and policies. However, in this article we have shown that these simply do not work and are at odds with the varied nature of learning in the disciplines, the complex, open-ended tasks that are typical for higher education, and the fact that codifications of academic standards almost always require interpretation. Accordingly, this approach, on its own, is unlikely to achieve

consistency in the subjective judgement of relative standards within and across local communities. In turn, this fundamentally challenges the assumptions that underpin the award of degrees, the notion of equivalence and of fairness to students and other stakeholders. We suggest that progress is more likely to be made through further professional development and rebalancing our focus away from the development of explicit assessment codifications towards their more informed use supported by dialogic moderation and calibration processes orchestrated within subject communities Such activities would be further enhanced if centred on both the dissimilarities and commonalities of individual assessment perspectives at identified points of discordant interpretation in the assessment process. This paper pulls together the limited research that has been undertaken on the impact of social moderation and calibration in higher education assessment praxis. Larger scale empirical evidence drawn from a broader array of subject communities investigating how activities can be most effectively structured and deployed is clearly needed. Such research may prompt changes to practice that are likely to require additional time and resources. However, higher education institutions are tasked to provide reliable and valid assessment, and as van der Vleuten et al. (2012, 211) quote 'if you think education is expensive, try ignorance'.

Note

1. For instance, in an essay assignment such attributes of quality may be the 'use of literature' or 'critical evaluation'.

Acknowledgements

We are indebted to the team of assessment specialists that contributed to the UK Degree StandardsProject (https://www.advance-he.ac.uk/degree-standards-project) whose discussions stimulated this article.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Berry O'Donovan http://orcid.org/0000-0001-9756-1522
Ian Sadler http://orcid.org/0000-0002-1636-2052
Nicola Reimann http://orcid.org/0000-0002-2674-2761

References

Adie, L., M. Lloyd, and D. Beutel. 2013. "Identifying Discourses of Moderation in Higher Education." *Assessment & Evaluation in Higher Education* 38 (8): 968–77. https://doi.org/10.1080/02602938.2013.769200

AdvanceHE. n.d. "The Degree Standards Project." AdvanceHE. https://www.advance-he.ac.uk/degree-standards-project. Accessed July 24, 2023.

Ajjawi, R., and M. Bearman. 2018. "Problematising Standards: Representation or Performance." In *Developing Evaluative Judgement in Higher Education: Assessment for Knowing and Producing Quality Work*, edited by D. Boud, R. Ajjawi, P. Dawson, and J. Tai, 41–50. Abingdon: Routledge.

Ajjawi, R., M. Bearman, and D. Boud. 2021. "Performing Standards: A Critical Perspective on the Contemporary Use of Standards in Assessment." *Teaching in Higher Education* 26 (5): 728–41. https://doi.org/10.1080/13562517.2019.1678579

Beutel, D., L. Adie, and M. Lloyd. 2017. "Assessment Moderation in an Australian Context: Processes, Practices, and Challenges." *Teaching in Higher Education* 22 (1): 1–14. https://doi.org/10.1080/13562517.2016.1213232

Biesta, G. 2008. "Good Education in an Age of Measurement: On the Need to Reconnect with the Question of Purpose in Education." *Educational Assessment, Evaluation and Accountability* 21: 33–46. https://doi.org/10.1007/s11092-008-9064-9

Bloxham, S. 2009. "Marking and Moderation in the UK: False Assumptions and Wasted Resources." *Assessment & Evaluation in Higher Education* 34 (2): 209–220. https://doi.org/10.1080/02602930801955978

Bloxham, S., and P. Boyd. 2012. "Accountability in Grading Student Work: Securing Academic Standards in a Twenty-First Century Quality Assurance Context." *British Educational Research Journal* 38 (4): 615–34. https://doi.org/10.1080/01411926.2011.569007

Bloxham, S., P. Boyd, and S. Orr. 2011. "Mark My Words: The Role of Assessment Criteria in UK Higher Education Grading Practices." *Studies in Higher Education* 36 (6): 655–70. https://doi.org/10.1080/03075071003777716

Bloxham, S., B. den-Outer, J. Hudson, and M. Price. 2016. "Let's Stop the Pretence of Consistent Marking: Exploring the Multiple Limitations of Assessment Criteria." *Assessment & Evaluation in Higher Education* 41 (3): 466–81. https://doi.org/10.1080/02602938.2015.1024607

