Authors: Peter Grainger and Katie Weir

TITLE: An alternative grading tool for enhancing assessment practise and quality assurance in higher education

Introduction

Criteria-referenced and standards-based assessment regimes are becoming the norm in educational systems throughout Australia and indeed the rest of the world. At the secondary school level in Australia these assessment regimes have been mandated in some state systems for over 30 years. However, in tertiary educational contexts the uptake is slower, even though criterion-referenced assessment has framed university assessment practices for at least a decade. According to Sadler (2005), higher education institutions have become increasingly committed to making assessment and grading more effective in promoting student learning. Nevertheless, despite some big improvements in assessment practices, the current emphasis on standards creates new challenges to some tertiary educators whose assessment practices need reframing and reforming. According to Boud and Associates (2010) any improvements will require particular support for academics to develop assessment expertise so they can meet university course and program requirements.

In response to a recently published report on the higher education in Australia known as *The Bradbury Report* (DEEWR, 2008), the federal government created a new agency for regulating the university sector called the *Tertiary Education Quality and Standards Authority* (hereafter referred to as TEQSA). A key goal of this new regulatory authority is the development of a set of threshold standards for every level of program offering at Australian universities. This set of standards is contained within the *Higher Education Standards Framework* (Department of Industry, Innovation, Science, Research and Tertiary Education, 2011, p.16). As a consequence of these recent reforms there is a need to investigate alternative assessment approaches that can accommodate TEQSA's new standards and that are accountable to any standards-based assessment mandates. For example, according to item 5.5 of the TEQSA framework there is a requirement to benchmark assessment standards against similar accredited courses of study offered by other higher education providers. In order for this cross-institutional benchmarking to occur a common understanding of assessment principles is a necessity (Boud and Associates, 2010), as is a common language

for describing and interpreting assessment criteria and standards that is shared across tertiary education contexts.

Equally important to a shared understanding of assessment principles, is a need to reform existing assessment practices so that tools used to evaluate student learning are truly criterion-referenced and standards-based. These shifts to a standards-based curriculum framework are in keeping with national and international attention on standards in research quality (e.g. OECD, 2011), and increasing accountability for research quality (e.g. the UK's HEFCE's Research Excellence Framework and Australia's 'Excellence for Research in Australia'). As with research, there is an increasing requirement for universities to demonstrate appropriate standards of teaching and learning plus quality assurance procedures to meet the requirements of professional accrediting agencies such as TEQSA. Grading tools, also known as criteria sheets, are a key to quality assurance but their design and efficacy for judging student work often varies within and across tertiary education contexts.

The aim of this paper is to report on the findings from a pilot research project that looked at the experiences of a community of scholars that were implementing a new style of grading tool for judging the quality of their students' work. The project focus was on how the new grading tool assisted staff in improving their capacity to implement standards-based assessment in line with current 'best practice'. The research took place at a regional university in South East Queensland where the design and deployment of assessment criteria sheets in teacher education courses was in need of reform. Although the context is Australia, we believe the lessons to be learned are applicable in all international contexts for tertiary assessment. Assessment reform is a major agenda in all educational systems as it is a significant driver of student motivation.

Literature review

Assessment is a key driver in fostering student learning and "is one of the most significant influences on students' experiences of higher education and all that they gain from it" (Boud and Associates, 2010, p. 1). This has led to a renewed focus on improving assessment practise in tertiary education because of the powerful impact it has on the quality of learning.

In tertiary educational institutions, assessors commonly assess student work against a set of pre-determined criteria and achievement standards. This system is referred to as

criterion-referenced assessment and contrasts with traditional, norm-based assessment system that ranks students' achievements against pre-defined outcome. Criterion-referenced, standards-based assessment tasks are normally accompanied by a grading tool that outlines the specified criteria being assessed, the attainable levels of achievement or standards, and standards descriptors that capture the expected quality of work. Existing matrix-style grading tools or rubrics (see Figure 1 below) tend to contain standards descriptors that lack explicitness about the quality of student performance expected. If the description of the expected behaviour is not explicit then judging a student's learning performance becomes problematic. This is especially the case when multiple assessors are involved who may have differing interpretations of 'quality' and the end result is inconsistent grading that produces unreliable evidence of student learning outcomes.

