9th International Conference on Higher Education Advances (HEAd'23)
Universitat Politècnica de València, València, 2023
DOI: http://dx.doi.org/10.4995/HEAd23.2023.16264

The complexity of grading student work and the reconstruction of the meaning of criterion-referenced assessment

Liao Liang
The Office of University General Education, The Chinese University of Hong Kong, Hong Kong.

Abstract

This study explores how assessment criteria are applied in grading student work. It is found that explicit assessment criteria do not work as authoritative guidance as expected and that tacit criteria are more decisive in awarding a certain grade. Various sources that form idiosyncratic tacit criteria are identified. These sources, including values on curriculum and assessment, teaching experiences, and the way of supervising grades, work as different social contextual elements that influence judgements on grading students' work. The study suggests that tacit criteria and the different sources that form the tacit criteria need to be identified, perceived, and communicated in the community of practice to reduce grade variability and achieve a shared understanding of grading.

Keywords: *Criterion-referenced assessment; grading process; assessment criteria; tacit criteria; shared understanding.*

1. Introduction

Criterion-referenced assessment (CRA) has been recognised as a common method for assessing what students have achieved in a specific course (Sadler, 2017; Svennberg, Meckbach, & Redelius, 2018). Compared with norm-referenced assessment, which evaluates a student's work by ranking it within a group (Lok, McNaught & Yong, 2016), CRA is based on scoring according to a series of explicit criteria (Popham, 1978, 2014). CRA has many positive impacts on student assessment. As grading in a bell curve is no longer required in criterion-referencing, pursuing higher grades becomes a matter of an individual's efforts, rather than competing with others. Additionally, by providing explicit assessment criteria, students can form a clearer understanding of what the assessment task entails, how it will be judged, and what the level of achievement is (O'Donovan, Price, & Rust, 2001).

Although CRA has many advantages, it is still need to note that the grading process is complex and intermingled with assessors' internal judgement (Bloxham, Boyd & Orr, 2011; Orr, 2007). Previous researches indicate that besides explicit criteria, tacit criteria commonly exist in grading judgement (Sadler, 2005, 2009). Tacit criteria are usually personally determined, which could easily lead to a substantial difference between different assessors when judging the same student's work (Bloxham, 2016a). Consequently, how an assessor applies the explicit and tacit criteria to award a grade greatly affects the reliability and quality of an assessment. To review and to reflect upon the quality of student assessment, a thorough investigation of the process of grading judgement is necessary.

2. Literature review

2.1. The complexity of grading process

Several studies reveal similar findings, stating that making judgements in grading is more likely to happen implicitly rather than being based on external sources. In other words, the real standards are locked inside the teacher's head (Bloxham et al., 2016a; Grainger, Purnell, & Zipf, 2008; O'Hagan and Wiggleswoth, 2015). Furthermore, the interpretation of each of the standards for the criteria is highly diverse among assessors, thus leading to a large gap in assessment results for the same student's work. For example, Bloxham et al. (2016a) invited 24 teachers from four isciplines to assign five students' writing works respectively; they found substantial differences in the awarded grades. Only one of the 20 pieces was graded in the same rank by six teachers in any of the disciplines. The assignment gap for the same student work was at least three positions. Bloxham points out that assessors' internal standards framework plays an important role in awarding grades. This internal framework, also called tacit criteria, commonly exists among assessors, even when explicit criteria are provided (Sadler, 2005, 2009; Gonzalez & Burwood, 2003). The nature of tacit criteria is recognised as idiosyncratic (Matshedisho, 2020; Adie, Lloyd, & Beutel, 2013), to be

perceived unconsciously and expressed uneasily (Sadler, 2009, 2013). Tacit criteria increase the opacity of the grading process and make grading complicated to understand.

Beside the effect of tacit criteria, grading judgement is also contextually influenced. Shay (2004) used conceptions of 'field' and 'habitus' to discuss the basis of decision making in assessment. It concludes that assessing complex tasks is a socially situated interpretive act. Assessors' interpretations are shaped by their disciplinary orientations, years of experience, and level of involvement with students. These interpretations are constituted not only to sustain systems of belief but also to maintain identities and interpersonal relations. Given social practice, grading is not merely decided individually. More precisely, it is closely related to the backgrounds and experiences of assessors and the institutional culture they encounter (Watty et al. 2014; Zahra et al. 2017).

