Benchmarking Assessment: breaking down barriers and building institutional understanding

Simon J Cross and Denise Whitelock
The Open University, UK

Abstract

Benchmarking offers a comprehensive way of measuring current practice in an institution; whilst also gauging achievement against external sources. Although e-learning has been benchmarked with a number of universities in the UK and abroad no one to date has tackled the area of assessment; which is now becoming of more concern with the advent of e-assessment. This paper describes the construction of a set of benchmarking measures/indicators and the outcome of early pilots which combine a survey instrument and semi-structured interview methodologies. The findings suggest that building a comprehensive and robust core of benchmark measures can have great utility and value to institutions; not just in external benchmarking but also in internal reviews. It can also assist with setting baselines, exploring the student experience, providing staff with data meaningful to their role and professional development together with supporting a continuous improvement trajectory.

Benchmarking the practice and processes that support, drive and deliver assessment should be an activity all universities periodically undertake. However, the very idea of 'benchmarking' can carry connotations of a detached, strategic, and time-intensive process offering little to practitioners and their immediate managers. Our approach is focused on assessment in Higher Education institutions where we are seeking to develop a more light-touch methodology for gathering data in an area that has not been investigated before. E-Learning has been subjected to a benchmarking scrutiny (Bacsich 2005, Marshall 2006; Higher Education Academy 2009) however assessment per se has been neglected. We believe with the advent of more e-assessment and changing pedagogies and learning designs in this area merits further investigation.

Our first aim is to develop a comprehensive set of benchmark measures (what others may refer to as indicators, or criterion) about assessment processes and practice in consultation with the wider Higher Education sector. Our second aim is use these measures to lead a benchmarking exercise at the Open University and other

interested institutions. We anticipate that, in addition to enabling valuable external benchmarking, such a project will help achieve internal benchmarking across faculty and awards, baseline assessment practice and process, which will drive continuous improvement and professional development and better understanding of the student experience of assessment.

In the two months since starting this project, we have made good progress towards achieving the first of our two aims. As the following two sections describe, we have created a core set of one hundred assessment benchmark measures and begun to pilot the methodology – a staff survey instrument – that we plan to use 'in anger' together with a number of other universities. First of all let us turn to the construction of the benchmarking measures within our assessment agenda.

1. Building assessment benchmarking measures

Initial enquiries could not locate a predefined and comprehensive set of benchmark measures for assessment although there are a plethora of assessment principles, guidelines, recommendation of best practices and quality assurance indicators. We instead decide to turn to methodologies for benchmarking e-learning with the expectation that assessment measures could be found within these. The five benchmark methodologies used by projects in the HEFCE funded Benchmarking and Pathfinder Programme (2005-2008) offer a representative selection of these, which include:

- Embedding Learning Technologies Institutionally (ELTI) methodology
- e-Learning Maturity Model (eMM);
- MIT90s conceptual framework;
- Observatory for Borderless education/Association of Commonwealth Universities (OBHE/ACU;
- and the Pick&Mix approach (HEA, 2009).

For our purposes the eMM seemed particularly appropriate as a starting point. It is essentially a process benchmarking method and was developed by Stephen Marshall at the Victoria University of Wellington. It is based on the principle that the maturity of a process in an institution is an indicator of how effective and accomplished the process is. This offers a continuum from partial 'ad hoc' processes through to those that are comprehensive and integrated. These can likewise be judged on a scale from 'not adequate' to 'fully adequate'. There are around forty overarching benchmark categories which eMM called 'processes' and under each is listed a series of around twenty to thirty discrete, specific measures called 'practices'. These practices define aspects of the process and therefore, when scored can be augmented to give a score for the process (Marshall, 2006).

The eMM method, therefore, offered both 'headline' process criterion and more finer measures of practice – the latter of a much greater granularity than other benchmarks we had encountered. This additional specification and clarity promised greater utility for our assembling of a core of assessment benchmark measures. A review of the approximately one thousand practices given in the eMM identified around 150 that included the words or concepts associated with assessment or that covered practice that would include assessment. In addition, two other sources were consulted: the QAA's Code of practice for the assurance of academic quality and

standards in higher education (2006) and work on formative feedback by Nicol & Macfarlane-Dick (2006). Each measure was recorded in an Excel spreadsheet.

Our next step was to begin to group these measures in to headline process categories. A thematic analysis identified fifteen broad groups and each measure was added to one or more of these groups. During this process some similar measures were combined or removed and it was reassuring to find overlap in measures from the three sources. A final rationalisation of groupings ended with the definition of just eight headline process criterion each containing between 13 and 33 measures (of practice). Some measures were common to two or more groups.

