Towards sustainable marking practices and improved quality of feedback in short-answer assessments

Jon Yorke
Office of Assessment, Teaching and Learning, Curtin University, j.yorke@curtin.edu.au

Will Gibson
Faculty of Health Sciences, Curtin University, w.gibson@curtin.edu.au

Heath Wilkinson
Business Systems, Curtin IT Services, Curtin University, h.wilkinson@curtin.edu.au

Student dissatisfaction with the quality and quantity of feedback received on their performance is a recurrent feature of student surveys (Scott, 2005; Williams & Kane, 2008). The study described here sets out to address this issue in a sustainable way, working within the context of the short-answer assessment format. The primary aims of this study were to reduce the time needed to mark work and to improve the quality of feedback received by students, without recourse to the kind of automated approaches that lack the personal dimension of assessor judgment. To achieve this, a prototype marking software was developed during 2009, and a preliminary trial was conducted in 2010 with 25 students responding to a quiz comprising 25 questions. Online marking using this tool was completed in a third of the time taken to mark the work conventionally some 10 weeks previously. Further investigation revealed additional (and in some cases serendipitous) advantages relating to moderation and administrative efficiency, suggesting that rapid feedback can be provided without overloading academic staff with repetitive, time-consuming marking tasks. Ultimately, this project may aid in the development of a sustainable marking approach within the short-answer context to help address the important issue of timely student feedback.

Keywords: automated; feedback; marking; short-answer; sustainable.

Introduction

Assessment is a complex process, and often a more challenging one than many in higher education appreciate (Knight, 2002; Gibbs, 2006; Yorke, 2008; Attwood, 2009). It is a key quality indicator in teaching and learning (Flowers & Kosman, 2008) and, as Ewan (2009) notes, there remains a need for a "strong oversight of assessment" in higher education. This need is repeatedly reflected in surveys that consistently identify assessment and feedback as the areas with which students are least satisfied (Scott, 2005; Williams & Kane, 2008). Conditions that support assessment for learning are well known and oft cited (Race, 2003; Gibbs & Simpson, 2004-5; Nicol, 2009). Unfortunately, the extent to which these good practices are realised is often attenuated by myriad competing pressures relating to availability of time and declining resources, among other things. The trend towards larger cohorts inevitably adds to these pressures, the risks of which are neatly summarised by Harris et al., who warn us that "Approaches to assessment that are manageable and effective for a class of forty may become virtually impossible for a class of five hundred", observing also that "many end-of-semester examinations rely heavily upon multiple-choice questions specifically in order to reduce the time and resources required for grading" (Harris et al., 2007).

Many would argue that this is an unacceptable and unsustainable situation in higher education. Concerns such as these are amplified in the seven propositions for assessment reform produced by Boud and Associates (2010), who observe that marks and grades alone provide scant information to learners, arguing that learners need "specific and detailed" responses. As Sadler (1989) points out, assessment for learning is more effective when students are given advice on how to 'close the gap' between their current level of performance and that to which they aspire. This worthwhile aim cannot be achieved without some effort, and timely feedback plays an important role here.

Good feedback comes from a range of sources, as effective and equitable appraisals of student performance are achieved through a variety of assessment tasks. It is necessary to achieve a representative balance of tasks to demonstrate a range of attributes (such as communication skills or project report writing). An incidental benefit of this spread is that it avoids giving an advantage to particular individuals by favouring any one form of assessment. Inevitably, each different type of assessment approach has its associated strengths and weaknesses, yet all share a common intent to support and quantify student learning. This study acknowledges the diverse range of assessments available, and endorses those designs that assess authentic tasks, such as the student accountant grappling with a balance sheet full of errors. However, the pressures associated with assessment (particularly where cohorts are large) are readily apparent and it is easy to see why approaches such as those based on multiple choice (MC) may come to be seen as a panacea.