It is acknowledged that academics who are experienced assessors possess tacit knowledge of what quality looks like in student work (Race, 2006). This concurs with Sadler's (2011) judgement that competent appraisers can consistently identify quality when they see it. This tacit knowledge has been shown to enable assessors to make accurate interpretations of sometimes vague descriptions of student behaviour in order to discriminate between standards or levels of achievement (Grainger, Purnell & Zipf, 2008) which means there is no compromise of assessment integrity and reliability in terms of teacher judgments.

However, not all academics understand or are experienced with sound assessment practices and it takes some years to get to know how to align evidence of quality with relevant achievement standards and achieve consistency of judgement over time. For assessors who are unclear about learning quality, vague assessment grading tools are not, in fact, objective arbiters of performance, nor are they defensible, nor do they encourage consistency of teacher judgments. Hence, these grading tool deficiencies represent major challenges to what Sadler (2010) refers to as "grade integrity".

An additional dimension to the problem with existing assessment practices is the multiple roles of a grading tool. Primarily, grading tools are designed by and for assessors to evaluate the quality of student work. They also enable assessors to construct feedback to students that aligns with the criteria for marking. From a pedagogical perspective the grading tool can act as a guide for students' learning by making explicit the evidence they must demonstrate through the assessment task. Because grading tools have more than one purpose and audience they are complicated texts to construct, especially the standards descriptors to

ensure they adequately differentiate between the levels of achievement. This often results in standards descriptors that are ambiguous, open to interpretation and make assumptions that the user is familiar with the language used (Sadler, 1987).

According to Sadler (2009a), standards descriptors must be precise enough to allow for unambiguous determinations and they must consist of statements that accurately describe the properties which characterise a learning behaviour at its designated level of quality. From author experience current practice with standards in universities do not meet these requirements and our claim is that reforming these practices is constrained by an almost universal reliance on matrix-style rubrics.

Insert Figure 1 here: Matrix model

The matrix format of these rubrics suggests that assessors must ‘fill up’ the grid which forces them to manufacture descriptors for every cell, even in cases where the criterion may not have the depth to derive (typically) five different qualities of performance. Matters (2005) claims this practice can be counterproductive because it tends to reduce the multidimensionality of complex tasks that measure a range of knowledges, skills and dispositions. The implications for marking are that some standards descriptors may not accurately capture the evidence assessors must look for in students’ work. In cases like this, the integrity of the final judgement is brought into question as assessors ignore the grading tool and evaluate student work based on subjective interpretations of what constitutes quality. Evidence of this is reported by Klenowski, et al (2007) who witnessed the practice when reviewing teacher moderation practices.

There are other problems with rubrics, particularly the simplicity of the matrix format which can disguise real complexities in its design and use. According to Matters (2005), underpinning the standard matrix are two spurious assumptions that have implications for the validity and reliability of the assessment outcome. The first assumption is that each move across the standards is the same ‘quantum leap’ which, in practice, is not the case mainly because the significance of the move is dependent on the discriminating capacity of each descriptor. The second assumption, that each assessment criteria carries equal weight, or that they can be assigned arbitrary numerical weighting, is not always perceived or interpreted as such by the user.

Other research that critiques matrix-style rubrics includes Klenowski's (2007) study which found a major issue with the matrix format is that assessors need to commit their judgement to a single cell and this creates a psychological barrier to making more holistic judgments about the quality of student work. Sadler's (2007) work demonstrates how holistic judgments are further constrained by 'fine grained' or atomised outcomes or standards indicators. Whereas Reddy's (2014) research questions the number of criteria in a rubric and finds that there is no consensus on this issue and limited studies into its implications.

It is our belief that some of these inherent deficiencies of the matrix-style rubric can be alleviated by an alternative grading tool, the *Continua Model of a Guide to Making Judgments (GTMJ)*, illustrated in Figure 2.

