Transformative Dialogues: Teaching and Learning Journal
Winter 2020, Volume 13, Issue 3
https://journals.kpu.ca/index.php/td/index

Examining Tacit Knowledge in Assessing International Postgraduate Students

Claudia Rosenhan
University of Edinburgh

Farah Akbar
University of Edinburgh

Takuya Numajiri
University of Edinburgh

Abstract
The study investigates student and assessor attitudes to a marking rubric used in a taught postgraduate programme to examine whether or not level descriptors (LDs) enhance students' and staff assessment literacy. A student cohort was surveyed at two time-points, with a response rate of 62% (n = 99) and 24% (n = 39), respectively. One focus group with four assessors was also conducted. Using exploratory factor analysis, we found students were confident in their understanding of the LDs, but also believed markers drew on tacit knowledge. This concern was confirmed, to an extent, by the focus group. The findings question the usefulness of LDs to foster assessment literacy, especially for international students, as they do not mitigate against tacit knowledge. Both data sets were small, therefore not generalizable. The findings are, however, indicative of recurring issues in academic assessment, in which international students struggle to attain the requisite understanding of quality necessary for their development as autonomous learners.

Keywords
postgraduate assessment; assessment literacy; higher education

Introduction

Students and assessors in higher education (HE) often have low assessment literacy as they do not understand assessment principles and practices (Norton et al., 2013; Price et al., 2012). At the same time the way assessors and students engage with a marking rubric may be considered a touchstone for examining and honing assessment literacy skills. Level descriptors (LDs) in the marking rubric provide quality definitions for evaluative criteria at predetermined levels. In the context of this investigation, a taught postgraduate (PGT) programme with a high level (over 90%) of international students, the marking scheme was designed as a task-type rubric for a critical academic essay. We aimed to use this rubric generically across a range of similar assignments (see Table 1).

Table 1
Marking Rubric

Criteria Marks	Knowledge and Understanding of Concepts	Knowledge and Use of Literature	Critical Reflection on Theory and Practice	Application of Theory to Practice	Academic Discourse	Planning and Implementation of the research (used mainly for the dissertation)
A (70-100) Distinction						
B (60 – 69) Merit	The work demonstrates understanding of the concepts and theories relevant to the discipline/area, as is shown by outlining and reporting on established issues with a view to explanation. Judgement is used to establish relationships between the various relevant concepts and theories, but these are not evaluated or framed by different perspectives or strands in the discipline. There is a tendency to list the concepts or					

Level Descriptors

	place them in the argument without further reflection					
C (50 – 59%) Pass						
D (40 – 49) Pass at Diploma level						
E (30 – 39) Fail						
F (below 30) Bad Fail	Not addressed in this investigation					

This rubric was intended to provide consistent feedback in accordance with the assessors' professional judgements, to guide students' learning, and enable the latter to improve in subsequent assignments. However, if students cannot relate to the LDs, or if assessors use them inconsistently, the assessment may fail to support students' learning. Hence, we wanted to investigate students' attitudes towards the marking rubric in general, and LDs in particular. By understanding students' attitudes, we should be able to infer whether or not students felt they were successfully engaging in a dialogue with their assessors via the LDs (Nicol, 2010). We were also interested in the attitude of assessors and how they utilised the marking rubric in their assessment of student work and in the feedback they provided. Our investigation was designed as a mixed-method study in which we investigated students' attitudes quantitatively through a survey and factor analysis and assessors' perspectives using thematic analyses of focus group data.

This investigation is considered relevant because assessment and feedback regularly receive the highest frequency of negative responses in PGT surveys. Students report, for example, a lack of clarity about what they are expected to achieve (PTES, 2017). To counteract students' concerns, we initially wanted to develop students' understanding of the quality of their work and to hone their active assessment literacy skills (Nicol et al., 2014). For this purpose, the LDs for the programme were rewritten with the expressed intention of making the assessment criteria more explicit, as it is assumed that this would enhance their use as a medium for developing assessment literacy skills through shared understandings and dialogue between assessors and students. In the subsequent investigation, students' and assessors' attitudes on a PGT programme towards the LDs in the assessment rubric were used as a means to infer their levels of active engagement with them and thereby understand the development of their assessment literacy skills.

Literature Review

Assessment literacy in higher education (HE) goes to the heart of how value is added for students in achieving their learning objectives. This is particularly relevant for international students, who frequently struggle with new academic cultures. Marking rubric LDs may be a direct way of scaffolding international students' understanding of how to develop their learning, especially if the LDs operate at course or programme level (Leask, 2011).

The debate on LDs originated, however, from a link between criterion-based assessment of educational outcomes and quality assurance at national level. National quality codes demand accountability, explicitness and constructive alignment from assessment processes (e.g., Quality Assurance Agency for Higher Education, 2018). Quality assurance aspects dominate assessment policies of HE institutions, and exhaustive documentation of assessment processes and moderation practices are standard procedure in many universities (Boud, 2007). Quality standards knowledge is encoded and disseminated via the numerous artefacts, of which the marking rubric is one example (Sadler, 2014).

Published criteria and LDs serve as a strategy to address public scepticism about educational standards (Grainger et al., 2008). The key idea is that standards are maintained through transparency and public accountability (Brown, 2010; Koh, 2011). If LDs align with established performance criteria, they are more likely to be accepted publicly as trustworthy indicators for quality. Criterion referencing further suggests assessment in HE is objective and robust analytical measurement, replacing a perceived arcane standards model of undefined assumptions (Stowell, 2004). This techno-rationalist paradigm, in line with the auditable outcomes-based ethos of HE, is meant to cast a veil of rigour over what remains a fundamentally subjective assessment method of complex intellectual performances required by PGT students (Hussey & Smith, 2002). Pre-set criteria, however, can only record a fuzzy signal of achievement and overlook unarticulated performances. While LDs may be useful in recording the essence of a performance standard across different levels, they do not automatically guarantee good quality assessment practices (Bloxham, 2009). Faith in criterion-referencing seems to be misplaced.

