

Developing evaluative judgement: enabling students to make decisions about the quality of work

Joanna Tai¹ · Rola Ajjawi¹ · David Boud^{1,2,3} ·
Phillip Dawson¹ · Ernesto Panadero^{1,4}

Published online: 16 December 2017

© The Author(s) 2017. This article is an open access publication

Abstract Evaluative judgement is the capability to make decisions about the quality of work of oneself and others. In this paper, we propose that developing students' evaluative judgement should be a goal of higher education, to enable students to improve their work and to meet their future learning needs: a necessary capability of graduates. We explore evaluative judgement within a discourse of pedagogy rather than primarily within an assessment discourse, as a way of encompassing and integrating a range of pedagogical practices. We trace the origins and development of the term 'evaluative judgement' to form a concise definition then recommend refinements to existing higher education practices of self-assessment, peer assessment, feedback, rubrics, and use of exemplars to contribute to the development of evaluative judgement. Considering pedagogical practices in light of evaluative judgement

✉ Joanna Tai
joanna.tai@deakin.edu.au

Rola Ajjawi
rola.ajjawi@deakin.edu.au

David Boud
david.boud@deakin.edu.au

Phillip Dawson
p.dawson@deakin.edu.au

Ernesto Panadero
ernesto.panadero@uam.es

¹ Centre for Research in Assessment and Digital Learning, Deakin University, Geelong, Australia

² Faculty of Arts and Social Sciences, University of Technology Sydney, Sydney, Australia

³ Work and Learning Research Centre, Middlesex University, London, UK

⁴ Developmental and Educational Psychology Department, Faculty of Psychology, Universidad Autónoma de Madrid, Madrid, Spain

may lead to fruitful methods of engendering the skills learners require both within and beyond higher education settings.

Keywords Assessment for learning · Sustainable assessment · Peer assessment · Feedback · Graduate attributes · Self-regulated learning

Introduction

How does a student come to an understanding of the quality of their efforts and make decisions on the acceptability of their work? How does an artist know which of their pieces is good enough to exhibit? How does an employee know whether their work is ready for their manager? Each of these and other common scenarios require more than the ability to do good work; they require appraisal and an understanding of standards or quality. This capability of ‘evaluative judgement’, or being able to judge the quality of one’s own and others’ work, is necessary not just in a student’s current course but for learning throughout life (Boud and Soler 2016). It encapsulates the ongoing interactions between the individual, their fellow students or practitioners, and standards of performance required for effective and reflexive practice.

However, current assessment and feedback practices do not necessarily operate from this perspective. Assessment design is frequently critiqued for being unidirectional, excessively content and task focused, while also positioning students as passive recipients of feedback information (e.g. Carless et al. 2011). Worse than failing to support the development of evaluative judgement, these approaches may even inhibit it, by producing graduates dependent on others’ assessment of their work, who are not able to identify criteria to apply in any given context. This may be because evaluative judgement is undertheorised and under-researched. In particular, although there have been some theoretical arguments about its importance (e.g. Sadler 2010), evaluative judgement has not been the focus of sustained attention and very little empirical work exists on the effectiveness of strategies for developing students’ evaluative judgement (Nicol 2014; Tai et al. 2016). Sadler (2013) observed that ‘quality is something I do not know how to define, but I recognise it when I see it’ (p. 8), a common experience for those dealing with abstract concepts. This perceived inability to articulate what quality is has likely impacted on our ambitions and capability to directly educate learners about quality.

Studies in the field of human judgement have brought to bear better understandings regarding our own fallibility in evaluating or judging student work (Brooks 2012). We may default to ‘quick thinking’, assessing intuitively rather than by thoughtful deliberation and employ referents for judgements which exist outside published criteria. Heuristics and biases are also likely to influence decision-making (Joughin et al. 2016). Given there are strategies to assist markers to develop their judgement (Sadler 2013), we may also be able to help learners develop strategies for refining their own judgement (i.e. surfacing potentially implicit quality criteria and comparing these to the espoused ones), through a range of activities and tasks which requires students to examine and interact with examples of work of varying quality.

In this paper, we trace the elaboration of evaluative judgement and related concepts in the literature and converge on a definition. We then discuss five pedagogical approaches that may support the development of evaluative judgement, and how they could be improved if evaluative judgement were a priority. We conclude with implications for future agendas in research and practice to understand and develop evaluative judgement. We argue that evaluative judgement is a necessary, distinct concept that has not been given sufficient attention in

higher education research or practice. We propose that evaluative judgement underpins and integrates existing pedagogies such as self-assessment, peer assessment, and the use of exemplars. We suggest that the development of evaluative judgement is one of higher education's ultimate goals and a necessary capability for graduates.

Tracing the evolution of 'evaluative judgement'

The origin of evaluative judgement can be traced back to Sadler's (1989) ideas of 'evaluative knowledge' (p 135), or 'evaluative expertise' (p 138), which students must develop to become progressively independent of their teachers. Sadler proposed that students needed to understand criteria in relation to the standards required for making quality judgements, before being able to appreciate feedback about their performance. Students then also required the ability to engage in activity to close the gap between their performance and the standards that were expected to be achieved. These proposals were specifically made in relation to the role of formative assessment in learning design: formative assessments were to aid students in understanding how complex judgements were made and to allow students to have direct, authentic experiences of evaluating others and being evaluated (Nicol and Macfarlane-Dick 2006). The overall goal was to enable students to develop and rely on their own evaluative judgements so that they can become independent students and eventually effective practitioners.