Table 1. Headline measures for the benchmarking Assessment

Headline Process Criteria	**Number of measures (some of which are in common with another Headline process)**
A1. Course design process and phases	22
A2. Teaching and teaching activity	29
A3. Evidence base, templates and examples	20
A4. Strategy, policy and institutional expectations	25
A5. Monitoring, measurement and evaluation	33
A6. Staff Guidelines and standards	25
A7. Student guidelines, support and communications	20
A8. Staff training and support	13

The outcome of this initial work was a document with 8 headline measures and 116 specific measures of practices could be benchmarked. A draft version of these measures can be found at http://kn.open.ac.uk/document.cfm?docid=13112.

The relationship between the headlines measures is shown in Figure 1 below and reveals that the measures probe three main areas that affect Assessment practice. These are:

- Institutional Policy
- Assessment development
- Checking Good Practice which not only deals with Quality assurance Measures but also includes staff training and support

Checking good practice also includes investigating whether the institution is engaging in practices that include redesigning approaches that leverage the use of new technologies as shown by the work of the REAP project. This Scottish research has revealed that technology supported assessment can result in 'improved learning, higher student satisfaction and more efficient use of staff time' (Nicol, 2007). We have also taken note of the findings of the REAQ project (Gilbert et al., 2009) and included quality issues in our measures.

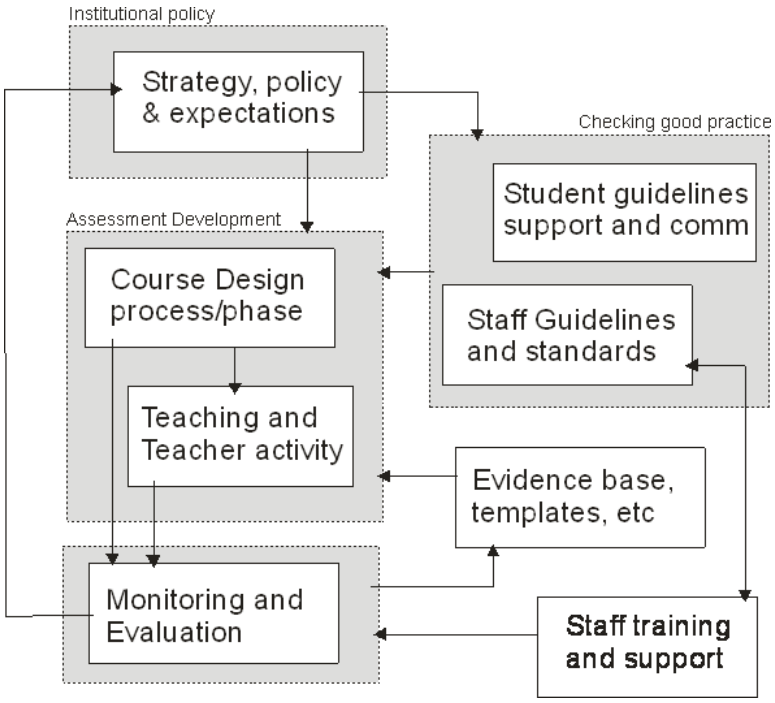

Figure1: The three main Benchmark categories and the relationship between them in terms of headline measures

2. Developing a methodology

From the outset it was clear that a light-touch methodology was required. This was not just because of the limited resources that were available for the project, but also because we did not want to overload the subjects we were planning to consult. We decided on a two stage methodology: where the first stage would consist of a questionnaire-style survey instrument and the second comprised of more in-depth interviews which would take into account the findings from the survey.

In developing a specification for the questionnaire survey we have retained the 4-point Likert scale used by the eMM benchmark (Marshall, 2006) and remained committed to including staff from all levels and key roles in the assessment process at the university together with the end consumers which are the students themselves. We have also remained sensitive to the need to gather information to support both intra-faculty and intra-institutional, as well as external benchmarking of staff and student perceptions and experience. Previous benchmark projects have shown the value of exploring areas of agreement but also where there is a divergence. Such data could provide a baseline from which improvement could be measured together with the targeting of resources for improvement activities in this domain.

3. Pilot of measures and methodology in to an institutional context

Four members of staff and one student participated in our pilot study: two senior managers, one course manager, one staff tutor (a role that supports teaching staff in based in the OU's regions) and one student.

The pilot involved a preliminary one hour interview where the purpose of the benchmarking process and the organisation of the benchmark measures were explained. The individual was then asked to take away a mock-up of the questionnaire survey which included all 116 measures (see earlier). They were asked to attempt to score each of the 116 measures of practice using the eMM 4-point scale and use a cross to show any measures they felt were not relevant to their role. Furthermore, they were asked to make a note if the meaning of the measure was not clear, if they were unsure how to score it, if there were measures that were missing from the benchmark, and if they felt they needed to quantify their score in any way. The completed self-assessment survey was returned and two follow up interviews have since taken place.

The outcome of this initial pilot phase has been to identify revisions and flag up issues we need to consider further. Some of these, grouped by theme, are outlined below.