Bloxham, S., J. Hudson, B. den Outer, and M. Price. 2015. "External Peer Review of Assessment: An Effective Approach to Verifying Standards?" *Higher Education Research & Development* 34 (6): 1069–82. https://doi.org/10.1080/07294360.2015.1024629

Bloxham, S., and M. Price. 2015. "External Examining: Fit for Purpose?" *Studies in Higher Education* 40 (2): 195–211. https://doi.org/10.1080/03075079.2013.823931

Boyd, P., and S. Bloxham. 2014. "A Situative Metaphor for Teacher Learning: The Case of University Tutors Learning to Grade Student Coursework." *British Educational Research Journal* 40 (2): 337–52. https://doi.org/10.1002/berj.3082

Brandenburg, R., A. Fletcher, A. Gorriss-Hunter, C. Van der Smee, W. Holcombe, K. Griffiths, and K. Schneider. 2023. "'More Than Marking and Moderation': A Self-Study of Teacher Educator Learning Through Engaging with Graduate Teaching Performance Assessment." *Studying Teacher Education* 19 (3): 330–50. https://doi.org/10.1080/17425964.2022.2164761

Broadfoot, P. 2002. "Dynamic Versus Arbitrary Standards: Recognising the Human Factor in Assessment." *Assessment in Education* 9 (2): 157–9.

Crimmins, G., G. Nash, F. Oprescu, K. Alla, G. Brock, B. Hickson-Jamieson, and C. Noakes. 2016. "Can a Systematic Assessment Moderation Process Assure the Quality and Integrity of Assessment Practice While Supporting the Professional Development of Casual Academics?" *Assessment & Evaluation in Higher Education* 41 (3): 427–41. https://doi.org/10.1080/02602938.2015.1017754

Delandshere, G. 2001. "Implicit Theories, Unexamined Assumptions and the Status Quo of Educational Assessment." *Assessment in Education* 8 (2): 113–33.

Dickinson, J. 2023. "An Avalanche Really is Coming This Time." Wonkhe. https://wonkhe.com/blogs/an-avalanche-really-is-coming-this-time/. Accessed March 17.

Gillis, S. 2023. "Ensuring Comparability of Qualifications Through Moderation: Implications for Australia's VET Sector." *Journal of Vocational Education & Training* 75 (2): 349–71. https://doi.org/10.1080/13636820.2020.1860116

Grainger, P., L. Adie, and K. Weir. 2016. "Quality Assurance of Assessment and Moderation Discourses Involving Sessional Staff." *Assessment & Evaluation in Higher Education* 41 (4): 548–59. https://doi.org/10.1080/02602938.2015.1030333

Habermas, J. 1996. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Translated by William Rehg. Cambridge, MA: The MIT Press.

Hancock, P., M. Freeman, A. Abraham, P. De Lange, B. Howieson, B. O'Connell, and K. Watty. 2015. *Achievement Matters: External Peer Review of Accounting Learning Standards*. Australian Government, , Sydney, NSW, Office for Learning and Teaching, Department of Education and Training.

Hanlon, J., M. Jefferson, M. Molan, and B. Mitchell. 2004. *An Examination of the Incident of 'Error Variation' in the Grading of law Assessments*. University of Warwick, United Kingdom Centre for Legal Education (UKCLE).

Harvey, L. 2005. "A History and Critique of Quality Evaluation in the UK." *Quality Assurance in Education* 13 (4): 263–76. https://doi.org/10.1108/09684880510700608

Herridge, M., J. Tashiro, and V. Talanquer. 2021. "Variation in Chemistry Instructors' Evaluations of Student Written Responses and its Impact on Grading." *Chemistry Education Research and Practice* 22 (4): 948–72. https://doi.org/10.1039/D1RP00061F

Hill, D., C. Lewis, and A. Maisuria. 2015. "Neoliberal and Neoconservative Immiseration Capitalism in England: Policies and Impacts on Society and on Education." *Journal for Critical Education Policy Studies* 13 (2): 38–82.

Krause, K. L., G. Scott, K. Aubin, H. Alexander, T. Angelo, S. Campbell, M. Carroll, et al. 2014. *Assuring Learning and Teaching Standards Through Inter-Institutional Peer Review and Moderation: Final Report of the Project: A Sector Wide Model For Assuring Final Year Subject and Program Achievement Standards Through Inter-University Moderation*. Canberra: Office for Learning and Teaching.

Lambert, H. 2019. "The Great University Con: How the British Degree Lost Its Value. The New Statesman." August 21. https://www.newstatesman.com/politics/2019/08/the-great-university-con-how-the-british-degree-lost-its-value.