The term 'sustainable assessment' was coined to indicate a purpose of assessment that had not been sufficiently encompassed by the previous dichotomy of the summative and the formative. Sustainable assessment was proposed as assessment 'that meets the needs of the present and [also] prepares students to meet their own future learning needs' (Boud 2000, p. 151). This was posited as a distinct purpose of assessment: summative assessment typically met the needs of other stakeholders (i.e. for grade generation and certification), and formative assessment was commonly limited to tasks within a particular course or course unit. Sustainable assessment focused on building the capacity of students to make judgements of their own work and that of others through engagement in a variety of assessment activities so that they could be more effective learners and meet the requirements of work beyond the point of graduation.

Similarly, ideas of 'informed judgement' also sat firmly within the assessment literature. In *Rethinking Assessment in Higher Education*, Boud (2007) sought to develop an inclusive view of student assessment, locating it as a practice at the disposal of both teachers and learners rather than simply an act of teachers. To this end, assessment was taken to be any activity that informed the judgement of any of the parties involved. Those being informed included learners; many discussions of assessment positioned learners as subjects of assessments by others, or as recipients of assessment information generated by others. Assessment as informing judgement was designed to acknowledge that assessment information is vital for students in planning their own learning and that assessment activities should be established on the assumption that learning was important. As a part of this enterprise, Boud and Falchikov (2007, pp. 186–190) proposed a framework for developing assessment skills for future learning. These steps towards promoting students' informed judgement from the perspective of the learner consisted of the following: (a) identifying oneself as an active learner, (b) identifying one's level of knowledge and the gaps in this, (c) practising testing and judging, (d) developing these skills over time, and (e) embodying reflexivity and commitment.

Returning to pedagogical practices, Nicol and Macfarlane-Dick (2006) interpreted Sadler's ideas around evaluative knowledge to mean that students 'must already possess some of the same evaluative skills as their teacher' (p. 204). Through making repeated judgements about the quality of students' work, teachers come to implicitly understand the necessary standards of competence. These understandings of quality (implicit or explicit) and striving to attain them are necessary precursors of expertise. Yet, this understanding had not previously been expected of students in such a nuanced and explicit way.

Sadler (2010) revisited the idea of evaluative judgement to further explore what surrounds acts of feedback. Here, he proposed that student capability in 'complex appraisal' was necessary for learning from feedback and that educators needed to do more to articulate the reasons for the information provided (i.e. quality notions) rather than providing information. Sadler also suggested that an additional method for developing this capability was peer review or assessment, which was also taken up by Gielen et al. (2011). By engaging students in making judgements, and interacting with criteria, it was thought that understandings of quality and tacit knowledge (which educators already held) would be developed. Nicol (2014) continued this argument and provided case studies to illustrate the power of peer review in developing evaluative capacity.

In parallel with the development of these ideas, the wider context of higher education had been changing internationally. The framing of courses and assessment practices around explicit learning outcomes and judging students' work in terms of criteria and standards had become increasingly accepted (Boud 2017). This move towards a standard-based curriculum provided a stronger base for the establishment of students developing evaluative judgement than hitherto, because having explicit standards helps students to clarify their learning goals and what they need to master.

We suggest that the authors identified in this section were discussing the same fundamental idea: that students must gain an understanding of quality and how to make evaluative judgements, so that they may operate independently on future occasions, taking into account all forms of information and feedback comments, without explicit external direction from a teacher or teacher-like figure. This also aligns with much of the ongoing discussions in higher education focused on preparing students for professional practice, which is sometimes couched in terms of an employability agenda (Knight and Yorke 2003; Tomlinson 2012). Employers have criticised for decades the underpreparedness of graduates for the workplace and the lack of impact of the various moves made to address this concern, e.g. inclusion of communication skills and teamwork in the curriculum (e.g. Thompson 2006). A focus on developing evaluative skills may lead to graduates who can identify what is needed for good work in any situation and what is needed for them to produce it.

This paper therefore makes an argument for evaluative judgement as an integrating and encompassing concept, part of curricular and pedagogical goals, rather than primarily an assessment concern. Evaluative judgement speaks to a broader purpose and the reason why we employ a range of strategies to facilitate learning: to develop students' notions of quality and how they might be identified in their actual work. Before we discuss practices which promote students' evaluative judgement, it is necessary to indicate how we are using this term.

Defining evaluative judgement

As a term, evaluative judgement has only relatively recently been taken up in the higher and professional education literature, as a higher-level cognitive ability required for life-long

learning (Cowan 2010). Ideas about evaluation and critical judgement being necessary for effective feedback, and therefore learning, have also been highlighted by Sadler (2010), Nicol (2013, 2014, 2014), and Boud and Molloy (2013). These scholars argue that evaluative judgement acknowledges the complexity of contextual standards and performance, supports the development of learning trajectories and mastery, and is therefore aimed at future capacities and lifelong learning.

Little empirical work has so far been conducted within an explicit evaluative judgement framework. Firstly, Nicol et al. (2014) demonstrated the ability of peer learning activities to facilitate students' judgement making in higher education settings. Secondly, Tai et al. (2016) also explored the role of informal peer learning in producing accurate evaluative judgements, which impacted on students' capacity to engage in feedback conversations, through a better understanding of standards. Thirdly, Barton et al. (2016) reframed formal feedback processes to develop evaluative judgement, including dialogic feedback, self-assessment, and a programmatic feedback journal.