Insert Figure 2 here: Exemplar of a Continua model of a GTMJ

Originating in secondary school contexts, this model of a GTMJ is not commonly applied at the university level although author experience shows it can be successfully implemented in some teacher education courses albeit adapted to accommodate different institutional assessment policies and practices

The design and layout of the continua model afford it some unique characteristics for a grading tool. The first of these is the use of arrows to more accurately depict learning as a continuum of increasing quality rather than a series of discrete, measurable units. Furthermore, the tip of the arrow is deliberately longer than the highest standard descriptor to acknowledge the predictive nature of this text and that the highest achievement possibly defies description. The second unique feature is positioning the standards descriptors along the continuum as a 'region' of quality rather than a fixed zone. According to Matters (2005), this is a major advantage of the continua model because this arrangement acknowledges that the even the best-written generalised standards can never be sharply defined nor communicated with absolute precision. This arrangement also means that the number of standards can vary from criterion to criterion, as can their relative placements, removing the necessity to 'fill up' a matrix.

Another distinguishing feature of the *Continua Model of a GTMJ* is the language used to construct the standards descriptors for each assessment criteria. This aspect of the continua model is called 'nestedness' which ensures that only the defining quality of each standard, the discernible difference, is identified for each descriptor. Nestedness also means that a student

who achieves the very highest standard also displays the behaviours of the previous pass standards. Hence, 'nestedness' precludes the necessity of repeating defining behaviours in multiple descriptors, so the standards descriptors become more efficient at eliminating superfluous and repetitious words that can clutter understanding about the assessment expectations.

In this study the *Continua Model of a GTMJ* is the grading tool that academic staff in a teacher education program at a regional university were asked to implement. The research aimed to examine staff experiences with this alternative model of grading tool and any influence it has in reforming their assessment practices and building their assessment capacity. Thus, the key research questions for this project are

Does the *Continua model of a GTMJ* enhance the grading process for staff?

Do staff perceive that this alternative grading tool enables them to build their capacity to understand and implement standards-based assessment in higher education contexts?

Methodology

In order to answer these research questions and get some sense of the efficacy and suitability of this alternative grading tool, a pilot study was conducted at a regional Australian university with 25 participants who were all teacher educators. These academics participated in professional development workshops that targeted the problems with current assessment practices and assisted staff in redesigning their marking rubrics using the *Continua model of a GTMJ*. A community website was set up for the project participants to share their ideas and the process of revising their grading tools. The website also served as a place where participants could share their experiences and assessment outcomes from implementing these new grading tools. A handful of project participants were interviewed so that further insights could be obtained regarding staff perceptions of the utility and efficacy of the *Continua model of a GTMJ* in relation to improving their assessment outcomes. For the purpose of this pilot study the staff selected to be interviewed was limited to five participants. These staff were self-selected and willing to offer further insights into their practise. Despite being a small sample the group had a range of tertiary experiences (i.e., they were employed at different academic levels). The interview questions covered a range of topics that would enable comparisons between existing approaches to grading student work and staff experience implementing the alternative, *Continua model of a GTMJ*. All participants had

previous experience with other marking rubrics, but none had any familiarity with the *Continua model of a GTMJ*. The data from these interviews was transcribed and analysed for emerging patterns using thematic analysis.

The idea of conducting this research as a pilot study was to ‘test’ for the effects of employing the *Continua model of a GTMJ* within a familiar context, that of teacher education, before broadening our research perspective and considering other tertiary learning contexts. Therefore, generalizability of the findings from this study is limited to informing future, larger-scale research. It can be argued however, that this pilot study has value outside this particular tertiary education context because it introduces an innovative approach to grading student work more holistically that potentially could improve quality assurance of assessment in universities more generally.

Results and discussion

The thematic analysis of the interview responses indicate that staff perceived the design of the ‘new’ tool to have some significant advantages over the traditional matrix-style rubric with regard to explicitly identifying different standards, simplifying and validating the grading process and streamlining moderation procedures. What the analysis highlighted was some symmetry in the data in that the issues that arose when participants employed matrix-style rubrics tend to be resolved by employing the alternative grading tool, the *Continua model of a GTMJ*. The following discussion presents the data in support of these claims beginning with look at staff perceptions of the grading process using matrix-style rubrics. Direct quotes from participants’ responses are italicised for ease of reading.