Nevertheless, constructive alignment (Biggs & Tang, 2010) has become the dominant assessment paradigm of the last 15 years (Hudson et al., 2017). The strength of a standards-based accountability framework lies in the public availability of the marking rubric that seems to guarantee consistency and unassailability of grades (DeLuca, 2012; Taras, 2009). Accordingly, measurable outcomes can be predicted and controlled, which in turn bolsters the institution's professional status as 'assessor' (Almqvist et al., 2017). An apparent social justice agenda is also operationalised through explicit criteria, as knowledge about assessment is thus conceivably accessible to all (Torrance, 2017). This is particularly pertinent in the context of the internationalisation of HE. However, institutional standards are not always reflective of the public's priorities. Rather, institutional standards reproduce mandated institutional knowledge (Alderman, 2009; Ashworth et al., 2010).

The entanglement of accessibility issues and public accountability is only part of the complex network of formalised academic assessment and feedback. Individual assessors' tacit frameworks repeatedly endure over disciplinary norms (Taras & Davies, 2012). Assessment is, in essence, judgement and involves heuristic methods (Brooks, 2012; Crisp, 2013). Tacit knowledge (i.e., connoisseurship) is instrumental in the process of judging but is frequently inarticulable (Tsoukas, 2003). Judgements may be unreliable, inconsistent and difficult to articulate, but this is not to be confused with bias or random judgements (Shay, 2005). Judging is, instead, a complex process of 'double reading,' in which the interpretative framework of the individual is entangled with implicit disciplinary norms.

It is clear assessors engage personal constructs when assessing a piece of work, and these constructs are fluid and intuitive (Bloxham et al., 2016; Hunter & Docherty, 2011). LDs are then used retrospectively to provide justification, and academic judgement becomes a source of bargaining, a 'shopping around for a grade' across assessors (Bloxham et al., 2011). This ultimately leads to indeterminacy in assessors' judgement which cannot serve the mythos of objectivity (Sadler, 2009). The separation between explicit LDs and private judgements creates a tension intended to be addressed by communities of practice.

The collective nature of tacit professional knowledge is situated in communities of practice, producing a contextualised 'guild knowledge' that builds expertise (Orr, 2010). Marking rubrics may provide structured guidance to the shared 'guild' understandings, and moderation dialogues help practitioners develop a common language, thus elucidating the fuzzy nature of LDs (Grainger et al., 2008; Adie et al., 2013). However, communities of practice do not automatically share a common understanding and moderation rarely aids calibration (Hudson et al., 2017). Additionally, moderation may draw on extra, uncalibrated and internalised criteria, such as specific characteristics of students (Orr, 2007). While the efficacy of moderation is not proven, continued faith in the moderation process lies in attempts to harmonise the intangible sense of personal and locally agreed standards (Bloxham & Boyd, 2012).

The lack of direct correspondence between the verbalisation of assessment criteria and tacit professional knowledge may be due to unquantifiable linguistic indeterminacy (Sadler, 2013). A salient question is whether or not assessment literacy is actually a linguistic issue (Forsyth et al., 2015). Students from international educational and linguistic backgrounds often struggle with academic concepts, such as analysis, synthesis, and critical reflection. The techno-rationalist language of LDs is further confounded by the fuzziness and malleability of standards; for example, many marking rubrics include qualifiers, modifiers, and hedge words that lack a clear grounding in the qualitative nature of the work. The use of relative and comparative terminology adds vagueness about the accomplishment of a criterion, especially at the threshold level (Payne & Brown, 2011). Moreover, marking rubrics are commonly based on intuitive and historical wordings (Greatorex et al., 2001). As such, there is no 'thing-in-itself' to which a description may point, and which may help students to direct their own learning.

Assessment has the power to direct students' learning, mainly through the benefits of feedback and feedforward (Jessop & Tomas, 2017; Sambell et al., 2013). An understanding of marking rubrics by the students enables self-regulation, empowerment, and autonomy (Popham, 2011; Price et al., 2012). Students need the transparency of the LDs, operationalised as feedback, to develop an understanding of the quality of their work. An ideal way of enhancing students' assessment literacy is, therefore, through partnerships between students and their assessors (Deeley & Bovill, 2017). Conversely, students' active involvement in their learning, through shared understandings with their assessors of transparent quality criteria, fosters assessment literacy (William & Thompson, 2008).

Since the language of LDs, and how it may be repeated in feedback, is frequently considered the main stumbling block to the development of shared understandings, ambiguities can be lessened through enabling students to get a 'feel' for a standard expressed in the LDs. This is only possible, however, if these standards are applied fairly and consistently by the assessors. The proposed 'nested hierarchy' of approaches to assessment literacy, however, frequently stops short of the ‘cultivated’ community of practice in which that knowledge is made explicit to students (O’Donovan et al., 2008). The key concern in the field is how students are excluded from tacit judgments of their work, given tacit judgments by assessors are the rule rather than the exception in assessment situations. This would threaten the partnership between students and their assessors. Hence, this study investigated firstly students’ attitudes towards LDs to see whether or not they trusted their work was evaluated fairly and transparently, and whether or not they believed they were able to act on the feedback being given to them, as a way of enhancing their assessment literacy. It also investigated whether the reliability of assessment was ensured by the consistent use of LDs by the assessors, signalling their assessment literacy. In view of the concerns reviewed in the field, our research questions were:

1. Do LDs enhance students’ assessment literacy skills through the development of clear understandings of assessment criteria?
2. Do LDs assist a clear understanding of the criteria amongst groups of assessors on assignments (thereby enhancing assessor reliability)?