Nevertheless, there is a wider literature on formative assessment that has not been framed under the evaluative judgement umbrella, but which can be repositioned to do so. This literature supports students developing capability for evaluative judgement either through strategies that involve self-assessment (e.g. Boud 1992; Panadero et al. 2016), peer assessment (e.g. Panadero 2016; Topping 2010), or self-regulated learning (e.g. Nicol and Macfarlane-Dick 2006; Panadero and Alonso-Tapia 2013). Although they do not use the term, these sets of studies share a common purpose with evaluative judgement: to enhance students' critical capability via a range of learning and assessment practices. We propose that a focus on developing evaluative judgement goes a step beyond these through providing a more coherent framework to anchor these practices while adding a further perspective taken from sustainable assessment (Boud and Soler 2016): equipping students for future practice. However, what are the features of evaluative judgement and how does it integrate these multiple agendas?

An earlier definition of evaluative judgement by Tai et al. (2016) was

'... the ability to critically assess a performance in relation to a predefined but not necessarily explicit standard, which entails a complex process of reflection. It has an internal application, in the form of self-evaluation, and an external application, in making decisions about the quality of others' work'. (p. 661)

This definition was formed in the context of medical students' learning on clinical placements. There are many variations of what 'work' is within the higher education setting, which are not limited to performance. Work may also constitute written pieces (essays and reports), oral presentations, creative endeavours, or other products and processes. Furthermore, many elements of the definition seemed tautological, and so, we present a simpler definition:

Evaluative judgement is the capability to make decisions about the quality of work of self and others

In alignment with Tai et al. (2016), there are two components integral to evaluative judgement which operate as complements to each other: firstly, understanding what constitutes quality and, secondly, applying this understanding through an appraisal of work, whether it be one's own, or another's. This second step could be thought of as the making of evaluative judgements, which is the means of exercising one's evaluative judgement. Through appraising work, an individual has to interact with a standard (whether implicit or explicit), potentially increasing their understandings of quality. The learning of quality and the learning of

judgement making are interdependent: when coupled, they provide a stronger justification for the development of both aspects, in being prepared for future work. While judgement making occurs frequently and could be done without any attention to the development of itself as a skill, we consider that the process of developing evaluative judgement needs to be deliberate and be deliberated upon. Evaluative judgement itself, however, may become unconscious when the individual has had significant experience and expertise in making evaluative judgements in a specific area.

We propose that the development of evaluative judgements can be fruitfully considered using the staged theory of expertise of decision-making (Schmidt and Rikers 2007). This has been illustrated through the study of how medical students learn to make clinical decisions by seeing many different patients and making decisions about them. This involves both analytical and non-analytical processes; however, as novices, this process is initially labour intensive, requiring the integration of multiple strands of knowledge in an analytical fashion. Learning to make decisions occurs through actually having to make decisions, justification of the decisions in relation to knowledge and standards, and discussion with peers, mentors, and supervisors (Ajjawi and Higgs 2008). As they become more experienced, medical trainees come to ‘chunk’ knowledge through making links between biomedical knowledge and particular clinical signs and symptoms. These are chunked further to form patterns or illness scripts about particular diseases, presentations, and conditions. Thus, the process becomes less analytical, with analytical reasoning used when problems do not seem to fit an existing pattern, as a safety mechanism.

In a similar way, we hypothesise that students and teachers develop the equivalent of such scripts where knowledge about what makes for quality becomes linked through repeated exposure to work, consideration of standards (whether these are implicit or explicit) and the need to justify decisions (as assessors do post hoc (Ecclestone 2001)). These help strengthen judgements in relation to particular disciplinary genres and within communities of practice. This may explain why experienced assessors find marking of a familiar task less effortful: they have had sufficient experience to become more intuitive in their judgements based on previously formed patterns, but they still use explicit criteria and standards to rationalise their intuitive global professional judgement to others. Indeed in health, the communication of clinical judgements has been found to be a post hoc reformulation to suit the audience rather than an actual representation of the decision-making process (Ajjawi and Higgs 2012). Similarly, we suggest that students can develop their evaluative judgement through repeated practice. Making an evaluative judgement requires the activation of knowledge about quality in relation to a problem space. By repeating this process in different situations, students build their quality scripts for particular disciplinary assessment genres which in turn means they can make better decisions.

We note that evaluative judgement is contextual: what constitutes quality, and what a decision may look like, will depend on the setting in which the evaluative judgement is made. Evaluative judgement is domain specific: one develops expertise pertaining to a specific subject or disciplinary area, from which decisions about quality of work are made. However, the capability of identifying criteria of quality and applying these to work in context may cross domains. Evaluative judgement is also more than just an orientation to disciplinary, community, or contextual standards as applied to current instances of work, it can be carried over to future work too. There are also likely to be differences between work expected at an entry level and that conducted by experts. In areas in which it is possible to make practice explicit, the notion of learners calibrating their judgement against those of effective practitioners or other

sources of expertise is also important (Boud et al. 2015). To develop evaluative judgement, feedback information should perhaps focus less on the quality of students' work and more on the fidelity of their judgements about quality. Feedback is, however, but one of several practices which may develop evaluative judgement.

Current practices which develop evaluative judgement

There are several aspects of courses that already contribute to the development of students' evaluative ability. However, they may not previously have been seen from this perspective and their potential for this purpose not fully realised. Through the formal and informal curriculum (including assessment structures), learners are likely to develop evaluative judgement. There are a range of assessment-related activities, such as self and peer assessment, which are now imbued with greater purpose through their link with evaluative judgement.