3.1 Measuring process or effectiveness

An assumption implicit in the eMM model was that a measure of the maturity of a process can be used as a surrogate measure of its effectiveness. Several of the staff involved in the pilot said that they occasionally had difficulty deciding on an appropriate score because whilst there was a robust process in place (and therefore could be considered as being 'fully adequate') the process and practice it promoted was not producing an effective outcome. For example, whilst one staff scored the criteria 'students are provided with opportunities to describe and reflect on their own learning' (under the headline measure A2) as 'fully adequate' they noted that 'there is a blog but no-one [is] involved in it – it's left to individuals'. Elsewhere Crook et al. (2004), amongst others, have looked at this tension of process and practice and voiced concern that the proceduralisation of assessment and demands of auditing may obstruct consideration of the student experience. Taken together, this evidence suggests that a focus solely on the practice of processes may not adequately reflect the effectiveness of those processes. This has led us to consider adding a second column to benchmark score sheet associated with quality of outcome.

3.2 Scales and language used

Moving on from focus on practice, our pilot found that staff were generally comfortable with the wording of the individual benchmark measures, with one commenting they were relatively 'fair and easy enough to answer by people who know their course or programme'. This would be expected as those in course, faculty or university management encounter languages associated with benchmarking and management indicators in their roles.

The issue of interpreting what some benchmark measure were actually getting at did present some issues for teaching staff and students alike. We had attempted to remain true to the original wording in the eMM where possible and this feedback from staff shows that, as others have indicated, a degree of revision of language may be required for the UK context. In respect to a question about whether to

include students in the benchmarking, one member of staff commented that they liked the idea of asking students 'but questions would need to be direct'. This indicates that there may need to be different versions of the questionnaire, each pitched at specific audiences and asking questions relating to each measure in an accessible and relevant way.

No one interviewed suggested any new measures however, the feedback jotted in the margins on the pilot score sheets/questionnaires showed that around 10% of measures were not clear to respondents – often the definition or terminology used was unclear or a measure was considered too 'dense' (that is to say, it had two or more conditions or sub-clauses). This suggests that measures need to be kept simple, even if this means that their number increases.

Coupled with this, we found the majority of scores given to the measures of practice were either 'fully adequate' or 'not present'. There were fewer 'partially adequate' or 'mostly adequate'. This may indeed be an accurate reflection of practice, although it could also indicate the need to brief staff more explicitly about the differences between, say, 'mostly adequate' and 'fully adequate' or consider a greater range in the scale, such as the 5- or 7- point scales used in the Quality on the Line report (2000). Given the importance of setting the appropriate criteria and ensuring these link to strategy (Bacsich 2006) we plan to make a revision before our second study commences.

3.3 Staff awareness and professional development

The very fact that staff were querying the meaning and terminology of a measure demonstrated that they were thinking quite deeply about what it meant. In respect to this engagement, it emerged from the interviews that, in having to score all the measures of practice, the respondents' attention was drawn to questions they would not normally be asked to reflect upon. This had a positive impact on the respondent who acknowledged that the Benchmarking survey was prompting them to reflect and question their current practice in new ways. This finding has also been documented by Jackson (1998) in a pilot benchmarking of assessment practice in engineering departments where he found that 'respondents perceived that the benchmarking process extended their capacity to evaluate themselves critically in a non-threatening way'. This would suggest that irrespective of what data was recorded for aggregation and analysis, the very process of having to score each benchmark measure could act as a useful professional development tool. This would raise awareness, help foster shared productive dialogue and terms of reference and support the setting of baseline and continuous improvement strategies.

3.4 Dealing with variation and multiple scales of practice

The issue of scale emerged in most of the initial pilot interviews. Some staff, such as programme managers, are involved with a number of courses which may differ in their design, delivery, monitoring of assessment etc. These staff were uncertain about how to accommodate this range or variation within the score they assigned to a measure: should they give a range of scores or perhaps a score that reflected the majority of courses? This would suggest that there will be several levels, or frames-of-reference to any benchmark scoring and that these should effectively be linked together: students and tutors would score in respect to a single course; programme

and faculty managers in respect to a programme; and senior management to the university as a whole.

Some variation occurred in the answers given by the same respondent. To test this one measure was included twice in the survey: under one headline process category it was scored as 'fully adequate' and under another 'largely adequate' by one respondent. This would indicate that a questionnaire should include some repeated measures so as to evaluate the accuracy of scoring.