Mason, J., and L. D. Roberts. 2023. "Consensus Moderation: The Voices of Expert Academics." *Assessment & Evaluation in Higher Education* 48 (7): 926–37. https://doi.org/10.1080/02602938.2022.2161999

Mason, J., L. D. Roberts, and H. Flavell. 2022. "A Foucauldian Discourse Analysis of Unit Coordinators' Experiences of Consensus Moderation in an Australian University." *Assessment & Evaluation in Higher Education* 47 (8): 1289–1300. https://doi.org/10.1080/02602938.2022.2064970

Mason, J., L. D. Roberts, and H. Flavell. 2023. "Consensus Moderation and the Sessional Academic: Valued or Powerless and Compliant?" *International Journal for Academic Development* 28 (4): 468–80. https://doi.org/10.1080/1360144X.2022.2036156

McCune, V., and D. Hounsell. 2005. "The Development of Students' Ways of Thinking and Practising in 3 Final-Year Biology Courses." *Higher Education* 49 (3): 255–89. https://doi.org/10.1007/s10734-004-6666-0

Miller, G. E. 1990. "The Assessment of Clinical Skills/Competence/Performance." *Academic Medicine* 65 (9): 63–7. https://doi.org/10.1097/00001888-199009000-00045

Mok, K. H. 2016. "Massification of Higher Education, Graduate Employment and Social Mobility in the Greater China Region." *British Journal of Sociology of Education* 37 (1): 51–71. https://doi.org/10.1080/01425692.2015.1111751

Moss, P. A., and A. Schutz. 2001. "Educational Standards, Assessment and the Search for Consensus." *American Educational Research Journal* 38 (1): 37–70. https://doi.org/10.3102/00028312038001037

Naidoo, R., and J. Williams. 2015. "The Neoliberal Regime in English Higher Education: Charters, Consumers and the Erosion of the Public Good." *Critical Studies in Education* 56 (2): 208–23. https://doi.org/10.1080/17508487.2014.939098

O'Byrne, D., and C. Bond. 2014. "Back to the Future: The Idea of a University Revisited." *Journal of Higher Education Policy and Management* 36 (6): 571–84. https://doi.org/10.1080/1360080X.2014.957888

O'Connell, B., P. De Lange, M. Freeman, P. Hancock, A. Abraham, B. Howieson, and K. Watty. 2016. "Does Calibration Reduce Variability in the Assessment of Accounting Learning Outcomes?" *Assessment & Evaluation in Higher Education* 41 (3): 331–49. https://doi.org/10.1080/02602938.2015.1008398

O'Donovan, B. 2019. "Patchwork Quilt or Woven Cloth? The Student Experience of Coping with Assessment Across Disciplines." *Studies in Higher Education* 44 (9): 1579–90. https://doi.org/10.1080/03075079.2018.1456518

O'Donovan, B., M. Price, and C. Rust. 2004. "Know What I Mean? Enhancing Student Understanding of Assessment Standards and Criteria." *Teaching in Higher Education* 9 (3): 325–35. https://doi.org/10.1080/1356251042000216642

O'Hagan, S. R., and G. Wigglesworth. 2015. "Who's Marking my Essay? The Assessment of non-Native-Speaker and Native-Speaker Undergraduate Essays in an Australian Higher Education Context." *Studies in Higher Education* 40 (9): 1729–47. https://doi.org/10.1080/03075079.2014.896890

Palermo, C., E. Volders, S. Gibson, M. Kennedy, A. Wray, J. Thomas, M. Hannan-Jones, D. Gallegos, and E. Beck. 2018. "Exploring Approaches to Dietetic Assessment of a Common Task Across Different Universities Through Assessment Moderation." *Journal of Human Nutrition and Dietetics* 31 (1): 41–6. https://doi.org/10.1111/jhn.12499

Pitts, J., C. Coles, P. Thomas, and F. Smith. 2002. "Enhancing Reliability in Portfolio Assessment: Discussions Between Assessors." *Medical Teacher* 24 (2): 197–201. https://doi.org/10.1080/01421590220125321

Price, M., and J. Carroll, O'Donovan, B., and Rust, C. 2010. "If I was Going There I Wouldn't Start from Here: A Critical Commentary on Current Assessment Practices." *Assessment & Evaluation in Higher Education* 36 (4): 479–92. https://doi.org/10.1080/02602930903512883

Rinne, I. 2024. "Same Grade for Different Reasons, Different Grades for the Same Reason?" *Assessment & Evaluation in Higher Education* 49 (2): 220–232.