Unfortunately, MC-based approaches may be more placebo than panacea. MC approaches suffer from a number of limitations, including the ever present risk that a candidate will achieve a pass by chance. Brown (2001) points out that a test needs to contain at least 50 questions with three answer options each in order to reduce the probability of a pass by chance to 1.1 per cent. Adding a fourth answer option reduces this probability to 0.01 per cent; however, it becomes increasingly difficult to write additional convincing distracter responses and, in reality, candidates will rule out improbable or obviously wrong answers, thereby considerably improving their odds of passing. Often, it is simple recognition of the correct answer (rather than recall of the correct answer: a subtly but qualitatively important distinction) that is rewarded. There is also a fundamental restriction placed on students' answers, as they are required to choose from the list of options, and other answers may also be correct. ('Multiple select' versions of this approach address this issue to some extent, although we note their relative lack of popularity with both staff and students.) It is significant that feedback provided in MC approaches is usually limited to explanations of why particular answers were correct or incorrect, and this can simply be insufficient, as it fails to 'close the gap' between current and desired performance.

For these reasons, short-answer (SA) approaches are often preferable to MC assessment. MC and SA approaches are essentially convergent (in that in each case there is one correct answer), although divergent responses (many possible answers) are to some extent catered for by the SA approach. Fundamentally, SA approaches are more logical, as they reward recall rather than a mixture of what might be termed 'recall, recognition and ruling out'. Furthermore, spelling, grammar and the appropriate use of nomenclature are all observable. The correct answer is not available to the candidate unless they are able to derive it, and the probability of a pass by chance is likely to be significantly lower for this reason. The SA format is often able to sidestep issues relating to nuances of the English language, the subtleties of which are often used in MC approaches to differentiate correct and incorrect responses with a consequent bias in favour of native English speakers. SA questions can often be asked in a straightforward fashion, and this approach may be more inclusive of those for whom English is not a first language.

Despite these apparent advantages, there are a number of reasons why SA-type assessments are avoided. Largely this reluctance is due to the time required to mark student responses and provide feedback. As class sizes increase, the time required to mark SA responses increases in a linear fashion to the point where the use of SA formats may be avoided in favour of MC alternatives, which offer quick feedback. The need for this is well documented (Maclellan, 2001; Nicol, 2009). Unfortunately, while such feedback may be quick, it is not necessarily helpful.

Although SA formats provide lecturers with the opportunity to give much richer and more detailed feedback, the issue inevitably returns to the time required to mark the work. Put simply, for larger classes the SA approach rapidly becomes unsustainable. Table 1 sums up the well-known assessment 'costs' of creation and marking for three broad approaches, including MC and SA.

Table 1: Comparison of assessment approaches and associated costs

Type	Creation cost	Marking cost
Multiple choice (MC)	High	Low
Short answer (SA)	Low to medium	Medium to high

Extended answer	Low	High

Source: Adapted from Ebel, 1972.

Other assessment approaches, such as group work, have been successfully supported through the use of new technologies, such as Raban and Litchfield's (2007) TecTRA system (used to support the high assessor loads associated with processing individual contributions to group activities) and the ReMark PDF approach to the marking of extended-answer formats (Colbran, 2009). Automated marking of short (and even long) answers have been extensively reported (see, for example, Wood et al., 2006), but these have tended to focus on sophisticated solutions involving semantic analysis or computational linguistics. We suggest that any approach that visibly removes individual assessor judgment from the equation will – in students' eyes – fall short of the mark. There is, we argue, a case for developing a technology-mediated approach to support the assessment niche of the short answer.

The short-answer feedback system (SAFS) described here places assessor judgment at the heart of the matter but also sets out to address the 'marking cost' issues associated with short-answer formats (where a short answer is defined as a student response of no more than one sentence). The aim is to reduce the marking load on the assessor and create a sustainable marking approach without detracting from the quality and quantity of feedback provided to the learner.

SAFS specifications and prototyping

Essentially, SAFS is a software tool used to replicate marking decisions across the student cohort and provide rapid and detailed feedback. It combines a marking engine with an extendable database that 'learns' (stores) student responses and the associated assessor feedback. All 'learnt' student responses are automatically processed and the remaining student responses are passed on to the assessor for review, grading and feedback (which are then used to further extend the database of student responses and assessor feedback).