Methods and Results

Quantitative Study: Level Descriptors

Overview

To investigate students’ attitudes towards LDs, and whether or not they assisted in clarifying the students’ understandings of the criteria (i.e. enhanced their assessment literacy), we conducted a cohort survey during the 2016/17 academic year at two time points. Our aim was to see if there was any change in their understanding of the LDs during the academic year, since growing familiarity with assessment processes may have increased their understanding and thus their assessment literacy. This involved 99 PGT students at the March time point and 39 at the June time point. There were three incomplete cases at the March time point. This reflected a response rate of 62% (n = 99) in March and 24% (n = 39) in June. The drop in return rates at the June time point was most likely due to ‘survey fatigue’ at the end of the academic year. Using a 35-item self-administered questionnaire, respondents rated items on a 4-point Likert scale. This scale had high reliability: Cronbach’s alpha was 0.94 at the March time point and 0.95 at the June time point. The questionnaire items were developed following a systematic literature review of the connection between assessment and learning, students’ confidence in how assessment aligns with curriculum, their confidence in decisions made based upon the LDs, their own self-regulation and understanding of the LDs. Additionally, the respondents were asked to describe their background (home/international student) and familiarity with assessment procedures as a baseline for familiarity. The questionnaire was piloted with the 2015/16 cohort, and changes in wording and the arrangement of the scales were made accordingly. Two cases with missing data were excluded from the pairwise analyses.

Ethics

All students who participated in the research did so on an opt-in basis, following a detailed explanation of the aims and objectives of the research. Students who filled in the questionnaire gave their informed consent to participate. Data was initially collected via an online survey, which ensured privacy and confidentiality. Due to poor response rates, this was subsequently changed to paper copies, and we asked students, if they had not yet participated, to fill in the questionnaire at a programme meeting. This may have exerted some pressure on students to comply, as both researchers were present at that meeting. However, it was clearly explained participation was voluntary, and completion of the questionnaire would not be monitored. The same strategy was used at June time point. Since the response rate remained relatively low in relation to the students present at the meetings, it can be assumed students freely exercised their right not to fill in the questionnaire. The questionnaire did not collect any identifying personal data beyond some general information about knowledge of assessment procedures and whether they were home or international students.

Factor Analysis

Exploratory factor analysis (EFA) was used for the March time point dataset but not in June since the number of participants was much smaller. EFA was used to explain a larger set of variables with a smaller set of latent constructs and to determine if the dataset could be reduced to a smaller set of factors (Field, 2013; Hair et al., 2010; Henson & Roberts, 2006). For conducting EFA, a Mahalanobis Distance (MD) for each case was computed to identify multivariate outliers (Hair et al., 2010). The critical value of $\chi^2_{(35)} = 66.62$, $p = 0.001$ shows there were no multivariate outliers among the cases. Moreover, distributions of the 35 variables (based on the questionnaire items) were examined with the frequencies. Although the sample size was small, each of the variables had skewness or kurtosis within acceptable ranges, ±1.

As the sample size was only 99 cases, an approach for factor analysis with small sample numbers designed by Zhao (2009) was applied. Kaiser-Mayer-Olkin (KMO) measure and Bartlett’s test were used to check the factorability and sampling. The sample size was adequate as the overall KMO was 0.831 and Bartlett’s test was statistically significant ($\chi^2_{(91)} = 1387.6$, $p = 0.000$), $p < 0.001$. All individual variables had an anti-image correlation matrix of less than 0.60, which further confirmed sample adequacy.

Principal component analysis, using both orthogonal and oblique rotations, was used on all 35 items. Items with the smallest communality were dropped in the analysis until the communalities of all variables were above 0.60. On this basis, 2 items were removed. The mean value of the communalities of 33 items was 0.72 (> 0.70). The scree plot test was applied to determine the number of factors and suggested that a 3-factor solution should be appropriate. The current study set the cut-off point of 0.55 and above for each factor loading as suggested by MacCallum et al. (2001). As a result, 22 items in Table 2 were retained, and there was no cross loading among the 3 factors. There were very weak or negligible correlations between the factors (Factor 1/Factor 2, r = -0.060; Factor 1/Factor 3, r = 0.482; Factor 2/Factor 3, r = -0.006). The variable to factor ratio is 7.3. This can be regarded as “a moderate to high degree of overdetermination” (Zhao, 2009, np).

Table 2
Factor Loadings for PGT Students from Explanatory Factor Analysis Using Varimax Method (n = 97)

	Factor Loadings		
Variable	**Conf.**	**Value**	**Conc.**
Q5-3. I clearly understand what is meant by a particular grade based on the standard guidelines available through the level descriptors.	0.851		
Q5-4. I can identify with the statements of achievement in level descriptors that are composed with the help of qualifiers, modifiers and hedge words.	0.824		
Q5-2. I find the qualifying words used in the level descriptors helpful for distinguishing grades.	0.807		
Q5-1. I clearly understand what is ‘good’ or ‘poor’ achievement of a criterion based on the level descriptors.	0.790		
Q5-5. I find level descriptors can equally be used for any written assignment that is required on the programme (does not apply to Research Methods).	0.726		
Q5-6. I find level descriptors provide fixed reference points of how the criterion has been achieved.	0.722		
Q6-8. I find that the level descriptors make me more satisfied with the marking process.	0.709		
Q6-2. I have a sense of empowerment and autonomy, because the level descriptors provide a clarified expectation of what I need to do in order to improve.	0.703		
Q6-5. I find that level descriptors provide me with a ‘feel’ for a standard and how standards are applied fairly and consistently.	0.700		
Q6-6. I find that level descriptors increase my confidence in the marking process.	0.681		
Q6-4. I find that the level descriptors help me understand what is behind higher-order skills, such as analysis, synthesis and critical reflection.	0.654		
Q6-7. I find that level descriptors enable me to manage my expectations about the marking process.	0.624		
Q2-1. The level descriptors are explicitly linked to the learning outcomes of the courses on the programmes.		0.784	
Q3-4. The level descriptors underpin the relationships between assessment, learning outcomes and course objectives.		0.691	
Q2-2. Each level descriptor relates to a discrete level of intellectual performance with which I am familiar.		0.664	
Q3-5. The overall quality of my work shows in terms of the multiple interconnected level descriptors for the criteria.		0.650	
Q2-5. The level descriptors refer to the mandated knowledge I have acquired in the courses on the programme.		0.642	
Q4-2. Assessors may sometimes use more constructs, or rank constructs differently or interpret shared constructs differently, than are stated in the level descriptors.			0.827