3. Research questions

This study aims to explore how teachers internalise, interact with, and interpret explicit criteria to illuminate the mechanism of using explicit criteria to frame internal decisions. The role of tacit criteria and its influence in grading judgement will also be investigated. The following research questions are considered:

1. What are the roles of explicit and tacit criteria in determining the final grade?
2. How are tacit criteria formed and how do they influence the assessor's grading considerations?
3. What insights are proposed to improve CRA practice by revealing the formation of tacit criteria and the nature of the grading process?

4. The case

4.1. Background

A general education programme of University A in Hong Kong was selected as the case study. There are currently 28 full-time teachers involved in this programme to teach two GE courses. In 2018, CRA was required to be fully implemented in all faculties and teaching units at University A. A task force was initiated in the GE programme to respond to the shift, which involved five teachers from two GE courses and one researcher. Grading rubrics were developed based on the intended learning outcomes to echo the outcome-based approach advocated by the Quality Assurance Council. In addition to developing grading rubrics, several meetings were also held to discuss CRA and how to use the grading rubrics.

4.2. Data collection

Interviews were the main sources of data collection. Seven GE Programme teachers were invited to attend one-on-one interviews, and each interview lasted 1–1.5 hours. Interview questions addressed teachers' considerations of using grading rubrics, understanding the assessment criteria, and their views on CRA, GE learning goals, and assessment. Five of the teachers had more than six years of teaching and assessment experiences in this programme. Two were novice teachers with less than one year of work experience. In addition to conducting interviews, the author joined the task force, developed grading rubrics with other task force members, and attended all meetings on CRA. Other sources of data included reflective notes of CRA meetings, emails of the discussions among teachers, and a brief survey of teachers. As a complement to teachers' thoughts on CRA and grading judgement, some informal talks were also collected and used to support teachers' viewpoints. In CRA meetings, the author was an observer and took notes on what happened in the field, such as questions raised, interactions, and responses among teachers.

5. Results

5.1. Role of explicit and tacit criteria in grading judgment

All teachers reviewed the descriptions of explicit criteria and expressed their evaluations on them. Grading is not like enacting orders under the guidance of external requirements; it is rather an application of internalised criteria. A typical problem of internalising explicit criteria was about abstract expressions such as 'focused topic', 'relevant evidence', and 'logical and specific conclusion'. Teachers noted that these wordings were very general and needed more explanations.

Because elaborating and communicating explicit criteria were not recognised as a regular practice, most teachers took individual interpretations on these explicit criteria as common; nevertheless, they did not realise the necessity of illuminating their own interpretations. Among the seven teachers, only one showed the notes, which explained in detail what the abstract descriptions in each criterion mean, whereas the other teachers did not specify the meaning of those abstract wordings. It seemed that teachers had examined the meaning of criteria implicitly, but they did not explicate and reflect their interpretations or seek other ways of interpretations.

Sometimes, personal interpretations were in conflict with the explicit criteria. Three teachers noted that they did not completely agree with the standards for each criterion. Even so, they had their way of reconciling the external standards with their internal standards. For example, a teacher noted that they would grade leniently on some criteria which were set higher in their views:

> *'I use different standards in grading reflective journal and term paper. For reflective journals, standards will be set lower since they're completed in the middle of the course and where in this stage, understanding is most important other than higher-order thinking skills. I would give the score leniently in the reflective journal. If students show a good understanding of texts, I will give them a good grade even if they do not exert complex thinking skills'.*

Not every teacher would refer to all criteria listed in the grading rubric. Among the seven teachers, three indicated that they would make a holistic judgement, rather than make one by referring to every criterion. One teacher stated disagreement in dividing criterion into several parts, for it will spoil the integrity of student achievement and make each part of performance separate and irrelevant to each other:

> *'All criteria should be related, and they actually reflect the whole. I will focus on the integrity of student's performance'.*

5.2. The formation of tacit criteria and its influence on grading considerations

What qualities should be cultivated for students affect the interpretation and usage of explicit criteria. Teachers held different expectations regarding the GE curriculum. These individual notifications were not fully covered either by learning outcomes or explicit criteria. Both learning outcomes and assessment criteria were worked out by the representatives of teachers, instead of all of them. Consequently, these uncovered expectations became tacit criteria. This may explain why personal interpretations and various grading strategies were common. For example, a teacher who regarded curiosity as an essential characteristic noted that 'he would adjust the descriptions of explicit criteria and inject the element of curiosity into them'. Another teacher who did not hold the value of curiosity would not interpret the criteria in this way.