The provision of assessment criteria is an attempt to represent standards to students. However, such representations are unlikely to communicate the complex tacit knowledge embedded in standards and quality (Sadler 2007). Teachers are considered able to make expert and reliable judgements because of their experiences with having to make repeated judgements forming patterns or instantiations, as well as socialisation into the standards of the discipline (Ecclestone 2001). However, even experts vary in their interpretation of standards and use of marking criteria, especially when they are not made explicit from the outset (Bloxham et al. 2016). Standards reside in the practices of academic and professional communities underpinned by tacit and explicit knowledge and are, therefore, subject to varied interpretation/enactments (O'Donovan et al. 2004). Students develop their understanding of standards, at minimum, through the production of work for assessment and then receiving a mark: this gives them clues about the quality of their work and they may form implicit criteria for improving the work regardless of the explicit criteria. We argue that through an explicit focus on evaluative judgement, students can form systematic and hopefully better calibrated evaluative judgements rather than implicit and idiosyncratic ones. The curriculum may then better serve the function of inducting and socialising learners into a discipline, providing them with opportunities to take part in discussing standards and criteria, and the processes of making judgements (of their own and others' work). This will help them develop appropriate evaluative judgement themselves in relation to criteria or standards and make more sense of and take greater control of their own learning (Boud and Soler 2016; Sadler 2010).

Many existing practices have potential for developing evaluative judgement. However, some of the ways in which they have been implemented are not likely to be effective for this end. Table 1 illustrates ways in which they may or may not be used to develop evaluative judgement. We focus attention on these five common practices, as they hold significant potential for further development.

Self-assessment

Self-assessment involves students appraising their own work. Earlier studies focused on student-generated marks, comparing them with those generated by teachers. Despite long-standing criticisms (Boud and Falchikov 1989), this strategy remains common in self-assessment research (Panadero et al. 2016). We suggest that persistence of this emphasis lies in several reasons. Firstly, such studies are easy to conduct and analyse. They involve little or

Table 1 Practices to develop evaluative judgement: sub-optimal strategies and potential improvements

| Practice | Sub-optimal strategies which may already be implemented | Potential improvements |
|-----------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Self-assessment | Self-assessments that solely generate grades or contribute to final grades; which do not involve students in identifying or considering criteria; which provide no feedback on the quality of the judgements involved. | Focus on how to identify and choose criteria for judgement; feedback to indicate how the self-assessment might be improved in future; use of self-assessments which span units and identify future learning needed. |
| Peer feedback/-review | Peer assessment which substitutes for teacher marking; peer feedback without consideration of criteria and the varieties of ways in which they may be applied; peer review without orientation and discussion of issues that may arise. | Emphasis on perspectives peers can uniquely bring to the work (e.g. extent to which it communicates to them quality); focus on the how work meets or does not meet agreed standards and criteria; appreciation that benefits might accrue more to the provider than the receiver of information. |
| Feedback | Feedback as ‘telling’ (Sadler 1989) or ‘corrections’ or as information handed to the student; information provided at times and on tasks, or in a manner which does not allow it to be easily used for future work; provision of information without follow-up on whether it was able to be used in subsequent tasks. | Feedback as dialogue which looks to future tasks rather than just present instances of work; sufficient detail and justification which help students understand how their work aligned or differed from the standards employed and information that expects a response or a plan for action; contains feedback that responds to students’ self-assessment of their own work; maintaining a feedback journal that spans a course with ipsative feedback per competence or learning outcome rather than task or unit focused. |
| Rubrics | Rubrics as checklists in a present/absent dichotomy rather than descriptors of quality of work; use of criteria that students have not been involved in discussing; rubrics supplanting evaluative judgement; standardised templates that do not respond to the different forms of judgement needed for each task. | Rubrics implemented with formative assessment purposes, rubrics co-created to develop a shared understanding of criteria and how they might be applied; rubrics that represent how evaluative judgements are actually made in the particular discipline; dialogic use of rubrics where there is discussion around the descriptors/criteria; use of rubrics in conjunction with exemplars; movement back and forth from work to rubrics then to the work to formulate deeper understandings of quality; rubric criteria that span multiple tasks across units. |
| Exemplars | Exemplars may be merely given to students or uploaded on a learning management system with little discussion of how quality is demonstrated within the exemplar. | Multiple, different exemplars discussed with peers and teachers to uncover what the range of quality indicators might be. |

no changes to curriculum and pedagogy. ‘Grade guessing’ can be added with little disruption to a normal class regime. Secondly, there remains a legitimate interest in understanding students’ self-marking; students exist in a grade-oriented environment. What is needed for the development of evaluative judgement is for the emphasis of self-assessment to be on eliciting and using criteria and standards as an embedded aspect of learning and teaching strategies (Thompson 2016). Self-assessment can more usefully be justified in terms of assisting students to develop multiple criteria and qualitatively review their own work against them and thus

refine their judgements rather than generate grades that might be distorted by their potential use as a substitute for teacher grades.

The focus of evaluative judgement is not just inwards on one's work but also outwards to others' performance, about the two in comparison to each other, and standards. Metacognition and 'reflective judgement thinking' (Pascarella and Terenzini 2005) also fall within this category of largely self-oriented ideas. Self-assessment is typically focused on a specific task, whereby a student comes to understand the required standard and thus improves their work on the specific task at hand (Panadero et al. 2016). Evaluative judgements encompass these elements while also transcending the immediate task by developing a student's judgement of quality that can be further refined and evoked in similar tasks.

