The datafication of higher education: evaluating tests that measure learning outcomes

Hamish Coates

Follow this and additional works at: http://research.acer.edu.au/joining_the_dots

Recommended Citation
Coates, Hamish, "The datafication of higher education: evaluating tests that measure learning outcomes" (2011).
http://research.acer.edu.au/joining_the_dots/24

This Journal Article is brought to you by the Higher Education at ACEReSearch. It has been accepted for inclusion in Joining the Dots (JTD) by an authorized administrator of ACEReSearch. For more information, please contact repository@acer.edu.au.
Introduction

‘Fireside chats’ remain part of higher education, but the industry is subject to increasing quantification, assessment and review. Australia is certainly at the forefront of this change, with the Bradley Review ushering in a new era of accountability, transparency and productivity initiatives. After much discussion, these are starting to play out in 2011, underscoring the significance of building capacity and resources within universities for critically consuming potential assessment tools and approaches.

Bearing in mind this apparently inexorable trend, this briefing explores one facet of what might be referred to as the datafication of higher education. In particular, the briefing reviews the kinds of tests institutions can use to measure undergraduate learning outcomes. The briefing is not a product review, but after a few contextualising remarks presents a broader analysis of the kinds of questions institutions must ask of assessments that could be adopted – or potentially imposed – to measure, and hence improve, higher education.

While Joining the Dots presents an evidence-based picture of key facets of higher education, no data are presented in this briefing as very little data exist. Despite and possibly because of its centrality to higher education, generating objective measures of learning remains an elusive task. Necessarily, the briefing touches on methodological ideas, but is practical in intent.

There are many and varied outcomes resulting from a university education – employment, graduate study, intellectual growth, social development, to name just a few. As these outcomes take different forms they are measured with different tools such as tests, surveys, interviews and administrative data. Given the telic nature of tests and the centrality of learning outcomes, this briefing concentrates on tests of learning outcomes.
In summary, the briefing provides information about:

- tests that exist to measure higher education learning outcomes;
- questions that critical consumers should ask; and
- qualities of productive assessments in higher education.

**Emerging policy contexts**

The Bradley Review of Australian Higher Education portended major system-wide changes to key policy settings, but the changes of most relevance to this briefing are only starting to result in development timelines scheduled across the next few years.

Recently the Australian Government Department of Education, Employment and Workplace Relations (DEEWR) announced plans to commit around $30m to develop a new data infrastructure to support more evidence-based planning and leadership of higher education. Attention is focused on building a new ‘University Experience Survey’, validating the United States Collegiate Learning Assessment, and enhancing Australia’s ‘Graduate Destination Survey’. (See www.deewr.gov.au/HigherEducation/Policy/Pages/AdvancingQuality.aspx).

Partly to support new regulatory and funding changes, data from these assessments will play into accountability mechanisms (like ‘mission-based compacts’ currently being negotiated), transparency initiatives (such as the proposed ‘My University’ website), continuous improvement (‘performance-based’ funding), and projects to support the promotion of excellence in learning and teaching.

Seen from above, the spirit of such developments is to better link students’ learning with system-level leadership. For aspirant learners, the developments are designed to help them self-locate within an expanding and more complex system. Taken together, they confirm the intensifying datafication of higher education.
What kinds of tests exist?

Several kinds of tests exist to measure higher education learning outcomes, so it is helpful to develop a taxonomy that defines the scope and scale of assessment tasks.

Local achievement assessments

Undoubtedly the most important data on learning outcomes come from assessments – examinations, assignments, laboratories, class presentations – that teachers give students in class. Such assessments have a local remit, however, limiting their relevance to institution or system-level quality indicators. Yet procedures can be developed to bolster the standard and generalisability of local assessments and ensure both that aggregated student feedback can be captured in broader quality deliberations, and that the quality of the tasks themselves is improved. An example is the formation of a well-managed item library from which academics borrow and contribute assessment materials. A prime example is the Medical Schools Council Assessment Alliance (MSC-AA) (www.medschools.ac.uk/MSC-AA/Pages/default.aspx); similar initiatives are underway in Australia.

Generalisable achievement assessments

Many university subjects make use of various forms of student assessment that are more generalisable in nature. Such assessments may be externally developed in response to mandates from industry or professional associations, or generalisability may stem from various forms of moderation injected into the assessment process. Often exams are adapted from teachers’ manuals supplied with standard textbooks. Clearly there are enormous benefits for teachers and students in drawing on standard assessment materials, to the extent that this does not conflate diversity, innovation and standards.