Compared to novice teachers, experienced teachers were more flexible in using explicit criteria. Experienced teachers were able to interpret the meaning of criterion from a broad perspective, instead of translating the keywords of a criterion literally. In addition, they were more confident in determining the 'right' grade for students. Experienced teachers seemed to know how to adapt explicit criteria to make grading decisions more consistent with their tacit criteria. Such an adaptive strategy was scarce in novice teachers. For novice teachers, interpreting explicit criteria involved many uncertainties. It was difficult for them to connect teaching experiences with the criteria to give an appropriate grading judgement. Unlike 'technically adjusting' the meaning of explicit criteria, as done by experienced teachers, novice teachers more often adopted direct methods. They added or deleted criteria to make explicit criteria more consistent with their internal judgement, although these methods would lead to various patterns of criterion-referencing. Confidence in grading was also inadequate

among novice teachers. Assessment training was more frequently mentioned by these participants.

> *'I hope I could receive some trainings on how to award a grade. Differences in each grade level, especially between A- and B+ or some similar cases, were not very clear for me to identify. Maybe we could invite some grading experts and share with us how to grade accurately'.*

The other difference between grading considerations of experienced and novice teachers was their adoption of holistic judgement. Two novice teachers indicated they would adopt analytic judgement, whereas, of five experienced teachers, three noted that they preferred holistic judgement. Although task force members communicated with experienced teachers that analytical judgement was indispensable in criterion-referencing, it was hard to change their views. Adopting a holistic judgement was related to the issue of trust and efficiency. For some teachers, holistic judgement would be more reliable. Teachers worried that the final grade might be unmatched with their original judgement by adding sub-scores to each criterion. The other consideration was time. Teachers noted that analytic judgement was too time-consuming:

> *'I have tried grading according to the criterion one by one and find it wasted too much time. After doing this, I still need to review the grade and examine if the judgement is appropriate with a holistic approach'.*

Besides the influence of teaching experiences that form the different basis of grading judgement, many teachers treated grade distribution as a boundary to keep assessment results 'safe'. In other words, although CRA called for 'giving students a real grade' and 'abandoning grading in a bell curve', in the grading practice, norm-referenced grading was still being tacitly used.

There were two patterns of tacit practices. One examined the entire grade distribution after grading and compared it with the previous policy. If any 'abnormal' grades were discovered, they would be adjusted. Some teachers chose to examine grade distribution in the mid-term to adjust the coming grading strategy:

> *'I would first examine the results of the reflective journal, if the grade distribution is not good, I would reconsider the grading of the term paper. If there are too much A's in term papers, I probably would make some adjustments'.*

Another way of applying the principle of grade distribution is to set a line to guide grading. In this way, grades were regulated during the grading process to satisfy the hidden criteria of the specific distribution. This approach was more undisclosed and even discerned that teachers were using norm-referencing. In sum up, Grading was still reckoned on norm-

referencing because of the obscure policy of monitoring grades. In the brief survey results, many teachers indicated that grade distribution was still applied in their grading practice.

6. Discussion

Although CRA has been practised over the years, this study revealed that the educational ideas of CRA are not thoroughly manifested in grading practice. The core idea of CRA is to consider the explicit criteria reflectively, so as to minimize the subjectivity in grading affected by student image (Shay, 2004), peer relationship (Zahra et al., 2017), department culture (Deneen & Boud, 2014) and so on. The marginalisation of explicit criteria and amplification of the power of tacit criteria may be due to a superficial and inappropriate understanding of CRA. Elaborating and reconstructing the meaning of CRA is, therefore, necessary to clarify some common misconceptions of criterion-referenced grading and to ensure a fairer grading process.