In the main, the academics who offered comment on their current practice using matrix-style rubrics often deferred to broad judgements such as, *not terribly good; too subjective; difficult for students to understand*. When respondents provided more specific details about the issues with the matrix format, the most common theme emerged to be the articulation of standards, and more specifically, the specificity of the description of quality for each standard. The data indicates that this is the main factor in rendering the grading tool difficult to communicate to students, difficult to mark their work and difficult to moderate between markers.

The other main issue to emerge from the responses relates directly the matrix format, its design and the implications for constructing these grading tools. One participant’s claim

illustrates this issue: *Having to put something in every box, for each level, can be hard sometimes; while another expressed the problem asrepetitive with many cells that use adverbs and adjectives to differentiate levels. This is not a good way to explain the differences and doesn't show what the improvement has to be to get to the next level.* These findings correlate with those cited in the literature reviewed for this study and support the observation that academics feel under pressure to 'fill the grid' rather than clearly indicate the evidence they are looking for.

In this regard, participants pointed to a number of advantages using the 'new' continua model, especially noting the language that is used:

It was easier to mark because of the shorter descriptions, the emphasis on new behaviour and just the key words like 'synthesize', 'critically reflect'. The vocabulary used in the new model is much better than what we see on many criteria sheets which is quite vague. This allows all of us to focus on what the level typically evidences and not have to interpret what the words mean.

The new model is much better in describing the action for each level of achievement. It is not just about filling a box with some description.

The responses also show that the alternative grading tool did not just make the assessment expectations more explicit for assessors, it was also perceived by participants to have an impact on students understanding of what is expected:

Recently I rewrote a criteria sheet using the new model-I visualised the task-What does it (student performance) look like? The new model made me do this because I had to write very explicit descriptions (as standards descriptors) on the new criteria sheet. Baseline behaviours can be defined and one can step up and step down.

This new model encourages this. I described the new re written criteria sheet to students who clearly understood what they had to do because of the explicitness. They said "that makes sense". Students could see the progression from one level to another. There is no need to guess what is in the teacher's head. It took me much less time to explain the criteria sheet because I did not have to explain the descriptions-they were explicit.

In addition, I have gone through the process of rewriting criteria sheets based on the new model. I have showed it to students and understood it well. They have commented that it is much clearer and more explicit-they knew what to do in the task as a result of reading the

descriptors. They liked the 'nestedness' aspect of the new model.-the step ups and step downs are obvious. It is easier to see if the student has demonstrated the standard.

What the interview data did reveal was the benefits of the continua model were not limited to its design and construction, the new model was also perceived as an important link between the processes of grading student work and then moderating the final marks awarded by each assessor. In relation to the grading process participants commented on the utility of the continua model in terms of the explicitness of descriptors, the continuum of quality and the 'nested' aspect of the standards descriptors:

I found the my new rewritten criteria sheet quicker to grade thanks to the collapsed criteria and the explicitness for the descriptors and the fact that I only articulated the new behaviour at each level rather than repeating everything.

As a staff member who is marking the task you can say a simple yes or no to each step within the criteria, if it is a yes then move up the continua, if it's a no then the grade is decided on the successful standard achieved. It removes the subjective of how well is it done and becomes has it been done.

It seems easier to make an on-balance judgement using this new model.

These responses indicate that if this new tool simplifies the grading process because it is easier to interpret the standards descriptors, then a consequent, positive flow on effect for the moderation process is to be expected. The data shows this assumption to be true as the following responses demonstrate:

Markers liked it during moderation and they found it good to use. I think it streamlines the moderation process, especially the 'nestedness' aspect which simplifies everything because the marker does not have to justify the decision (it's clear and explicit on the sheet). Visually, it has advantages-it is easier to show the level using the new model.

In terms of moderation, no one had taught this course before. The process was really simplified using the new criteria sheet-it was easy to engage with.

As there is little subjective assumptions made on "how well" a task is done moderation is simplified and only really required for 'borderline' achievements. Moderation becomes a pleasant task rather than a discussion to defend what you think is good or very good.