	Factor Loadings		
Variable	**Conf.**	**Value**	**Conc.**
Q4-1. Assessors may sometimes have different expectations and relative standards that are not specified in the level descriptors.			0.759
Q4-7. Level descriptors do include a 'hidden curriculum', i.e. interpretations of constructs that are invisible to me.			0.756
Q4-4. Assessors may use 'guild knowledge' (Orr 2010), i.e. professional knowledge that is situated and local, which differs from my own knowledge about the assessment.			0.713
Q4-5. Level descriptors refer to slippery and opaque concepts that can only be known through experience and training.			0.582
Eigen values	8.85	2.83	1.55
% of variance accounted for	40.23	12.84	7.02
Cronbach's α	0.94	0.82	0.78

At the March time point, the factors Confidence (12 items), Value (5 items), and Concern (5 items) accounted for nearly 60 percent of the total variance in the dataset.

Confidence items were about the LDs' language (Question 5) and how confident students were about decisions made based upon these descriptors (Question 6). Value items focused on whether or not students believed feedback based on the LDs could be used to direct their learning and connects with learning outcomes (Questions 2 and 3). Concern items were those in which students voiced beliefs assessors drew on tacit and guild knowledges and hidden curricula. These factors allow insights into assessment literacy in that they reveal students' understanding of the principles of assessment.

The mean scores for each factor ranged from a minimum value of 1.50 to a maximum value of 4.00. Overall, the respondents had the most positive attitude towards Value (range = 1.50–4.00, $\bar{x} = 3.11$, sd = 0.52), but less positive attitude towards Confidence (range = 1.50–4.00, $\bar{x} = 2.96$, sd = 0.60). However, students' felt excluded from tacit knowledges of their assessors as displayed for the factor, Concern ($\bar{x} = 3.03$, sd = 0.50).

This three-factor model was subsequently applied to the June time point data to compare the differences between March and June time points, and to see if increased familiarity (i.e. an enhanced assessment literacy), may have changed the overall tendency to be concerned about tacit knowledges.

T-tests

Independent sample T-tests were used to analyse differences between home and international students and students who were familiar and those who were not familiar with the LDs used on the PGT course.

March Time Point

There were no significant differences in mean scores on any of the three factors between Home and International students (see Table 3). However, students, who reported familiarity with the LDs used on the programme, had significantly higher scores on the Confidence subscale ($\bar{x} = 3.19$, sd = 0.53) and Value subscale ($\bar{x} = 3.23$, sd = 0.47) than those who were unfamiliar with the LDs (confidence: $\bar{x} = 2.73$, sd = 0.60, p<0.001; value: $\bar{x} = 2.99$, sd = 0.55, p<0.05), thus indicating the value of familiarity for assessment literacy (see Table 4).

Table 3
Background at March Time Point

	Background		
	Home Students (n = 11) $\bar{x}$ (sd)	International Students (n = 85) $\bar{x}$ (sd)	p
Confidence	2.86 (0.62)	2.97 (0.61)	0.575
Value	2.98 (0.55)	3.13 (0.52)	0.392
Concern	2.95 (0.71)	3.04 (0.47)	0.691

Table 4
Levels of Familiarity with LDs at March Time Point

	Students' Familiarity Level with LDs		
	Familiar (n = 47) $\bar{x}$ (sd)	Unfamiliar (n = 49) $\bar{x}$ (sd)	p
Confidence	3.19 (0.53)	2.73 (0.60)	0.000
Value	3.23 (0.47)	2.99 (0.55)	0.026
Concern	2.95 (0.50)	3.09 (0.50)	0.171

Comparing March and June Time Points

Next, we examined differences between March and June time point results across all three factors. Mean scores for the June cohort were higher than those of the March cohort on Confidence ($\bar{x} = 2.97, sd = 0.53$ and $\bar{x} = 2.95, sd = 0.60$, respectively), and Value ($\bar{x} = 3.17, sd = 0.49$ and $\bar{x} = 3.10,\ sd = 0.52$, respectively), but not on Concern ($\bar{x} = 3.01$, sd = 0.60 and $\bar{x} = 3.03, sd = 0.50$, respectively). However, no significant difference was found (p>0.05). Similarly, there were no significant differences between Home and International students across all three factors at both the March and June time points.

Compared to the March cohort in which 47 of 96 students reported not being familiar with the LDs, the majority (32 out of 39) in the June cohort reported they were not familiar with the LDs when they started the programme. The June cohort, who reported they were familiar with the LDs used on the PGT course, had higher scores on the Confidence subscale ($\bar{x}$ = 3.37, sd = 0.50) than those in the March cohort ($\bar{x}$ = 2.95, sd = 0.50, p<0.05) (see Table 5). Conversely, there were no significant differences between participants who were not familiar with LDs across all three factors at both the March and June time points (see Table 6).