First, explicit criteria should not be regarded as the absolute standard, nor should personal interpretations be taken for granted without examining them in the community of practice. As judgement making based on teachers' professional experiences is in terms of whether academic freedom is respected and protected (Sadler, 2011; Lee, 2006), while explicit criteria provide a basis for calibrating various individual judgements (Sadler, 2013; Dracuo, 1997), balancing the function of tacit criteria and explicit criteria is critical in deciding a consistent and fair grading. Criterion-referencing does not favour the approach of maximising unified standards, nor does it favour the opacity of the grading process. Letting explicit criteria and tacit criteria dance together would make them the true criterion-referencing. To balance the power of tacit criteria, it is critical to have empathy regarding understanding the rationality of grading judgement by discussing it in the community of practice (Bloxham, Hughes, & Adie, 2016b). It requires putting aside individual liberty in making grading judgements and embracing the liberty of the whole (Sadler, 2011; Berlin, 1969).

Second, assessment criteria are in the central position of CRA. Therefore, the appropriate usage of the assessment criteria needs to be clarified. This study argues that CRA is a social constructive practice (Rust, O'Donovan, & Price, 2005), which contains two levels of meaning. The first level represents how assessment criteria should be used among assessors. It suggests that assessment criteria should be identified, selected, discussed, and communicated in the assessment team. Furthermore, teachers' value in the curriculum needs to be elaborated and absorbed into the learning outcomes and assessment criteria pool (Watty et al., 2014). The second level represents how assessment criteria should be used among teachers and students. Assessment criteria and how they are interpreted by teachers should be communicated with students (Bearman& Ajjawi, 2021). Social constructive practice means to initiate a dialogue with students to address the meaning of criteria and build an

understanding on grading judgement. In this way, CRA plays a role of assessment for learning which emphasises the concept of student-centred approach and empowers students' experiences of assessment by involving them into the grading process (Sadler, 1987, 1989).

Third, CRA is a different grading approach other than the regularly-adopted norm-referencing approach. How grades are supervised and what should be accountable for a given grade substantially affects whether assessors will adopt real criterion-referencing. If departments or universities still take grade distribution as the accountability objective, it is no surprise that teachers will finally depend on grade distribution to award a grade. In this case, explicit criteria would be marginalised and, more notably, students may not get the real grade. Therefore, grade distribution as the only standard for reviewing students' grades should be abandoned. The focus can be on how the assessment criteria were set. Do they match course content? Does the assessment task reflect the assessment criteria well? How do students regard the assessment criteria? To answer these questions, the paradigm of grade review needs to abandon the result-oriented approach and focus more on the grading process, the decision making of the assessment team, and students' feedback on grades and grading.

7. Conclusion

For a long time, many educators believe that the variability of grading is normal and inevitable due to the existence of tacit criteria. However, should this passive attitude toward grading continue, it would increase students' negative perceptions of assessments. The current study argues that criterion-referencing can reduce subjectivity in grading. This is realized by reflectively examining the grading judgement, better understanding our own tacit criteria, and discussing the explicit criteria in the community of practice. Revealing the formation of tacit criteria and illuminating the sources can help assessors become aware of the unperceived hidden standards influenced by complex social elements, consider grading more carefully, pay attention to grade variability issues, conduct grade moderation effectively, thus making grading and grades more fair. By recognising the characteristics of explicit and tacit criteria and build a balance between them, the spirit of CRA, which emphasises the assessment for learning, can be achieved.

References

Adie, L., Lloyd., Beutel, D. 2013. Identifying discourses of moderation in higher education. *Assessment and Evaluation in Higher Education* 38(8): 968-977

Bearman, M., and Aiiawi, R. 2021. Can a rubric do more than be transparent? Invitation as a new metaphor for assessment criteria. *Studies in Higher Education* 46(2): 359-368.

Berlin, I. 1969. Two concepts of liberty, In: I. Berlin, *Four Essays on Liberty*. Oxford University Press: London, 118-172.

Bloxham, S., P. Boyd, and S. Orr. 2011. Mark my words: the role of assessment criteria in UK higher education grading practices. *Studies in Higher Education* 36(6): 655-670.