3.5 Presentation of the benchmarking to stakeholders and participants

Some staff had mixed feeling as to the direct, practical value to themselves of benchmarking at the external macro-level. However, presenting the exercise as a tool that could provide baseline data about their course/programme/faculty and enable them to benchmark themselves against others in the university was well received. This stresses the need to present the benchmarking in terms of value to the stakeholder/participant and how the findings could be used to improve /change practice

An additional consideration when presenting the benchmarking to staff is being aware of the historical and cultural organisational context in which the benchmarking is to be introduced. For example, one of those interviewed had assumed our project was linked to an initiative proposed a few years earlier. This highlights the danger, as well as benefit, of a mistaken association.

3.6 Comparison of responses

In addition to the individual responses to the survey outlined above, we also wanted to explore how it would help make visible similarities and differences between staff responses. This we anticipated would provide evidence about the implementation of assessment policy and its effect on relevant staff and students. Our initial pilot already indicates great promise and potential to understand this type of scenario. This is demonstrated when tutor and students responses to measures under headline A2 and A7 (which both concern student-facing aspects of assessment) are contrasted. There was agreement on 17 measures and disagreement on 11. For example: whilst the tutor rated 'fully adequate' the measure 'those involved in designing teaching of the course ensure learning objectives are linked explicitly throughout learning and assessment activities using consistent language', the student scored this 'partially adequate' and conversely, where the student 'the course provides an explicit description of the pedagogical approach being used' was 'largely adequate' the tutor only rated this 'not/partially adequate'. This hints at the potential analysis achievable with a larger dataset of responses from across and beyond an institution facilitating the answers to such questions as:

- What could explain the differences detected?
- Which perspective is most accurate?
- Where do staff agree that there is a process or practice that is not adequate?

4. Next phases of benchmark development and implementations

This project, although only two months old, has built a core set of just over one hundred measures of practice. The measures are grouped in to eight headline process categories/criteria ranging from course design to teaching; strategy and policy to monitoring and evaluation; staff and student guidelines to training and templates. The measures have been pilot tested with a small group of staff and this piloting will be expanded in the forthcoming months.

Early indications are that the measures and survey instrument could provide a comprehensive and robust collection of measures for use by Higher Education Institutions. We anticipate that, in addition to enabling valuable external benchmarking, it will help achieve internal benchmarking across faculty and awards, baseline assessment practice and process, drive continuous improvement and professional development and better understand student perspectives and experience of assessment.

Over the next three months we plan to:

- Make the benchmarking measures available to the Higher Education community and promote a conversation about them. This paper represents part of this process. We hope that one result of this consultation would be that other institutions express an interest in joining our pilot and benchmark process, thereby promoting benchmarking across the sector.
- Extend our pilots within the university consulting with staff in other roles and across other faculties.
- Continue to review relevant literature in order to test our benchmark measures, and the language used in their wording, against other research.

We then intend to refine and revise the core set of measures in response to this consultation before developing a survey instrument, or rather a set of survey instruments for appropriate audiences. At this point we would be ready to join other interested institutions in undertaking the benchmarking of university assessment.

References

Bacsich, P. (2005). *Theory of Benchmarking for e-Learning: A Top-Level Literature Review*. [Online]. Available at: from http://www.matic-media.co.uk/benchmarking/Bacsich-benchmarking-2005-04.doc [Accessed 20 June 2010].

Bacsich, P. (2006) Higher Education Academy e-Learning Benchmarking Project: Consultant Final Public Report. [Online]. Available at: http://elearning.heacademy.ac.uk/weblogs/benchmarking/wp-content/uploads/2006/09/bacsich-report-public20060901.doc [Accessed 20 June 2010].

Crook, C., Gross, H. & Dymott, R. (2004). Assessment relationships in higher education: the tension of process and practice. *British Educational Research Journal*, 32 (1), 95-114.

Gilbert, L., Gale, V., Wills, G. & Warburton, B. (2009). *JISC Report on E-Assessment Quality (REAQ) in UK Higher Education*. LSL: University of Southampton.

Higher Education Academy (2009). *E-learning benchmarking + pathfinder programme.* York: Higher Education Authority.

Institute for Higher Education Policy (2000). *Quality on the Line: Benchmarks fo Success in Internet-Based Distance Education*, Washington DC.

Jackson, N. (1998). Pilot benchmarking study of assessment practice in seven engineering departments. *Pilot studies in benchmarking assessment practice*, Gloucester: The Quality Assurance Agency for Higher Education.

Marshall, S. (2006). *E–learning Maturity Model Process Assessment Workbook*, New Zealand: Ministry of Education.

Nicol, D.J. & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199-218.

Nicol, D. (2007) JISC Report on REAP: Re-engineering Assessment Practices in Scottish Higher Education, [Online] JISC. Available at http://www.jisc.ac.uk/media/documents/programmes/elearningsfc/sfcbookletreap.pdf [Accessed 20 June 2010].

Quality Assurance Agency for Higher Education (2006) *Code of Practice for the assurance of academic quality and standards in higher education - Section 6: Assessment of Students*. Gloucester: Quality Assurance Agency for Higher Education.