Table 5
Familiarity with LDs at March Time Point Compared with the June Time Point

	Students Familiar with LDs		
	March (n = 47) $\bar{x}$ (sd)	June (n = 7) $\bar{x}$ (sd)	*p*
Confidence	3.19 (0.53)	3.12 (0.45)	0.750
Value	3.23 (0.47)	3.26 (0.40)	0.884
Concern	2.95 (0.50)	3.37 (0.50)	0.042

Table 6
Level of Non-Familiarity with LDs at March Time Point Compared with June Time Point

	Students Not Familiar with LDs		
	March (n = 49) $\bar{x}$ (sd)	June (n = 32) $\bar{x}$ (sd)	*p*
Confidence	2.73 (0.60)	2.97 (0.53)	0.114
Value	2.99 (0.55)	3.18 (0.49)	0.175
Concern	3.09 (0.50)	3.01 (0.60)	0.172

At the March time point, the degree of familiarity with the LDs was related to Confidence, Value, and Concern scores. Students with high familiarity reported higher mean scores on Confidence and Value, and lower mean scores on Concern. However, higher mean levels of Concern were reported by those with high familiarity in the June cohort. This suggests increasing familiarity with the LDs did not alleviate concerns about tacit knowledge but may have increased them.

In summary, Concern was the only factor which was significantly different between the two time points for students familiar with the LDs and was weakly associated with increasing familiarity with the LDs. It illustrates that enhanced assessment literacy suggests an enhanced concern about assessors’ tacit knowledge.

Qualitative Study: Focus Group

The role of practitioners’ beliefs for the topic was considered important to the extent standards have been internalised by assessors and affect their scoring practices (Bloxam & Boyd, 2012). We explored this in a qualitative investigation including a focus group interview.

Focus Group Interview

Four assessors who marked on the programme participated in the focus group interview. Two of the assessors were experienced full-time staff and the other two were final year PhD students, new to the marking process. The aim of the interview was to have participants share their experiences with and opinions on how they used the LDs to assess assignments and see how much LDs helped with shared understandings between assessors and, indirectly, between assessors and students, as a symbol of assessment literacy.

The focus group interview was recorded, and notes (participants' responses) were taken simultaneously. The recording was professionally transcribed, and the data was analysed using the seven stages of Framework Method (Gale et al., 2013). This systematic approach was useful so all assessors' data could be compared and contrasted, allowing different perspectives to arise while closely reflecting on the contexts from which they emerged. The data was coded separately by two research team members; although the codes were packaged differently, they showed substantial agreement in terms of emerging themes. These themes were independently developed by the two researchers deductively and inductively from assessment criteria in HE literature and from the narratives of the participants and will be discussed below.

Ethical Issues

The familiarity of all focus group participants with each other and with the researchers was identified as a potential ethical risk. Participants may have felt cautious communicating truthfully about how they used LDs when marking assignments. Junior participants may have felt pressured to conform to expectations from more senior staff, and the researcher who mediated the focus group. These ethical concerns were mitigated by creating a friendly and collegiate atmosphere during the focus group. Participants were invited to respond to each other and to create their own dynamic in the focus group to alleviate interviewer bias (Browne, 2016). As guaranteeing anonymity was problematic reporting of data from the focus group research was not attributed to any identifiable variable, such as experience or gendered pronouns.

Findings

Six themes emerged from the focus group data: accountability; rigour; interpretation; language; familiarity; and interpretation.

Accountability. Some participants expressed concerns assessors tended to give marks based on holistic impressions, which made it difficult to account for the marks awarded and may have contributed to unfairness. They also felt there was a gap in the LDs with the absence of specific features to distinguish the upper and lower bands (i.e. bands A, B, C, D and E) in the assessment criteria (i.e., to discriminate clearly between, for example, 58%, upper C, and 61%, lower B). Thus, questions remain as to how assessors accounted for high or low marks within a band. Despite this, participants agreed LDs provided a means to justify scores and feedback. Moreover, having LDs meant they were able to dispel any uncertainties they may have had with their essays. This clearer understanding of the criteria amongst groups of assessors on assignments arguably enhances the reliability of scoring.

Rigour. The focus group participants reported scoring became efficient and robust only when the LDs were well understood and had been internalised. Initially, they reported slow progress with marking as they had to take time to study the language and reflect on its meanings. They also reported having to rationalise the LDs to understand the allocation of marks. One participant said, when working with the LDs, they needed to establish their relationship with the assignment requirement, which they did by matching the different aspects of the assignment to the assessment criteria. It is clear familiarity with the LDs led to more informed judgements as a result of the assessors' improved assessment literacy.

Despite this, some participants reported resorting to 'common sense' and 'general impression' rather than using the LDs as their guide. Even when the LDs for each band were prescribed, assessors tended to assess based on their knowledge, expertise and working standard. One participant expressed having deep understanding of the course and, "…knowing what the students should actually portray in the essay…" allowed them to have an impression of what an A paper should look like. Others took a more rationalistic view on scoring; they used the presence or absence of a criterion to make sense of their own grading system, "…so if all the suggested changes have no consideration, that's a low B or low C…if there is some consideration then it actually would be a borderline".

Language. Most focus group participants found the language of the LDs played a role in the accessibility of the assessment criteria, and thus for enhancing assessment literacy. They reported having to work with quite a lot of information within each criterion, at times with language they considered very academic, vague and verbose. They felt the LDs could have been more intelligible to assessors and international students. Despite this, two participants expressed appreciation for the more specific phrases in the LDs as they were able to match them against their own assessment requirement. The examples provided were helpful in illustrating specific criteria and contributed to better understandings. These factors affected the identification of criteria and how reliably assessors used the LDs.

Familiarity. The participants reported having varying levels of familiarity with the LDs. One participant, in particular, as a first-time assessor, felt their lack of familiarity with the LDs made it difficult to work with them. Another participant, having internalised previous LDs, found it hard to adapt to the current ones, "…I had internalised these, I was familiar with them, nobody likes change..." Others, who considered themselves familiar with the LDs, were able to match them quite quickly against the course requirements and one reported using the criteria and LDs for the purpose of students' self-assessment in their teaching, "...I try to scaffold the students' use of the learning descriptors, so I have like self-assessment checklists..." Another assessor stressed the importance of knowledge and awareness of the LDs, and the need to be trained to become familiar with the assessment process, as this would enhance assessment literacy.

