IJEA

Authors Denise Whitelock and Simon Cross

Address for correspondence The Institute of Educational Technology
The Open University
Walton Hall, Milton Keynes, MK7 6AA
d.m.whitelock@open.ac.uk

Published in International Journal of e-Assessment (IJEA), 1(1).
http://journals.sfu.ca/ijea/index.php/journal/article/view/18

Assessment Benchmarking: accumulating and accelerating institutional know - how for best practice

Denise Whitelock and Simon Cross
Institute of Educational Technology, The Open University

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 data from a survey instrument of these measures with semi-structured interviews. The findings indicate that the benchmark measures this project has identified can form a solid foundation for benchmarking and that a mixed methods approach built around thisa comprehensive and robust core of benchmark measures can have 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.

Keywords

Assessment;. Benchmarking; e-Learning Maturity Model; Good Practice; Quality assurance; staff training

Introduction

Providing quality feedback around assessment has become a benchmark for all Higher education institutions since it is one of the key factors highlighted on the National student’s survey and the whole of the HE community is seeking to improve their performance on these national indicators. However in order to understand how to deliver good feedback the whole assessment process needs to be considered and one way to do this is to find a set of key indicators(benchmarks) that will throw light on both process and practice with evidence gathered from the major stakeholders in the assessment arena. Yet there remains a need for a single set of measures to support institutions and practitioners in benchmarking their assessment processes and practices. For HE and FE institutions, the challenge is how to ensure that process and practices associated with assessment are visible, sufficient, of good quality, and effective; from teaching and learning to staff skills, integrated design processes to strategy and monitoring.

Yet it is not from the institutional perspective that assessment is valued. The view that assessment is important to driving learning is common (Rowntree 1987) and acquiring assessment literacy is often top of the students' agenda. Indeed, understanding the 'rules of the game' with respect to assessment practices becomes a goal in itself or a shortcut to better marks. Whereas, the acquisition of a meta knowledge about learning, which should run in tandem with assessment literacy can be bypassed altogether. Understanding how to acquire a specific subject discourse and move from being a novice to an expert, in a given domain, can be one facet of learning which can be shaped through practice and dialogue. This means the students' learning activities/assignments should provide sufficient feedback to encourage /sustain a learning dialogue with peers tutors and even self. In this way the learning design promotes an assessment for learning pedagogy, as advocated by the Assessment for Learning group (2002) Formative e Assessment has started to become an influential tool in the assessment for learning agenda primarily because it can provide timely and effective feedback embedding practice in an interesting and efficient manner in a number of different types of electronic learning materials (Kleeman et al. this issue).

This paper reports on a study, undertaken at the Open University, which set out to construct and test a 'light-touch' approach to Assessment Benchmarking. The research questions for this study included:

1. What are the main aspects of assessment process and practice that need to be represented in a series of assessment benchmark indicators?
2. How well do existing benchmarks and indicators fit these process categories?
3. Can a Benchmarking tool which adopts a low-resource, 'light-touch' approach successfully capture adequate data?
4. What are the potential uses and issues for embedding Benchmark indicators across an institution?

Central to our interest in evaluating such an approach was a desire to adequately capture the authentic voice of key stakeholders and to contrast and probe responses from different stakeholders. This was predicated on the ambition to provide an opportunity for the institution to pause and reflect on current practice and then translate the findings into a viable action plan for improvement.

Background

The term benchmarking was originally used by surveyors to compare elevations but became a quality management tool in the US in the 1970s. The Xerox Corporation were the first to utilize a set of benchmarks when they were confronted by the low cost of Canon's ,their main competitors ,copier machines (Horvath and Herter, 1992).It then became an instrument used by many US companies such as Motorola, Ford, GTE, IBM, AT&T (Kouzmin et al 1999). The Benchmark evolved from being a set standard into one of an identification of industry's best practice (Camp 1989)

Self assessment became one of the ways benchmarking has been used by industry but it was only employed in the UK for educational purposes after the publication of HEFCE's e-learning strategy (2005). The aim here was to identify a way in which HE establishments could be supported in understanding their own e-learning achievements and aspirations and then to benchmark their progress against others in the sector.

E-Learning has been subjected to a benchmarking scrutiny (Bacsich 2005, Marshall 2006; Higher Education Academy 2009). . 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).

Within these, and other, benchmark indicators, however assessment per se has been relatively neglected. Furthermore, many often work to support the perception of 'benchmarking' as a detached, strategic, and time-intensive process offering little to practitioners and their immediate manager.

Vlãsceanu,et al's (2007) takes the view that benchmarking can be used for :

- a diagnosis which provides judgments about quality
- self-improvement through comparison with others;
- evaluation of the assessment service received by the students
- informing an institution on how to improve its practice

These four uses align with our original vision for our Assessment benchmarking tool and broaden the value in undertaking an overtly 'benchmarking' process.