In addition to the streamlined aspect, one participant noted its impact on the decision making process, in terms of ‘evidence-based’ decision making. Sadler (2009b) noted an ‘impost of authority/seniority’ when describing some moderation activities which are impacted by the status of the course coordinator which served to make the decision making less than valid or reliable. This impost of authority is evidenced in the following comment:

Moderation is streamlined because it is fairly cut and dried. In the past, moderation was about acquiescence, not evidence based, because the tutors would basically agree with the expert course coordinator. The new model does not encourage this because of the focus on explicit behaviours which are identified in the descriptors. It’s evidence-based, not opinion based because there is no need for an explanation of terms like good or very good.

It can be inferred from the above comments that the academics who took part in this study were not satisfied with their current rubric design and implementation practices. Participants indicated that the layout of the continua model was more visually appealing in two ways and enabled a focus on the discriminating qualities of the descriptors. The use of a continuum and the concept of a ‘region of quality’ rather than the defined boundaries between standards that matrices create is a relatively recent innovation which is rarely mentioned in a review of existing literature around rubric design (for an example, see Readman & Allen, 2013). Hence, it could be concluded that the design of the *Continua model of a GTMJ*, using arrows as a visual cue to signify learning as a continuum of quality, may be the most significant aspect of this model. The impact of layout of the new model is evidenced in the following participant comments:

I find the new model to be very visual and user friendly with visible behaviours linking directly to evidence in student work. This link is possible because of the explicitness of the standards descriptors. I like the focus on describing the behaviour in positive terms, not negative. It makes the learning more visible and explains what the expectations are.

When you don’t know what a level looks like it is easier to use this new model because you have freedom to tick anywhere in the continuum.

As noted above, paramount to the advantages of the continua model is its interpretability by different audiences which is perceived to be based on the model’s two unique features: the visual appeal created by the continuum approach (described above) and, the ‘nested’ positioning of the standards descriptors. It is acknowledged that matrix style rubrics can also accommodate this notion of ‘nestedness’ and that the quality of any rubric is dependent upon

the quality of its standards descriptors. However, the findings from this study suggests that this feature of the new model serves to ‘de-clutter’ the grading tool and enable a focus on definitive standards descriptors that assist both assessor and student. The following responses are additional to previously mentioned comments about the advantages of the ‘nested’ aspect of the standards descriptors in the new model:

The new model creates something that is much more succinct and refined and I like the ‘nestedness’ aspect which is really helpful.

The students have responded positively and stated that they like the scaffolding of the ‘nested’ criteria as they know exactly what they are expected to do to achieve each standard.

Overall the results from the thematic analysis point to the *Continua model of a GTMJ* having considerable impact on the small sample of staff who agreed to be interviewed. In this way the pilot study has achieved its goal in introducing a innovation to the grading process and experimenting with its implementation to get a ‘feel’ for its efficacy, utility and interpretability.

Conclusions

Assessment serves a number of purposes but Boud and Associates (2010) claim that it must never lose the purpose of ensuring students are able to identify quality and be able to relate this to their own work. If students are not able to do this as a result of assessment practices, the educational purpose of assessment is lost. Hence, the tool used to make judgments about quality and then ‘justify’ quality using standards descriptors, must be seen as paramount in this process.

The marking rubric is not only important for student learning, it is also important in moderating students’ results in the quality assurance process used to ensure consistency of teacher judgments. Moderation brings the assessors together to reach a consensus about what evidence actually constitutes quality in an assessment task. Poorly constructed rubrics with ambiguous standards descriptors do not encourage consistency nor do they encourage smooth moderation meetings. The marking rubric could be considered as the most significant aspect of the moderation process and is crucial to accurate and consistent assessment judgments.

This paper has reported on a pilot project designed to engage a select group of academic peers in raising the quality of their teaching and learning through an assessment-

related initiative that centred on an alternative model of grading tool. The findings from our research indicate that the design of the *Continua model of a GTMJ* has significant advantages over traditional matrix-style rubrics with regard to understanding what constitutes quality, constructing the standards descriptors, simplifying the grading process, streamlining moderation procedures and communicating assessment expectations to students.