Interpretation. Our analyses of the data highlighted concerns about different assessors' interpretations of the LDs, especially during the standardisation and moderation process. Participants felt this may have been because the same generic criteria were used for different courses with varying foci and requirements. One assessor reported the more pressing issue was not how the LDs were used, but rather, "…how you interpret the criteria into the context of your course..." It was believed the lack of common understanding of the LDs had resulted in inconsistencies in marks awarded. Another assessor expressed their shock over the level of subjectivity surrounding the understanding of the LDs, which led to differing scores awarded for one single assignment, "…I was shocked by the way some people would give the same essay 70 and some people would give 60 and some people would give 50, and some would give like 48…so there will always be subjectivity…".

Participants reported using various strategies for making sense of the LDs to inform their judgments. Most times, interpretation is subjective, "…I work my way out actually to how to try to understand them..." One person said they focused on one criterion at a time and identified a defining phrase and feature for each criterion and band by paraphrasing the descriptors, "…if this essay has A and B, then it is within the knowledge and understanding it belongs to category A...” Another participant reported they rely on personal judgement, “…that’s where the personal judgement does come in, because I did use a lot of it and it made it so much easier…” and common sense, “…I use common sense, just common sense and I know that it must have sounded bad…but I think it’s important to me...” Other techniques included relying on examples to interpret meanings, cross referring the new LDs to the ones used previously, formulating their own understanding of the LDs based on own topic knowledge and previous marking experience, and also matching LDs against assignment context. These individual responses could be considered a threat to reliable assessment.

In addressing such threats to reliability, a participant suggested assessors could unpack the LDs and discuss each other’s understandings to reach an agreement. However, it was unanimously agreed building a community of practice amongst assessors requires time, willingness and commitment, which was not always forthcoming.

Conversely, some participants felt the standardisation and moderation procedure was rigorous and enabled assessors to go through a norming process. It was a process by which assessors could discuss the LDs according to the requirements of the course assignment and come to shared understandings, “…you need generic criteria, but the way people kind of interpret them, that makes the difference.”

Discussion

Analyses of the qualitative data suggested the LDs enhanced assessors’ understandings of the criteria, and thus their assessment literacy, to a certain extent, but the assessment process was fraught with individual issues due to the different ways assessors viewed their own professional knowledge, their topic knowledge, the level at which they were working, and their relationship with other assessors. Assessors also tended to draw on their own expertise when marking assignments by formulating their own interpretations of the LDs as well as by relying on personal judgements and common sense (Bloxham et al., 2016; Shay, 2005). This seems to suggest, while standards have been internalised by assessors, they also resorted to tacit judgements, which threatens reliability. This concurred with students’ concerns about the use of knowledge and criteria not reflected in the LDs for the assessment.

Both quantitative and qualitative analyses suggested distance between students and assessors in matters of assessment. LDs, as part of the marking rubric, are generally seen as having an important role in alleviating student dissatisfaction with assessment and feedback, as they could improve students’ clarity about assessment requirements. In short, LDs may enhance assessment literacy. However, the mere existence of LDs is not enough to serve as a training ground for assessment literacy if they are not used reliably (Bloxham & West, 2004; Rust et al., 2005; Blair & McGinty, 2013; Mulder et al., 2014). The distance between students and assessors is created through unreliable assessment practices. Our findings suggest this must be overcome so they can become partners in assessment (Smith et al., 2013; Deeley & Bovill, 2017). This is especially important for international students who may not be familiar with assessment concepts. Students and assessors should be constantly engaged in a dialogue about the practices of assessment and the interpretation of criteria as well as the language used. Exemplars shared between assessors and students, may, for example, encourage the development of connoisseurship of both partners (Handley & Williams, 2011).

As our research has indicated, the main hindrance to this ideal was the perceived and acknowledged existence of tacit professional knowledge. The use of LDs alone did not consistently facilitate commonality in understanding between students and assessors. LDs have emerged as a space where shared understandings between assessors and students seem to diverge rather than converge. However, only if students are aware of what is expected of them, and assessors are transparent and accountable in their assessment, commonality of understanding may be achieved.

To enhance assessment literacy of assessors and students, they need to have shared ownership of the marking rubric they are using; assessment tasks need to be negotiable and contextual. Familiarity with assessment procedures needs to be honed based on the socialisation of students, especially for those from other academic cultures. Feedback should draw on students' multicompetences to analyse, discuss and apply assessment criteria to work, (e.g., via dialogic reflection). Assessment literacy is an iterative process, which depends on unhurried chances to develop complex understandings. Ongoing active engagement with assessment practices is essential to fostering assessment literacy. Further research is needed to explore how this can be achieved against increasing demands on academics to assess and provide feedback with ever decreasing resources.

Limitations

This research was originally conducted as a pilot validation study for newly-designed LDs on a PGT course. Robust analyses of the literature and the data provided some assurances of the external validity of the research beyond a mere validation of the artefact and allowed a critical analysis of assessment literacy of the stakeholders in this assessment. As a small-scale project, it does not claim generalisability. It does, however, confirm the salient themes discussed in the literature. Further investigations of professional judgement in assessment are warranted, and of the extent to which it may be possible to include international students as parties to those judgements. Whether LDs in criterion-based marking rubrics ultimately provide the right pathway, however, is questioned.