Identifying the main aspects of assessment process and practice

In order to develop a comprehensive set of benchmark indicators, our first step was to identify the main categories that were essential to any scrutiny of current HE assessment practice. The three main areas operating in the HE environment that affect Assessment practice are:

- Institutional Policy
- Assessment development
- Monitoring and delivering Good Practice

Figure 1 expresses the relationship between these three areas and further unpacks these in to eight categories. From the interconnections in this diagram it is clear none operates in isolation and that each must be included if the Benchmark indicators are to encompass all the process and practices in operation.

The diagram, for example, shows that checking good practice would include 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 realized that quality issues should also be included in our measures.

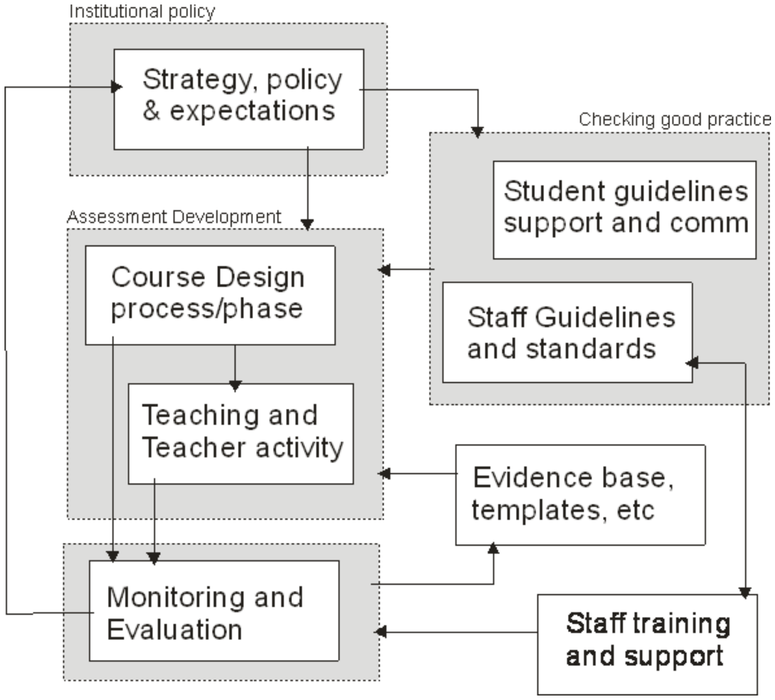
Institutional policy
Strategy, policy & expectations
Checking good practice
Student guidelines support and comm
Staff Guidelines and standards
Assessment Development
Course Design process/phase
Teaching and Teacher activity
Evidence base, templates, etc
Monitoring and Evaluation
Staff training and support

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

Identifying and adapting existing measures

The headline measures outlined in Figure 1 provide a framework in to which existing benchmark measures or indicators can be mapped. How well do existing benchmarks adequately measure the salient attributes of these categories?

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 decided to turn to the methodologies for benchmarking e-learning, mentioned above, with the expectation that assessment measures could be found within these.

For our purposes the e-Learning Maturity Model (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 finer measures of practice that could be accommodated well within our category framework. These were of a much greater granularity than other benchmarks we had encountered and 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. These were compared with other e-learning benchmarks and 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 within the headline process categories. During this process some similar measures were combined or removed and it was reassuring to find overlap in measures from the three sources. Furthermore, it was necessary to rephrase and often unpack compound e-learning measures in to constituent parts.

A final rationalisation of groupings and revision of category names ended with the definition of just seven headline process criterion each containing between 11 and 17 benchmark measures (of practice). Some measures appeared relevant for two or more categories so were situated in the category to which they most aligned. The outcome of this process was the identification of a set of 99 benchmark measures.

Table 1: Overview of our project 7 headline categories and 99 benchmark measures for benchmarking Assessment

Headline Process Criteria	**Number of measures in category measures**
A1. Teaching and teaching activity	17
A2. Student guidelines, support and communications	14
A3. Monitoring, measurement and evaluation	13
A4. Staff training and support	14
A5. Evidence base, template and examples	11
A6. Course design process and phases	16

A7. Strategy, policy, guidelines and standards	14

A full copy of the 99 benchmark measures (including revisions made as a consequence of the pilot reported below) can be found at http://kn.open.ac.uk/document.cfm?docid=xxx13112.

Developing the Benchmark questionnaire

The aim of our project was to pilot a 'light-touch' methodology for collecting data on which benchmarking could take place. This was to be a process that could take place with limited resource and which would minimize workload demands on stakeholders consulted. Consequently, emphasis was placed on deploying a questionnaire-style survey instrument after which targeted in-depth interviews could take place which would take into account the findings from the survey. The objective here was to gather, combine and compare views from many staff from different levels and key roles in the assessment process at the university together with the end consumers which are the students themselves. In this way, whilst an individual perspective may not reflect a full understanding of the process, a combination of perspectives would represent a more reliable indication.