Peer assessment and peer review

Similar to the literature regarding self-assessment, a significant proportion of studies of peer assessment have focused on the accuracy of marks generated and the potential for a number calculated from peer assessments to contribute to grades (Falchikov and Goldfinch 2000; Speyer et al. 2011). Simultaneously, others have pointed to the formative, educative, and pedagogical purposes of peer assessment (Dochy et al. 1999; Liu and Carless 2006; Panadero 2016) and consider both the receiving and the giving of peer feedback to be important for learning (van den Berg et al. 2006), particularly for the development of evaluative judgement (Nicol et al. 2014). By engaging more closely with criteria and standards, and having to compare work to these, students have reported a better understanding of quality work (McConlogue 2012; Tai et al. 2016). Many peer assessment exercises, whether formal or informal, may already contribute to the development of evaluative judgement. By a more explicit focus on evaluative judgement, students can pay close attention to what constitutes quality in others' work and how that may transfer to their own work on the immediate task and beyond. Learning about notions of quality through assessing the work of peers may reduce the effect of cognitive biases directed towards the self when undertaking self-assessment (Dunning et al. 2004). It is also likely that the provider of the assessment may benefit more from the interaction than the receiver, as it is their judgement which is being exercised (Nicol et al. 2014).

Feedback

Scholarship on feedback in higher education has progressed rapidly in recent years to move beyond merely providing students with information about their work to emphasising the effects on students' subsequent work and to involving them as active agents in a forward-looking dialogue about their studies (Boud and Molloy 2013; Carless et al. 2011; Merry et al. 2013). This work suggests that feedback (termed Feedback Mark 2) should go beyond the current task to look to future instances of work and develop students' evaluative judgement. Boud and Molloy (2013) introduced this model to supplant previous uses of 'feed forward'. Feedback to develop evaluative judgement would emphasise dialogue to clarify applicable standards and criteria that define quality, to help refine the judgement of students about their work and to assist them to formulate actions that arise from their appreciation of their work and the information provided about it from other sources. Feedback may be oriented specifically towards evaluative judgement across a curriculum, highlighting gaps in understandings of quality and how quality is influenced by discipline and context.

A focus on evaluative judgement requires new ways to think about feedback. Comments to students are not only about their work per se but about the judgements they make about it. The information they need to refine their judgements includes whether they have chosen suitable criteria, whether they have reached justifiable decisions about the work, and what they need to do to develop these capabilities further. There is no need to repeat what students already know about the quality of their work, but to raise their awareness of where their judgement has been occluded or misguided. Guidance is then also required towards future goals and judgement making. This new focus for feedback could be thought of as occupying the highly influential ‘self-regulation’ level of Hattie and Timperley’s (2007) model of feedback.

Rubrics

Rubrics can be thought of as a scaffold or pedagogy to support the development of evaluative judgement. There is extensive diversity of rubric practices (Dawson 2015), some of which may better support the development of evaluative judgement (e.g. formative use (Panadero and Jonsson 2013)) than others.

If rubrics are to be used to develop students’ evaluative judgement, they should be designed and used in ways that reflect the nuanced understandings of quality that experts use when making judgements about a particular type of task while acknowledging the complexity and fluidity of the practices, and hence standards, they represent. Current understandings of influences on rater judgement should also be communicated to give students a sense that there may also be unconscious, tacit, and personal factors in play. Existing work suggests that students can be trained to use analytic rubrics to evaluate their own work and the work of others (e.g. Boud et al. 2013); however, the ability to apply a rubric does not necessarily imply that students have developed evaluative judgement without a rubric. Hendry and Anderson (2013) proposed that marking guides are only helpful for students to understand quality when the students themselves use them to critically evaluate work. For evaluative judgement to be a sustainable capability, it needs to work in the absence of the artifice of the university. Although students’ notions of quality may be refined with each iteration of a learning task and associated engagement with standards and criteria, not all aspects of quality can be communicated through explicit criteria; some will remain tacit and embodied (Hudson et al. 2017).

Student co-creation of rubrics may support the development of evaluative judgement (Frailie et al. 2016). These approaches usually involve students collectively brainstorming sets of criteria and engaging in a teacher-supported process of discussing disciplinary quality standards and priorities. These processes may develop a more sustainable evaluative judgement capability, as they involve students in the social construction and articulation of standards: a process that may support them to develop evaluative judgement in the new fields and contexts they move into throughout their lives.

Exemplars

Exemplars provide instantiations of quality and an opportunity for students to exercise their evaluative judgement. A challenge in using exemplars is ensuring they are sustainable, that is, they represent features not just of a given task but of quality more broadly in the discipline or profession. This may be resolved by ensuring assessment tasks are authentic in nature. However, exemplars are likely to require some commentary for students to be able to use them to full advantage in developing their evaluative judgement.

There are a small number of research studies on the use of exemplars to develop student understanding of quality. Earlier work done by Orsmond et al. (2002) identified that student discussion regarding a range of exemplars without the revelation of grades aided in understanding notions of quality. To and Carless (2015) more recently have advocated for a ‘dialogic’ approach involving student participation in different types of discussion, as well as teacher guidance. However, teachers may need to reserve their own evaluative judgement to allow students to co-construct understandings of quality (Carless and Chan 2016). Students have been shown to value teacher explanation more than their own interactions with exemplars (Hendry and Jukic 2014). The use of exemplars in developing evaluative judgement is likely to require a delicate balance, where students are scaffolded to notice various examples of quality, without explicit ‘telling’ from the teacher.

Discussion and conclusion

We propose that developing students’ evaluative judgement should be a goal of higher education. Starting with such an integrative orientation demand refinements to existing pedagogical practices. Of the five discussed, some have a close relationship with assessment tasks; others can be part of any teaching and learning activity in any discipline.