A prototype of the system was developed at Curtin University during 2009, and a pilot evaluation was undertaken in 2010. The prototype is shown in Figure 1. Student responses are captured using conventional e-assessment tools or simple web forms, populated with questions drawn from a bank of questions. Expected responses are stored in the SAFS database from the outset and, initially, the correct student responses are automatically marked if they match a stored model answer. The remaining responses are presented to the marker for grading and comment. Matching 'unique' responses are clustered together and presented to the marker in order of frequency of occurrence. Using SAFS, the marker does not mark each script in turn: instead they mark question by question, working from the most common student responses to those that are rare. It can be seen from Figure 1 that the marker will never mark the same student response twice – encouraging the marker to give more detailed feedback, especially where the cluster size (number of students) giving this particular response is high. Furthermore, alternative and correct responses can be readily accepted, with especially innovative or creative work rewarded appropriately through praise in feedback or an increase in numerical score.

Grading decisions and associated feedback are stored in the database for each unique student response and automatically applied to all replicates of that particular response. These decisions and feedback items are then reassembled for each student assessment to provide individual feedback for each item in the form of a personalised response to their assessment. Additional information can be readily added to allow students to benchmark their performance against others', either at the level of the question or at the level of the assessment.

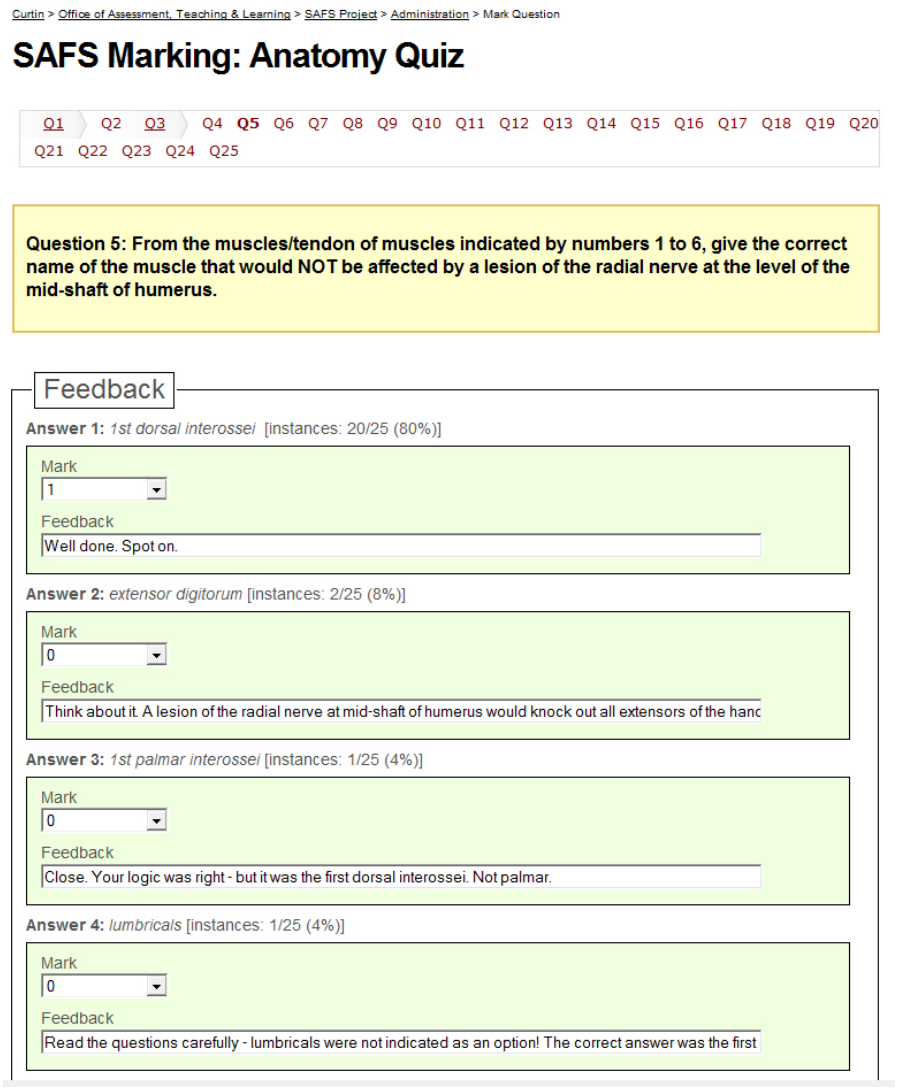

Figure 1: Prototype of the SAFS marking engine

In the pilot evaluation, a 25-question formative SA assessment was undertaken by 25 students (producing a theoretical maximum of 625 unique responses) at Curtin University. These were marked both traditionally and using SAFS. The online marking using SAFS was completed in five hours and 45 minutes, compared to 18 hours taken to mark the work conventionally some 10 weeks previously. The SAFS and conventional marking activities were separated by this period in an attempt to reduce practice effects, but these cannot be entirely discounted and further comparative work is required. Some of this seemingly large reduction in marking time was also due to the automatic collation and presentation of candidate responses grouped by question. Other incidental benefits were also observed: for example, the marker was able to swiftly review all responses to a particular question prior to commencing marking, thereby potentially improving intra-assessor reliability and identification of issues:

> … patterns could be picked in terms of common student misconceptions or interpretations … I don't think I could have picked up on these patterns so readily if the answers had not been collated. It's harder to see the big picture when dealing with one paper at a time (SAFS marker notes).

Moderation of marking in SAFS can also be achieved in a rapid and sustainable manner. A second marker is able to independently review the first marker's decisions for each unique student response. Significantly, owing to the way student responses are grouped by question and ordered by frequency of occurrence, moderation (and marking) is truly anonymous with respect to the identity of the individual students whose work is being marked. Furthermore, by sampling clusters of responses, a larger sample is effectively reviewed compared to alternative approaches that might sample just 10 per cent of student work: a practice that Bloxham (2009) argues is intrinsically "dubious".