In the questionnaire survey design we decided to retain the 4-point Likert scale structure used by the eMM benchmark (Marshall, 2006) and have two such scales for each measure. The first asked 'to what extent do you think this practice or process is going on at the institution' and the second 'what should be the minimal acceptable level of practice or process at the institution.' The first was to assess existing process and the second to better understand what was considered acceptable. We also considered but did not include here two other scales; one that asked about the importance of the process and another asking about how effective the process was.

Questionnaire design also remained sensitive to the potential range of uses identified by Vlãsceanu,et al's (2007). It was important to confirm that such a methodology be scalable and capable supporting both intra-faculty and intra-institutional, as well as external benchmarking of staff and student perceptions and experience.

Piloting and evaluating the benchmark measures

Eight people participated in our pilot study: three academics, one senior manager, one curriculum manager, one course manager, one staff tutor (a role that supports teaching staff in based in the OU's regions) and one student. Each was sent a copy of the benchmark questionnaire and were asked to respond to all items as best they could and to note any which proved problematic. Each participant was also invited to comment on their experience in completing the questionnaire: four interviews conducted with emails responses received from three others.

Some key results of the pilot are outlined below. As discussed above, the pilot represents a trial of both the wording of benchmark indicators and of the response scales. Overall, there was widespread agreement that the measures captured well the range of process and practices associated with assessment, but issues arose around how to best capture participants' responses and the specific phrasing of some measures.

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.

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.

Whilst 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.

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.

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.

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.

Comparison of responses

In addition to the individual responses to the survey, we also wanted to explore issues of consistency and uniformity in the responses given. Was there much variation in responses to the same measure? How could our questionnaire approach help make visible similarities and differences between staff? This we anticipated would provide evidence about the implementation of assessment policy and its effect on relevant staff and students and 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.

Our initial pilot already indicates great promise and potential to understand this type of scenario despite there being just eight responses. This is demonstrated when tutor and students responses to measures under headline A12 and A27 (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'. Conversely, where the student responded ‘largely adequate’ to 'the course provides an explicit description of the pedagogical approach being used' 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?

Discussion

Identifying the three main criteria which are essential components to assessment processes and practice was a useful start to modeling the relationship between these criteria and how their individual salient attributes could be represented, (see Figure 1). These three criteria were:

- Institutional Policy
- Assessment development
- Checking Good Practice

The eMM method offered a starting point for finer measures of practice which we could adapt to our model of assessment practice. In particular it offered a much greater granularity than other benchmarks we had encountered. This additional specification and clarity proved useful while assembling a core set of assessment benchmark measures.

Although the interviews were primarily about the tool itself, several participants reflected on how and why they had given the responses they did. This would indicate that a mixed methodology of combining the survey instrument which included the full set of benchmarks with semi structured interviews would prove to be a good way of prompting reflection for change. Alternatively, the individual interview can be replaced by a meeting of stakeholders discussing together the findings from the survey instrument and making sense of them with respect to any individual differences.

One of the important findings from this study was the increased understanding of the potential uses of our benchmarking tool raises. In addition to institutional benefits, the direct practical value of internal benchmarking became immediately apparent to the participants. The Benchmarking survey prompted staff to reflect and question their current practice in new ways and they can come to see the value of how the many strands of the HE's policies and strategies for Assessment meld together. These findings suggest the benefits of internal benchmarking which uses the measures as indicators or prompts for reflection and continuous improvement can encourage more stakeholder buy in to a change process which can build confidence in moving forward with an assessment for learning agenda.

References

ARG (2002). Assessment for Learning: 10 principles. Retrieved May 20, 2011, from assessment-reform-group.org

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].

Camp, R. (1989), Benchmarking: The Search for Best Practices that Lead to Superior Performance, ASQC Quality Press, Milwaukee, WI.

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.

Horvath and Herter, (1992), ``Benchmarking: Comparison with the best of the best'', Controlling, Vol. 4 No. 1, January-February, pp. 4-11.

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

Institute for Higher Education Policy (2000). *Quality on the Line: Benchmarks of 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.

Kouzamin,A. LoEffler,E., Klages,H. & Korac-Kakabadse,N. (1999) Benchmarking and performance measurement in public sectors: Towards learning for agency Effectiveness The International Journal of Public Sector Management, Vol. 12 No. 2, pp. 121-144.

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.

Rowntree.D.(1987) *Assessing students: how shall we know them*? London: Kogan Page

Vlãsceanu, L., Grünberg, L., and Pârlea, D., 2007, Quality Assurance and Accreditation: A Glossary of Basic Terms and Definitions (Bucharest, UNESCO-CEPES) Revised and updated edition. ISBN 92-9069-186-7. http://www.cepes.ro/publications/pdf/Glossary_2nd.pdf, accessed 29 January 2011.