Common to these practices is active and iterative engagement with criteria, the enactment of judgements on diverse samples of work, dialogic feedback with peers and tutors oriented towards understanding quality which may not be otherwise explicated, and articulation and justification of judgements with a focus on both immediate and future tasks. Evaluative judgement brings together these different activities and refocuses them as pedagogies to be used towards producing students who can make effective judgements within and beyond the course. We identified these pedagogies as developing evaluative judgement through the consideration of previous conceptual work which touched on the notion that students need a sense of quality to successfully undertake study and conduct their future lives as graduates (e.g. Sadler 2010; Boud and Falchikov 2007). In doing this, we have gone beyond using evaluative judgement as an explanatory term for student learning phenomena (Tai et al. 2016; Nicol et al. 2014) and sought to broaden its use to become a key justification for requiring students to engage in certain pedagogical activities.

Developing students’ evaluative judgement is important for several reasons. It provides an overarching means to communicate with students about a central purpose of these activities, how they may relate to each other and to course learning outcomes. With sceptical student-consumers, a strong educational rationale which speaks to students’ future use is necessary (Nixon et al. 2016). It may also persuade educators that devoting time to work-intensive activities such as using exemplars, co-creating rubrics, and undertaking self and peer assessment and feedback is worthwhile. Since the development of evaluative judgement is a continuing theme to underpin all other aspects of the course (Boud and Falchikov 2007), opportunities are available at every stage: in classroom and online discussions, in the ways in which learning tasks are planned and debriefed, in the construction of each assessment task and feedback process, and, most importantly, in the continuing discourse about what study is for and how it can be effective in preparing students for all types of work.

The explicit inclusion of evaluative judgement, in courses, might also actively engage student motivations to learn. The development of evaluative judgement complements and adds to other graduate attributes developed through of course of study. The notion of evaluative

judgement is not a substitute for well-established features of the curriculum but a way of relating them to the formation of learners' judgement. This additional purpose is likely to empower students in their learning context. Students will transition from being dependent novices and consumers of information to becoming agentic individuals able to participate in social notions of determining quality work. This sort of journey will require students to be supported in becoming self-regulated learners who take responsibility for their own learning and direct it accordingly (Sitzmann and Ely 2011).

Researchers and educators should be concerned not only with the effectiveness of interventions to develop evaluative judgement but also understanding the quality of student evaluative judgement. Existing 'grade guessing' research comparing student self-/peer marking against expert/teacher marking provides only a thin measure of evaluative judgement. Determining whether evaluative judgement has been developed could occur through particular self- and peer assessment activities in which students judge their own work and that of others and the qualities of the judgements made. The quality of an evaluative judgement lies not only in its result but in the thought processes and justifications used. High quality evaluative judgement involves students identifying and using appropriate criteria (both prescribed and self-generated) and discerning whether their own work and that of others meets these requirements. Through this process, students can be supported to refine their understanding of quality for future related tasks and come to know how to manage the ambiguities inherent in criteria when applied in different contexts. This presents a complex phenomenon for researchers and educators to understand or measure, for which correlations between teacher and student marks are inadequate.

This paper has developed a new definition and focus on evaluative judgement to describe why we might want to develop it in our students and how to go about this. Evaluative judgement provides the reason 'why' for implementing certain pedagogies. Further theoretical work may assist in better understanding how to develop evaluative judgement amongst students. While we have characterised it as a capability, there may also be motivational, attitudinal, and embodied aspects of evaluative judgement to consider. Empirical studies are also required to test aforementioned strategies for developing evaluative judgement, tracking the longitudinal development of evaluative judgement and professional practice outcomes over several years. Returning to Sadler's (2013) statement that quality is something I do not know how to define, but I recognise it when I see it, we have argued that a focus on evaluative judgement might enable a frank and explicit discussion of what constitutes quality and explored a range of common practices which may develop this capability. The exercise of evaluative judgement is a way of framing what makes for good practice.

Acknowledgements Our thanks go to attendees of the Symposium on Evaluative Judgement on 17 and 18 October 2016 at Deakin University, who contributed to discussions on the topic of this paper. In particular, Jaclyn Broadbent contributed to initial conceptual discussions, while Margaret Bearman, David Carless, Gloria Dall'Alba, Gordon Joughin, Susie Macfarlane, and Darrall Thompson also commented on a draft version of this paper.

Funding Ernesto Panadero is funded by the Spanish Ministry (Ministerio de Economía y Competitividad) via the Ramón y Cajal programme (file id. RYC-2013-13469).