References

Adie, L., Lloyd, M. & Beutel, D. (2013). Identifying discourses of moderation in higher education. *Assessment & Evaluation in Higher Education*, *38*(8), 968–977. https://doi.org/10.1080/02602938.2013.769200

Alderman, G (2009). Defining and measuring academic standards: A British perspective. *Higher Education Management and Policy, 21*(3), 9–22. doi:10.1787/hemp-21-5ksf24ssz1wc

Almqvist, C. F., Vinge, J., Väkevä, L. & Zandén, O. (2017). Assessment *as* learning in music education: The risk of "criteria compliance" replacing "learning" in the Scandinavian countries. *Research Studies in Music Education*, *39*(1), 3–18. https://doi.org/10.1177/1321103X16676649

Ashworth, M., Bloxham, S. & Pearce, L. (2010). Examining the tension between academic standards and inclusion for disabled students: The impact on marking of individual academics' frameworks for assessment. *Studies in Higher Education*, *35*(2), 209–223. https://doi.org/10.1080/03075070903062864

Biggs, J. B. & Tang, C. (2010). Applying constructive alignment to outcomes - based teaching and learning. Retrieved 23 September 2017 from https://intranet.tudelft.nl/fileadmin/Files/medewerkersportal/TBM/Onderwijsdag_2014/What-is-ConstructiveAlignment.pdf.

Blair, A. & McGinty, S. (2013). Feedback-dialogues: Exploring the student perspective. *Assessment & Evaluation in Higher Education, 38*(4), 466–476. https://doi.org/10.1080/02602938.2011.649244

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. & Boyd, P. (2012). Accountability in grading student work: Securing academic standards in a twenty-first century quality assurance context. *British Educational Research Journal, 38*(4), 615–634. https://doi.org/10.1080/02602930801955978

Bloxham, S., Boyd, P. & Orr, S. (2011). Mark my words: the role of assessment criteria in UK higher education grading practices. *Studies in Higher Education*, *36*(6), 655–670. https://doi.org/10.1080/03075071003777716

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

Bloxham, S. & West, A. (2004). Understanding the Rules of the Game: Peer Assessment as a Medium for Developing Students' Conceptions of Assessment. *Assessment & Evaluation in Higher Education, 29*(6), 721–733. https://doi.org/10.1080/0260293042000227254

Boud, D. (2007). Reframing Assessment as if Learning were Important. In D. Boud & N. Falchikov (Eds.), *Rethinking Assessment in Higher Education* (pp. 181–197). Routledge.

Bradley, M. (2017). Postgraduate Taught Experience Survey 2017 - Understanding the experiences and motivations of taught postgraduate researchers. York: The Higher Education Academy. Retrieved 21 October 2019 from https://s3.eu-west-2.amazonaws.com/assets.creode.advancehe-document-manager/documents/hea/private/hub/download/ptes_2017_national_report_1568037562.pdf

Brooks, V. (2012). Marking as judgment. *Research Papers in Education, 27*(1), 63–80. https://doi.org/10.1080/02671520903331008

Brown, R. (2010). The current brouhaha about standards in England. *Quality in Higher Education, 16*(2), 129–137. https://doi.org/10.1080/13538322.2010.487699

Browne, A. L. (2016). Can people talk together about their practices? Focus groups, humour and the sensitive dynamics of everyday life. *Area*, *48*(2), 198–205. https://doi.org/10.1111/area.12250

Crisp, V. (2013). Criteria, comparison and past experiences: How do teachers make judgements when marking coursework? *Assessment in Education: Principles, Policy and Practice*, *20*(1), 127–144. https://doi.org/10.1080/0969594X.2012.741059

Deeley, S. J. & Bovill, C. (2017). Staff student partnership in assessment: Enhancing assessment literacy through democratic practices. *Assessment & Evaluation in Higher Education, 42*(3), 463–477. https://doi.org/10.1080/02602938.2015.1126551

DeLuca, C. (2012). Preparing teachers for the age of accountability: Toward a framework for assessment education. *Action in Teacher Education, 34*(5-6), 576–591. https://doi.org/10.1080/01626620.2012.730347

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). SAGE Publications.

Forsyth, R., Cullen, R., Ringan, N. & Stubbs, M. (2015). Supporting the development of assessment literacy of staff through institutional process change. *London Review of Education, 13*(3), 34 – 41. Retrieved 22 October 2019 from https://files.eric.ed.gov/fulltext/EJ1160159.pdf

Gale, K. N., Heath, G., Cameron, E., Rashid, S. & Redwood, S. (2013). Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC Medical Research Methodology, 13*(117), 1–8. doi:10.1186/1471-2288-13-117

Grainger, P., Purnell, K. & Zipf, R. (2008). Judging quality through substantive conversations between markers. *Assessment & Evaluation in Higher Education, 33*(2), 133–142. https://doi.org/10.1080/02602930601125681

Greatorex, J., Johnson, C. & Frame, K. (2001). Making the grade: Developing grade descriptors for accounting using a discriminator model of performance. *Westminster Studies in Education, 24*(2), 167–181. https://doi.org/10.1080/0140672010240207

Hair, J.F., Anderson, R.E., Tatham, R.L. & Black, W. C. (2010). *Multivariate data analysis* (7th ed.). Prentice-Hall.