The constraints of this pilot project limit discussion of these results beyond the specific university teaching context in which they were generated. However, the results do indicate that academics are ready for an alternative to the matrix-style rubric and have the capacity to reform their assessment practices provided they have access to appropriate professional development and a forum for sharing their ideas. Enhancing course assessment strategies in this way has the potential to improve student satisfaction in university courses across a number of measures. To support this claim and consolidate the promising research findings of this pilot study, similar research using the *Continua model of a GTMJ* needs to be conducted across a range of university teaching contexts. An additional focus will be on expanding the data set to not only including more lecturers, but also to include the student voice in terms of their reaction to this model. Finally, we acknowledge that the context of this pilot project is Australia, but we believe the lessons to be learned from the introduction of an alternative grading tool are universal.

References

Boud, D. & Associates (2010). *Assessment 2020: Seven propositions for assessment reform in higher education*. Sydney: Australian Learning and Teaching Council. Canberra, Australia: Department of Education, Employment and Workplace Relations.

Department of Education, Employment and Workplace Relations (DEEWR). (2008). *Review of Australian Higher Education (The Bradbury Report)*. Available at <http://www.deewr.gov.au/highereducation/review/pages/reviewofaustralianhighereducationreport.aspx>

Department of Industry, Innovation, Science, Research and Tertiary Education (2011). *Higher Education Standards Framework*. Tertiary Education Quality and Standards Agency.

Education Queensland (2001). *New Basics - The why, what, how and when of rich tasks*. Available at:
<http://216.78.200.159/RandD/New%20Basics%20Project/NewBasicsrichtasksbklet.pdf>

Grainger, P., Purnell, K., & Zipf, R. (2008). Assessing the quality of student work: The role of “substantive conversation” between markers. *Assessment and Evaluation in Higher Education*, 33(2), 133-142.

Klenowski, V. (2007). *Evaluation of the effectiveness of the consensus-based standards validation*. Available at
<http://education.gld.gov.au/cornorate/newbasics/html/lceeval.html>

Klenowski, V., Adie, L., Gunn, Looney, Elwood, Wyatt-Smith, C., & Colbert, P. (2007). ‘Moderation as Judgement Practice: Reconciling System Level Accountability and Local Level Practice’. *Australian Association for Research in Education Conference: Research Impacts: Proving or Improving?* Fremantle, Western Australia (25-29 November). Available at <http://www.aare.edu.au/07pap/wya07015.pdf>

Matters, G. (2005). The grading master: A simpler way. *EQ Australia*, 2, 12-15. Available at <http://www.eqa.edu.au/site/thegradingmaster.html>

Organisation of Economic and Cultural Development (OECD). (2011). *Testing student and university performance globally: OECD's AHELO*. Retrieved 10.8.12 from http://www.oecd.org/document/22/0,3746,en_2649_35961291_40624662_1_1_1_1,00.html

- Race, P. (2006). *The Lecturer's Toolkit, 3rd Edn.* London: Routledge.
- Readman, K., & Allen, R. (2013). *Practical planning and assessment.* Oxford University Press. South Melbourne.
- Reddy, M. Y. (2014). Design and development of rubrics to improve assessment outcomes. *Quality Assurance in Education, 19*(1), 84-104.
- Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education, 13*, 191-209.
- Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment and Evaluation in Higher Education, 30*, 175-194.
- Sadler, D. R. (2007). Perils in the meticulous specification of goals and assessment criteria, *Assessment in Education: Principles, Policy and Practice, 14*(3), 387-392.
- Sadler, D. R. (2009a). Grade integrity and the representation of academic achievement. *Studies in Higher Education.* Vol. 34, No. 7, November 2009, 807–826.
- Sadler, D. R. (2009b). *Moderation, grading and calibration. Edited Keynote Address.* Good Practices in Assessment Symposium. Griffith University, 24 November. Available at http://www.griffith.edu.au/__data/assets/pdf_file/0017/211940/GPA-Symposium2009-Edited-Keynote-Address-FINAL.pdf
- Sadler, D. R. (2010). Fidelity as a precondition for integrity in grading academic achievement, *Assessment & Evaluation in Higher Education, 35*: 6, 727 -743.
- Sadler, D. R. (2011). *The role of academic teacher judgments in assuring achievement standards.* Oxford University Centre for Educational Assessment Invited Research Seminar, St Anne's College, 20-22 June. Available at: oucea.education.ox.ac.uk/wordpress/wp-content/.../Sadler-FINAL.pdf