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Ajjawi, R., & Higgs, J. (2008). Learning to reason: a journey of professional socialisation. *Advances in Health Sciences Education*, 13(2), 133–150. <https://doi.org/10.1007/s10459-006-9032-4>.
- Ajjawi, R., & Higgs, J. (2012). Core components of communication of clinical reasoning: a qualitative study with experienced Australian physiotherapists. *Advances in Health Sciences Education*, 17(1), 107–119. <https://doi.org/10.1007/s10459-011-9302-7>.
- Barton, K. L., Schofield, S. J., McAleer, S., & Ajjawi, R. (2016). Translating evidence-based guidelines to improve feedback practices: the interACT case study. *BMC Medical Education*, 16(1), 53. <https://doi.org/10.1186/s12909-016-0562-z>.
- Bloxham, S., Den-Outer, B., Hudson, J., & Price, M. (2016). Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education*, 41(3), 466–481. <https://doi.org/10.1080/02602938.2015.1024607>.
- Boud, D. (1992). The use of self-assessment schedules in negotiated learning. *Studies in Higher Education*, 17(2), 185–200. <https://doi.org/10.1080/03075079212331382657>.
- Boud, D. (2000). Sustainable assessment: rethinking assessment for the learning society. *Studies in Continuing Education*, 22(2), 151–167.
- Boud, D. (2007). Reframing assessment as if learning was important. In D. Boud & N. Falchikov (Eds.), *Rethinking assessment in higher education: learning for the longer term* (pp. 14–25). London: Routledge.
- Boud, D. (2017). Standards-based assessment for an era of increasing transparency. In D. Carless, S. Bridges, C. Chan, & R. Glofcheski (Eds.), *Scaling up assessment for learning in higher education* (pp. 19–31). Dordrecht: Springer. https://doi.org/10.1007/978-981-10-3045-1_2.
- Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: a critical analysis of findings. *Higher Education*, 18(5), 529–549. <https://doi.org/10.1007/BF00138746>.
- Boud, D., & Falchikov, N. (2007). Developing assessment for informing judgement. In D. Boud & N. Falchikov (Eds.), *Rethinking assessment in higher education: learning for the longer term* (pp. 181–197). London: Routledge.
- Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in Higher Education*, 38(6), 698–712. <https://doi.org/10.1080/02602938.2012.691462>.
- Boud, D., & Soler, R. (2016). Sustainable assessment revisited. *Assessment & Evaluation in Higher Education*, 41(3), 400–413. <https://doi.org/10.1080/02602938.2015.1018133>.
- Boud, D., Lawson, R., & Thompson, D. G. (2013). Does student engagement in self-assessment calibrate their judgement over time? *Assessment & Evaluation in Higher Education*, 38(8), 941–956. <https://doi.org/10.1080/02602938.2013.769198>.
- Boud, D., Lawson, R., & Thompson, D. G. (2015). The calibration of student judgement through self-assessment: disruptive effects of assessment patterns. *Higher Education Research & Development*, 34(1), 45–59. <https://doi.org/10.1080/07294360.2014.934328>.
- Brooks, V. (2012). Marking as judgment. *Research Papers in Education*, 27(1), 63–80. <https://doi.org/10.1080/02671520903331008>.
- Carless, D., & Chan, K. K. H. (2017). Managing dialogic use of exemplars. *Assessment & Evaluation in Higher Education*, 42(6), 930–941. <https://doi.org/10.1080/02602938.2016.1211246>.
- Carless, D., Salter, D., Yang, M., & Lam, J. (2011). Developing sustainable feedback practices. *Studies in Higher Education*, 36(4), 395–407. <https://doi.org/10.1080/03075071003642449>.
- Cowan, J. (2010). Developing the ability for making evaluative judgements. *Teaching in Higher Education*, 15(3), 323–334. <https://doi.org/10.1080/13562510903560036>.
- Dawson, P. (2017). Assessment rubrics: towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education*, 42(3), 347–360. <https://doi.org/10.1080/02602938.2015.1111294>.
- Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: a review. *Studies in Higher Education*, 24(3), 331–350. <https://doi.org/10.1080/03075079912331379935>.
- Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: implications for health, education, and the workplace. *Psychological Science in the Public Interest*, 5(3), 69–106. <https://doi.org/10.1111/j.1529-1006.2004.00018.x>.
- Ecclestone, K. (2001). “I know a 2:1 when I see it”: understanding criteria for degree classifications in franchised university programmes. *Journal of Further and Higher Education*, 25(3), 301–313. <https://doi.org/10.1080/03098770126527>.
- Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: a meta-analysis comparing peer and teacher marks. *Review of Educational Research*, 70(3), 287–322. <https://doi.org/10.3102/00346543070003287>.