The results of this small-scale pilot are extremely encouraging, especially given the relatively small size of this group, as it would be expected that larger classes would be required to start seeing substantial gains in efficiency. Yet even in this small-scale pilot efficiencies were readily noted:

> Collation and representation of answers meant I could attend to repeated answers very quickly, 'knocking off' large chunks of the marking load. This made the task somehow more manageable and allowed me then to go on and deal with the large number of unique answers, knowing I had already (and very quickly) dealt with a fair percentage of the marking already (SAFS marker notes).

As the work returned to students was exclusively based on conventional paper marking, it was not possible to draw conclusions between the two approaches with respect to student feedback on the quality of comments received: this is an issue we plan to address in further work. We would expect students to respond favourably to the 'anonymous marking' nature of SAFS, in that that assessors do not see responses grouped (and thus identified) by individual when marking.

Conclusion and future directions

This project addresses the longstanding problem of how meaningful feedback can be provided to larger classes in the short-answer domain. Initial evaluation work with a prototype marking software developed at Curtin University has been promising, suggesting that marking time can be significantly reduced without loss of feedback fidelity and with probable improvements in marker consistency. Significantly, each student response is always appraised by the assessor, but replication of assessor judgment for matching answers removes much of the marking overhead otherwise associated with this form of assessment.

Further research is planned to evaluate the:

- time taken for lecturers to give feedback using this system compared to traditional approaches, measured through comparative analyses
- quality and quantity of feedback, measured through self-reporting and peer review
- impact on the student experience, measured by survey and interview
- comparative efficacy of the system for classes of varying sizes, measured using above methods repeated across classes of differing size
- strengths and limitations of the project in a range of other contexts and disciplines
- issues that may facilitate or hinder uptake in other institutions.

In particular, evaluation activities will establish ways in which question sets may be managed using existing e-assessment systems or utilities such as Respondus. Particular attention will be paid to ways of sharing assessment content through the use of IMS QTI (IMS Global Learning Consortium, 2010) and how integration with Shibboleth Identity Providers can enable inter-institutional authentication and integration with commonly used learning management systems such as Blackboard and Moodle.

SAFS also has the potential to work with larger databases of questions and responses shared collaboratively across institutions. Such question sets may already be in existence and available for import into SAFS, although we acknowledge that some work may be needed to adapt or convert them from an alternative format. Nevertheless this is, we argue, an avenue worth exploring. In a higher education landscape defined by a sharpening focus on quality assurance but blighted by concerns about academic workload and declining resources, sustainable assessment practices that also provide high-quality feedback are highly attractive.

References

Attwood, R. (2009). 'HE in FE' holds its own in National Student Survey. *The Times Higher Education.* Retrieved July 30, 2010, from http://www.timeshighereducation.co.uk/story.asp?sectioncode=26&storycode=405247&c=1.