Handley, K. & Williams, L. (2011). From copying to learning: Using exemplars to engage with assessment criteria and feedback. *Assessment & Evaluation in Higher Education 36*(1), 95–108. https://doi.org/10.1080/02602930903201669

Henson, R. K. & Roberts, J. K. (2006). Use of exploratory factor analysis in published research. *Educational and Psychological Measurement, 66*(3), 393–416. https://doi.org/10.1177/0013164405282485

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

Hunter, K. & Docherty, P. (2011). Reducing variation in the assessment of student writing. *Assessment & Evaluation in Higher Education, 36*(1), 109–124. https://doi.org/10.1080/02602930903215842

Hussey, T. & Smith P. (2002). The trouble with learning outcomes. *Active Learning in Higher Education, 3*(3), 220–233. https://doi.org/10.1177/1469787402003003003

Jessop, T. & Tomas, C. (2017). The implications of programme assessment patterns for student learning. *Assessment & Evaluation in Higher Education, 42*(6), 990–999. https://doi.org/10.1080/02602938.2016.1217501

Koh, K. H. (2011). Improving teachers' assessment literacy through professional development. *Teaching Education, 22*(3), 255–276. https://doi.org/10.1080/10476210.2011.593164

Leask, B. (2011). Assessment, learning, teaching and internationalisation–engaging for the future. *Assessment, Teaching & Learning Journal*, 11, 5–20. Retrieved 20 October 2019 from http://eprints.leedsbeckett.ac.uk/1191/1/Assessment,%20learning,%20teaching%20and%20internationalisation.pdf

MacCallum, R. C., Widaman, K. F., Preacher, K. J. & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research, 36*(4), 611–637. doi:10.1207/S15327906MBR3604_06

Mulder, R. A., Pearce, J. M. & Baik, C. (2014). Peer review in higher education: student perceptions before and after participation. *Active Learning in Higher Education, 15*(2), 157–171. https://doi.org/10.1177/1469787414527391

Nicol, D. (2010). 'The Foundation for Graduate Attributes: Developing Self-Regulation through Self and Peer Assessment'. Gloucester: The Quality Assurance Agency for Higher Education. Retrieved 21 October 2019 from https://www.reap.ac.uk/Portals/101/Documents/PEER/Project/QAA_GA_SR.pdf

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

Norton, L., Norton, B. & Shannon, L. (2013). Revitalising assessment design: What is holding new lecturers back? *Higher Education, 66*(2), 233–251. doi:10.1007/s10734-012-9601-9

O'Donovan, B., Price, M. & Rust, C. (2008). Developing student understanding of assessment standards: A nested hierarchy of approaches. *Teaching in Higher Education, 13*(2), 205–217. https://doi.org/10.1080/13562510801923344

Orr, S. (2010). 'We kind of try to merge our own experience with the objectivity of the criteria': The role of connoisseurship and tacit practice in undergraduate fine art assessment. *Art, Design and Communication in Higher Education, 9*(1), 5–19. https://doi.org/10.1386/adch.9.1.5_1

Orr, S. (2007). Assessment moderation: Constructing the marks and constructing the students. *Assessment & Evaluation in Higher Education, 32*(6), 645–656. https://doi.org/10.1080/02602930601117068

Payne, E. & Brown, G. (2011). Communication and practice with examination criteria. Does this influence performance in examinations? *Assessment & Evaluation in Higher Education, 36*(6), 619–626. https://doi.org/10.1080/02602931003632373

Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator, 46*(4), 265–273. https://doi.org/10.1080/08878730.2011.605048

Price, M., Rust, C., O'Donovan, B., Handley, K. & Bryant, R. (2012). *Assessment literacy*. Oxford Brookes University.

Quality Assurance Agency for Higher Education, The (QAA) (2018). The revised UK quality code for higher education. Retrieved 21 October 2019 from https://www.qaa.ac.uk/docs/qaa/quality-code/revised-uk-quality-code-for-higher-education.pdf?sfvrsn=4c19f781_8

Rust, C., O'Donovan, B. & Price, M. (2005). A social constructivist assessment process model: How the research literature shows us this could be best practice. *Assessment & Evaluation in Higher Education, 30*(3), 231–240. https://doi.org/10.1080/02602930500063819

Sadler, D. R. (2014). The futility of attempting to codify academic achievement standards. *Higher Education*, *67*(3), 273–288. doi:10.1007/S10734-013-9649-1

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. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*(2), 159–179. https://doi.org/10.1080/02602930801956059

Sambell, K., McDowell, L. & Montgomery, C. (2013). *Assessment for Learning in Higher Education*. London: Routledge.

Shay, S. (2005). The assessment of complex tasks: A double reading. *Studies in Higher Education, 30*(6), 663–679. https://doi.org/10.1080/03075070500339988

Smith, C. D., Worsfold, K., Davies, L., Fisher, R. & McPhail, R. (2013). Assessment literacy and student learning: The case for explicitly developing students' 'assessment literacy'. *Assessment & Evaluation in Higher Education, 38*(1), 44–60. doi:10.1080/02602938.2011.598636

Stowell, M. (2004). Equity, justice and standards: Assessment decision making in higher education. *Assessment & Evaluation in Higher Education*, *29*(4), 495–510. https://doi.org/10.1080/02602930310001689055

Taras, M. (2009). Summative assessment: The missing link for formative assessment. *Journal of Further and Higher Education*, *33*(1), 57–69. https://doi.org/10.1080/03098770802638671

Taras, M. & Davies, M. S. (2012). Perceptions and realities in the functions and processes of assessment. *Active Learning in Higher Education, 14*(1), 51–61. https://doi.org/10.1177/1469787412467128

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 M. Easterby-Smith & M. Lyles (Eds.), *Handbook of Organizational Learning and Knowledge Management*, (pp. 411–427). Blackwell.

William, D. & Thompson, M. (2008). Integrating assessment with learning: What will it take to make it work? In C. Dwyer (Ed), *The future of assessment: Shaping teaching and learning* (pp. 53–84). Lawrence Erlbaum Associates.

Zhao, N. (2009). The minimum sample size in factor analysis. Retrieved 21 October 2019 from https://www.encorewiki.org/display/~nzhao/The+Minimum+Sample +Size+in+Factor+Analysis

Corresponding Author

Claudia Rosenhan, Moray House School of Education and Sport, University of Edinburgh, Edinburgh, EH8 8AQ, UK. Email: claudia.rosenhan@ed.ac.uk

Acknowledgements
We are grateful for the constructive comments of the editor and reviewers of this paper. We thank the University of Edinburgh IAD for their support in terms of the PTAS grant to carry out this research.