- Frailé, J., Panadero, E., & Pardo, R. (2017). Co-creating rubrics: the effects on self-regulated learning, self-efficacy and performance of establishing assessment criteria with students. *Studies in Educational Evaluation*, 53(June), 6–76. <https://doi.org/10.1016/j.stueduc.2017.03.003>.
- Gielen, S., Dochy, F., Onghena, P., Struyven, K., & Smeets, S. (2011). Goals of peer assessment and their associated quality concepts. *Studies in Higher Education*, 36(6), 719–735. <https://doi.org/10.1080/03075071003759037>.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>.
- Hendry, G. D., & Anderson, J. (2013). Helping students understand the standards of work expected in an essay: using exemplars in mathematics pre-service education classes. *Assessment & Evaluation in Higher Education*, 38(6), 754–768. <https://doi.org/10.1080/02602938.2012.703998>.
- Hendry, G. D., & Jukic, K. (2014). Learning about the quality of work that teachers expect: students' perceptions of exemplar marking versus teacher explanation. *Journal of University Teaching & Learning Practice*, 11(2), 5.
- Hudson, J., Bloxham, S., den Outer, B., & Price, M. (2017). Conceptual acrobatics: talking about assessment standards in the transparency era. *Studies in Higher Education*, 42(7), 1309–1323. <https://doi.org/10.1080/03075079.2015.1092130>.
- Joughin, G., Dawson, P., & Boud, D. (2017). Improving assessment tasks through addressing our unconscious limits to change. *Assessment & Evaluation in Higher Education*, 42(8), 1221–1232. doi:<https://doi.org/10.1080/02602938.2016.1257689>.
- Knight, P. T., & Yorke, M. (2003). *Assessment, learning and employability*. Maidenhead: Open University Press.
- Liu, N.-F., & Carless, D. (2006). Peer feedback: the learning element of peer assessment. *Teaching in Higher Education*, 11(3), 279–290. <https://doi.org/10.1080/13562510600680582>.
- McConlogue, T. (2012). But is it fair? Developing students' understanding of grading complex written work through peer assessment. *Assessment & Evaluation in Higher Education*, 37(1), 113–123. <https://doi.org/10.1080/02602938.2010.515010>.
- Mery, S., Price, M., Carless, D., & Taras, M. (2013). *Reconceptualising feedback in higher education*. London: Routledge.
- Nicol, D. (2013). Resituating feedback from the reactive to the proactive. In D. Boud & E. Molloy (Eds.), *Feedback in higher and professional education: understanding and doing it well* (pp. 34–49). Milton Park: Routledge.
- Nicol, D. (2014). Guiding principles for peer review: unlocking learners' evaluative skills. In C. Kreber, C. Anderson, N. Entwistle, & J. McArthur (Eds.), *Advances and innovations in university assessment and feedback* (pp. 197–224). Edinburgh: Edinburgh University Press.
- Nicol, D., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218. <https://doi.org/10.1080/03075070600572090>.
- Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: a peer review perspective. *Assessment & Evaluation in Higher Education*, 39(1), 102–122. <https://doi.org/10.1080/02602938.2013.795518>.
- Nixon, E., Scullion, R., & Hearn, R. (2016). Her majesty the student: marketised higher education and the narcissistic (dis)satisfactions of the student-consumer. *Studies in Higher Education*, 50(7), 1–21. <https://doi.org/10.1080/03075079.2016.1196353>.
- O'Donovan, B., Price, M., & Rust, C. (2004). Know what I mean? Enhancing student understanding of assessment standards and criteria. *Teaching in Higher Education*, 9(3), 325–335. <https://doi.org/10.1080/1356251042000216642>.
- Orsmond, P., Merry, S., & Reiling, K. (2002). The use of exemplars and formative feedback when using student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education*, 27(4), 309–323. <https://doi.org/10.1080/0260293022000001337>.
- Panadero, E. (2016). Is it safe? Social, interpersonal, and human effects of peer assessment: a review and future directions. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 247–266). New York: Routledge.
- Panadero, E., & Alonso-Tapia, J. (2013). Self-Assessment: theoretical and Practical Connotations. When It Happens, How Is It Acquired and What to Do to Develop It in Our Students. *Electronic Journal of Research in Educational Psychology*, 11(2), 551–576.
- Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: a review. *Educational Research Review*, 9, 129–144. <https://doi.org/10.1016/j.edurev.2013.01.002>.
- Panadero, E., Brown, G. T. L., & Srijbos, J.-W. (2016). The future of student self-assessment: a review of known unknowns and potential directions. *Educational Psychology Review*, 28(4), 803–830. <https://doi.org/10.1007/s10648-015-9350-2>.

- Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students* (Vol. 2). San Francisco: Jossey-Bass.
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119–144. <https://doi.org/10.1007/BF00117714>.
- Sadler, D. R. (2007). Perils in the meticulous specification of goals and assessment criteria. *Assessment in Education: Principles, Policy & Practice*, 14(3), 387–392. <https://doi.org/10.1080/09695940701592097>.
- Sadler, D. R. (2010). Beyond feedback: developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, 35(5), 535–550. <https://doi.org/10.1080/02602930903541015>.
- 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>.
- Schmidt, H. G., & Rikers, R. M. J. P. (2007). How expertise develops in medicine: knowledge encapsulation and illness script formation. *Medical Education*, 41(12), 1133–1139. <https://doi.org/10.1111/j.1365-2923.2007.02915.x>.
- Sitzmann, T., & Ely, K. (2011). A meta-analysis of self-regulated learning in work-related training and educational attainment: what we know and where we need to go. *Psychological Bulletin*, 137(3), 421–442. <https://doi.org/10.1037/a0022777>.
- Speyer, R., Pilz, W., Van Der Kruis, J., & Brunings Wouter, J. (2011). Reliability and validity of student peer assessment in medical education: a systematic review. *Medical Teacher*, 33(11), e572–e585. <https://doi.org/10.3109/0142159X.2011.610835>.
- Tai, J., Canny, B. J., Haines, T. P., & Molloy, E. K. (2016). The role of peer-assisted learning in building evaluative judgement: opportunities in clinical medical education. *Advances in Health Sciences Education*, 21(3), 659. <https://doi.org/10.1007/s10459-015-9659-0>.
- Thompson, D. G. (2006). E-assessment: the demise of exams and the rise of generic attribute assessment for improved student learning. In T. S. Roberts (Ed.), *Self, peer and group assessment in e-learning* (pp. 295–322). USA: Information Science Publishing.
- Thompson, D. G. (2016). Marks should not be the focus of assessment—but how can change be achieved? *Journal of Learning Analytics*, 3(2), 193–212. <https://doi.org/10.18608/jla.2016.32.9>.
- To, J., & Carless, D. (2015). Making productive use of exemplars: peer discussion and teacher guidance for positive transfer of strategies. *Journal of Further and Higher Education*, 40(6), 746–764. <https://doi.org/10.1080/0309877X.2015.1014317>.
- Tomlinson, M. (2012). Graduate employability: a review of conceptual and empirical themes. *Higher Education Policy*, 25(4), 407–431. <https://doi.org/10.1057/hep.2011.26>.
- Topping, K. J. (2010). Methodological quandaries in studying process and outcomes in peer assessment. *Learning and Instruction*, 20(4), 339–343. <https://doi.org/10.1016/j.learninstruc.2009.08.003>.
- van den Berg, I., Admiraal, W., & Pilot, A. (2006). Peer assessment in university teaching: evaluating seven course designs. *Assessment & Evaluation in Higher Education*, 31(1), 19–36. <https://doi.org/10.1080/02602930500262346>.