Generic capability tests

At a higher level of abstraction, the last ten years have seen the production of tests designed to assess learners’ generic skills. These skills – covering communication, reasoning, problem solving, and interpersonal understanding – are usually a complement to formative teaching activities and provide aggregate data primarily for accountability and improvement. As the name suggests, the focus on ‘generic skills’ is pitched above (or beyond) the particularity associated with disciplinary contexts. Examples include Australia’s Critical Reasoning Tests (see: www.acer.edu.au/tests/crt) and the US-based Collegiate Learning Assessment (see: www.cae.org).

Tests of discipline capability

Similar capability assessments can also be defined to test discipline-specific skills in a generalisable way that ‘hovers above’ detailed curriculum content. Rather than focus on exhaustive measurement of curriculum such tests focus on topics considered core to the discipline at the final-year level. The results from this kind of standard assessment complement data produced by local examinations, facilitating various forms of reporting and improvement. Examples include the OECD Assessment of Higher Education Learning Outcomes (AHELO) discipline assessments (see: www.oecd.org/edu/ahelo) and the ETS Major Field Tests (see: www.ets.org/mft).

Licensing examinations

Licensing examinations offer the most stringent form of generalisable assessment, providing comprehensive measurement of specific enumerated competencies. Tightly aligned with industry and professional accreditation, licensing examinations have complex and often unexpected links with curriculum, teaching and assessment. They can enable generalisability to national or international standards, and tend to be created by expert test developers who subject each item to rigorous psychometric evaluation and field trials to optimise validity and reliability. Licensing examinations exist in Australia in several industries, for example architecture (see: www.arbv.vic.gov.au/Pages/registration.aspx) and basic public medicine (for foreign trained doctors – see: www.amc.org.au/index.php/img).
Questions critical consumers should ask

Numbers carry an allure of certainty and can be used, at the extreme, to substantiate highly dubious propositions. Hence it is imperative to set criteria that help institutions undertake informed analysis of data used to reflect the quality of learning and teaching. This is a new field, particularly in Australia, and it is necessary to sketch principles that set ground rules for development. Each proposition explored below is relatively simple in itself. Together, they provide a framework for ensuring that data help to lead rather than mislead the higher education user.

Is the data collection well positioned and does it add value to existing sources?

While data from an assessment of learning outcomes should cross-validate with alternative evidence on outcomes, to minimise redundancy they should also add value and enrich the stock of existing information. Such triangulation prevents convergence around limited indicators, which hinders diversification and innovation. Collecting data is an expensive and complex task, so it is essential that assessments are well positioned and are worth the cost, time and effort.

Is there clarity around what is being measured?

Tests can be developed to measure all sorts of things, but before building the assessment it is important to secure buy-in to a good conceptual understanding of precisely what is being measured. Specifically, it is important to probe whether the test is underpinned by a documented assessment framework. Such frameworks specify the contexts and rationales for measurement, critically chart the content being assessed, and detail how the test converts the framework into a usable test.

It is of course imperative that any assessment is able to measure the knowledge and skills that really count for future academic and professional practice. Reports should inspire and promote rather than constrain growth. Any outcomes assessment should not reduce focus to knowledge and skills that are more easily tested. Finding ways to measure broader professional attributes is vital.

Is the data collection instrument technically sound?

Questions also need to be asked about the technical properties of any outcomes assessment. Are the right areas of the curriculum – however ‘curriculum’ may be defined – targeted for assessment? Do the questions provide a valid and reliable measurement of target areas? Do results provide valid estimates of individual and hence group performance? Importantly, it is vital to ensure that materials and related protocols are designed to promote sound rather than simple forms of reporting and interpretation. Much has been learned from the global rankings movement over the last decade, and new more effective approaches are being developed – particularly those that adopt more multidimensional views of educational quality.

Is the assessment instrument culturally appropriate?

In all assessments, but particularly where a test has been developed in another cultural context, it is essential to confirm relevance and appropriateness to administration and reporting contexts. Measurement science has progressed significantly in the last century, and particularly in the last few decades robust methods have been developed to guide test adaptation work. Such adaptation needs to account for several forms of variability, not least that which arises from linguistic, cultural, conceptual and psychometric difference. Shifting tests across sociolinguistic communities is a complex and expensive process, and must be executed to the highest technical standards to ensure validity of processes and outcomes.