Rust, C. 2011. "The Unscholarly Use of Numbers in Our Assessment Practices: What Will Make Us Change?" *International Journal for the Scholarship of Teaching and Learning* 5 (1): 4. https://doi.org/10.20429/ijsotl.2011.050104

Sadler, D. R. 1987. "Specifying and Promulgating Achievement Standards." *Oxford Review of Education* 13 (2): 191–209. https://doi.org/10.1080/0305498870130207

Sadler, D. R. 2013. "Assuring Academic Achievement Standards: From Moderation to Calibration." *Assessment in Education: Principles, Policy & Practice* 20 (1): 5–19. https://doi.org/10.1080/0969594X.2012.714742

Sadler, D. R. 2014. "The Futility of Attempting to Codify Academic Achievement Standards." *Higher Education* 67 (3): 273–88. https://doi.org/10.1007/s10734-013-9649-1

Shay, S. 2004. "The Assessment of Complex Performance: A Socially Situated Interpretive Act." *Harvard Educational Review* 74 (3): 307–29. https://doi.org/10.17763/haer.74.3.wq16l67103324520

Shay, S. 2005. "The Assessment of Complex Tasks: A Double Reading." *Studies in Higher Education* 30 (6): 663–79. https://doi.org/10.1080/03075070500339988

Shay, S. 2008. "Beyond Social Constructivist Perspectives on Assessment: The Centring of Knowledge." *Teaching in Higher Education* 13 (5): 595–605. https://doi.org/10.1080/13562510802334970

Smaill, E. 2020. "Using Involvement in Moderation to Strengthen Teachers' Assessment for Learning Capability." *Assessment in Education: Principles, Policy & Practice* 27 (5): 522–43. https://doi.org/10.1080/0969594X.2020.1777087

Spender, J. C. 1996. "Organisational Knowledge, Learning and Memory: Three Concepts in Search of a Theory." *Journal of Organizational Change Management* 9 (1): 63–78. https://doi.org/10.1108/09534819610156813

Stobart, G. 2008. *Testing Times: The Uses and Abuses of Assessment*. Abingdon: Routledge.

Stolpe, K., L. Björklund, M. Lundström, and M. Åström. 2021. "Different Profiles for the Assessment of Student Theses in Teacher Education." *Higher Education* 82 (5): 959–76. https://doi.org/10.1007/s10734-021-00692-w

Torrance, H. 2017. "Blaming the Victim: Assessment, Examinations, and the Responsibilisation of Students and Teachers in Neo-Liberal Governance." *Discourse: Studies in the Cultural Politics of Education* 38 (1): 83–96. https://doi.org/10.1080/01596306.2015.1104854

Tsoukas, H. 2003. "Do We Really Understand Tacit Knowledge?" In *Handbook of Organisational Learning and Knowledge Management*, edited by M. Easterby Smith and M. Lyles, 410–27. Cambridge, MA: Blackwell.

Van Der Vleuten, C., L. Schuwirth, W. Driessen, J. Dijkstra, D. Tigelaar, L. Baartman, and J. Van Tartwijk. 2012. "A model for programmatic assessment fit for purpose." *Med Teach* 34 (3): 205–214.

Watty, K., M. Freeman, B. Howieson, P. Hancock, B. O'Connell, P. de Lange, and A. Abraham. 2014. "Social Moderation, Assessment and Assuring Standards for Accounting Graduates." *Assessment & Evaluation in Higher Education* 39 (4): 461–78. https://doi.org/10.1080/02602938.2013.848336

Witheridge, A., G. Ferns, and W. Scott-Smith. 2019. "Revisiting Miller's Pyramid in Medical Education: The Gap between Traditional Assessment and Diagnostic Reasoning." *International Journal of Medical Education* 10: 191–92. https://doi.org/10.5116/ijme.5d9b.0c37

Wyatt-Smith, C., V. Klenowski, and S. Gunn. 2010. "The Centrality of Teachers' Judgement Practice in Assessment: A Study of Standards in Moderation." *Assessment in Education: Principles, Policy & Practice* 17 (1): 59–75. https://doi.org/10.1080/09695940903565610

Yorke, M. 2002. "Subject Benchmarking and the Assessment of Student Learning." *Quality Assurance in Education* 10 (3): 155–71. https://doi.org/10.1108/09684880210435921