Bloxham, S. (2009). Marking and moderation in the UK: False assumptions and wasted resources. *Assessment & Evaluation in Higher Education, 34*(2), 209-220.

Boud, D., & Associates (2010). *Assessment 2020: Seven propositions for assessment reform in higher education.* Sydney: Australian Learning and Teaching Council. Retrieved July 30, 2010, from http://www.iml.uts.edu.au/assessment-futures/Assessment-2020_propositions_final.pdf .

Brown, R. W. (2001). Multi-choice versus descriptive examinations. *Proceedings of the 31st ASEE/IEEE Frontiers in Education Conference*. Reno, US, 10-13 October 2001.

Colbran, S. (2009) *The ReMarks PDF Markup Editor: Stage 1 Final Report.* Retrieved July 30, 2010, from http://www.altc.edu.au/system/files/resources/PP7-542%20ReMarksPDF%20Stage%201%20Final%20Report%20Sept09.pdf.

Ebel, R. L. (1972). *Essentials of education measurement.* Englewood Cliffs, New Jersey: Prentice Hall.

Ewan, C. (2009). *Learning and teaching in Australian universities: A thematic analysis of Cycle 1 AUQA audits.* Australian Universities Quality Agency and the Australian Learning and Teaching Council. Retrieved July 30, 2010, from http://www.auqa.edu.au/files/publications/learning_and_teaching.pdf.

Flowers, J., & Kosman, B. (2008). Teaching matters: Developing indicators of teaching quality, a comparative study. *Proceedings of AUQF2008 – Quality and Standards in Higher Education: Making a Difference.* Canberra, 9-11 July 2008.

Gibbs, G., & Simpson, C. (2004-5). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education, 1*(1), 3-31.

Gibbs, G. (2006). *Departmental leadership for quality teaching: An international comparative study of effective practice*. Retrieved July 30, 2010, from http://www.learning.ox.ac.uk/oli.php?page=72.

Harris, K-L., Krause, K., Gleeson, D., Peat, M., Taylor, C., & Garnett, R. (2007). *Enhancing assessment in the biological sciences: Ideas and resources for university educators.* Retrieved July 30, 2010, from http://www.bioassess.edu.au/key-issues/engaging-large-classes-through-assessment.

IMS Global Learning Consortium (2010). *IMS question and test interoperability specification.* Retrieved July 30, 2010, from http://www.imsglobal.org/question.

Knight, P. T. (2002). The Achilles' heel of quality: The assessment of student learning. *Quality in Higher Education, 8*(1), 107-115.

Maclellan, E. (2001), Assessment for learning: The differing perceptions of tutors and students. *Assessment and Evaluation in Higher Education, 26*, 4, pp. 307-318.

Nicol, D. (2009). Quality enhancement themes: The first year experience. Transforming assessment and feedback: enhancing integration and empowerment in the first year. Scotland: The Quality Assurance Agency for Higher Education.

Raban, R., & Litchfield, A. (2007). Supporting peer assessment of individual contributions in groupwork. *Australasian Journal of Educational Technology, 23*(1), 34-47.

Race, P. (2003). Why fix assessment? In L. Cooke & P. Smith (Eds.), Seminar: Reflections on learning and teaching in higher education. Buckinghamshire, UK: Chilterns University College.

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

Scott, G. (2005). Accessing the student voice: Using CEQuery to identify what retains students and promotes engagement in productive learning in Australian higher education. University of Western Sydney.

Williams, J., & Kane, D. (2008). *Exploring the National Student Survey: Assessment and feedback issues. Higher Education Academy*. Retrieved April 7, 2010, from http://www.heacademy.ac.uk/assets/York/documents/ourwork/research/NSS_Assessment_and_Feedback_ExecSummary_31.04.08.pdf .

Wood, M. M., Jones, C., Sargeant, J., & Reed, P. (2006). Light-weight clustering techniques for short text answers in HCC CAA. *Proceedings of the 10th International Conference on Computer Aided Assessment*, Loughborough, UK.

Yorke, M. (2008). Grading student achievement: signals and shortcomings. Abingdon, Oxon: Routledge.