Bloxham, S., B. den-Outer, J. Hudson, and M. Price. 2016a. Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria. *Assessment and Evaluation in Higher Education* 41(3): 466-481.

Bloxham, S., C. Hughes., & L. Adie. 2016b. What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices. *Assessment and Evaluation in Higher Education* 41(4): 638-653.

Deneen C, Boud D. 2014. Patterns of resistance in managing assessment change. Assessment and Evaluation in Higher Education,39(5):577-591.

Dracup, C. 1997. The reliability of marking on a psychology degree. *British Journal of Psychology* 88(4): 691-708.

Gonzalez Arnal, S., and S. Burwood. 2003. Tacit knowledge and public accounts. *Journal of Philosophy of Education* 37(3): 377-391.

Grainger, P., K. Purnell, and R. Zipf. 2008. Judging quality through substantive conversations between markers. *Assessment and Evaluation in Higher Education* 33(2): 133-142.

Lok,B., McNaught,C.,&Yong,K. 2016. Criterion-referenced and norm-referenced assessments: compatibility and complementarity *Assessment and Evaluation in Higher Education,*41 (3): 450-465.

Lee,D.E. 2006. Academic freedom, critical thinking and teaching ethics. *Arts and Humanities in Higher Education: an International Journal of Theory, Research and Practice,*5(2):199-208.

Matshedisho, R. 2020. Straddling rows and columns: students' (mis)conceptions of an assessment rubric. *Assessment and Evaluation in Higher Education* 45(2): 169-179.

O'Donovan, B., Price, M., & Rust, C. The students experience of criterion-referenced assessment(Through the introduction of a common criteria assessment grid)[J]. *Innovations in Education and Teaching International*,2001,38(1):74-85.

O'Hagan,S.,& Wigglesworth,G. 2015. Who's marking my essay? The assessment of non-native speaker and native-speaker undergraduate essays in an Australian Higher Education Context. *Studies in Higher Education,*40(9):1729-1747.

Orr, S. 2007. Assessment moderation: Constructing the marks and constructing the students. *Assessment and Evaluation in Higher Education* 32(6): 645-656.

Popham, J. 1978. Criterion-referenced measurement. Prentice-Hall: Englewood Cliff.

Popham, W.J. 2014. Criterion-referenced measurement: half a century wasted? *Educational Leadership* 71(6): 62-66.

Rust, C., B. O'Donovan, and M. Price. 2005. A social constructivist assessment process model: how the research literature shows us this could be best practice. *Assessment and Evaluation in Higher Education* 30(3): 231-240.

Sadler, D.R. 1989. Formative assessment and the design of instructional systems. *Instructional Science* 18(2): 119-144.

Sadler, D.R. 1987. Specifying and promulgating achievement standards. *Oxford Review of Education* 13(2): 191-209.

Sadler, D.R. 2005. Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education* 30(2): 175-194.

Sadler, D.R. 2009. Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education* 34(2): 159-179.

Sadler, D.R. 2011. Academic freedom, achievement standards and professional identity. *Quality in Higher Education* 17(1): 85-100.

Sadler, D.R. 2013. Assuring academic achievement standards: from moderation to calibration. *Assessment in Education: Principle, Policy & Practice* 20(1): 5-19.

Sadler, D.R. 2017. Academic achievement standards and quality assurance. *Quality in Higher Education* 23(2): 81-99.

Shay, S. 2004. The Assessment of Complex Performance: A Socially Situated Interpretive Act. *Harvard Educational Review* 74(3): 307-329.

Svennberg, L., Meckbach, J., Redelius, K. Swedish PE teachers struggle with assessment in a criterion-referenced grading system. 2018. Sport, *Education and Society*, 23(4): 381-393.

Watty. K., Freeman. M., Howieson. B., Hancock. P., O'connell. B., Lange. P., and Abraham. A. 2014. Social moderation, assessment and assuring standards for accounting graduates. *Assessment and Evaluation in Higher Education* 39(4): 461-478.

Zahra, D., I. Robinson, M. Roberts, L. Coombes, J. Cockerill, S. Burr. 2017. Rigor in moderation processed is more important than the choice of method. *Assessment and evaluation in Higher Education* 42(7): 1159-1167.