Are data collected in a reliable manner?

Obviously, it is essential that the collection of performance data be conducted in reliable, efficient and transparent ways. Rigor and fairness is paramount, but testing is an expensive process and trade-offs are inevitably made around the management of testing contexts, security, and controlling candidate preparation. Ultimately, particularly with devolved tests that are of low overall significance to students (“low-stakes”) and administered in several contexts, it is essential to ask if the consistency of administration supports any intended generalisability. Online deployment introduces further considerations about
identity confirmation, system reliability, and the equivalence of testing platforms.

In low-stakes testing in post-compulsory environments it is important to be clear about who sits a test, and administrators need to ensure that a test-taker is not an impostor. Ultimately, there is a need to implement controls that ensure a sufficient and representative number of students, from a defined cohort, provide valid responses to a minimum number of items. Statistical controls have been implemented, but ultimately randomised selection is required to ensure the quality and relevance of processes and outcomes.

For the purposes of isolating educational effectiveness, such sampling needs to control for context and demography, taking into account factors such as institution selectivity, student background, field of education, and level of program design.

**Are procedures for scoring data transparent and verifiable?**

To ensure parity across respondents and administration contexts it is essential that the processes used to convert student responses into data are scalable and robust.

Multiple-choice and short-response questions can be used to measure highly complex and abstract knowledge and skills. Scoring such questions can be managed in efficient, clear and rigorous ways.

Scoring extended student responses to highly conceptual tasks, particularly when pitched close to graduate level, is expensive, involving numerous practical and technical complexities. Large numbers of scorers can be employed, either at or independent of participating institutions. Alternatively, sophisticated computer-based scoring applications can be deployed, especially for low-stakes applications. With such operational complexity, careful consideration needs to be given to quality assuring scoring processes themselves, a requirement that carries specific implications for test design itself.

It is essential that standard psychometric procedures are used to produce test scores. The use of such procedures is particularly essential if comparisons are to be made across contexts and sophisticated equation is required. While useful for in-class assessments, simply ‘adding up’ correct responses is an invalid approach and culminates in biased estimates.

**Are data being reported in valid and relevant ways?**

It is imperative that assessment results be reported in contextually relevant ways; otherwise they create a danger of provoking inappropriate interpretations and costly mistakes. As noted, a certain amount of contextualisation should be ‘baked into’ the assessment, and instruments must be linguistically and culturally appropriate. But even data that are well formed in this way can be reported in better or worse ways. Without care and detailed understanding, aggregating data to very high levels (such as to the institution) carries the risk of masking significant internal variation and difference. Conversely, disaggregating reports to small subgroups can produce statistical estimates that lack power and generalisability.

Even though Australian universities have become more centralised in recent years in terms of management, institution-level aggregations are still generally insufficient to record and prompt change. This is particularly the case with reports of context-free generic skills, which, without careful educational contextualisation, can fail to capture the interest and imagination of deans, heads, managers, teachers, and support staff. Aggregation to the field of education (faculty) level is required, perhaps even to a lower departmental level. Several empirical analyses underscore the significant technical and practical errors that can emerge by collapsing discipline assessments together in any way other than a simplistic ranking.

**Can results be linked to helpful benchmarks?**

For a small system such as Australia (with just 0.2% of the world’s universities), given the highly international nature of university education, and given the growth in international forms of assessment, it is desirable for any assessment to be linked to, or at least comparable with, international initiatives. One or more multilateral arrangements can be established with key benchmark markets (United Kingdom, Canada, United States, China, India, etc.), or national/
institutional approaches can lean on international initiatives (for instance: OECD AHELO – see www.oecd.org/edu/ahelo – or U-Multirank, see: www.u-multirank.eu).

At the same time, education is a very local and personal business – students learn in physical or virtual classes, institutions liaise with those within Australia, and employers recruit local graduates. Hence there is value, at a minimum, in ensuring that test outcomes can be generalised across institutional or disciplinary groups.

The need for benchmarking sets a baseline need to share items, overlap exams, benchmark practices, and implement cross-national collaborations. Synergies exist, but the essential requirement for validity means that extensive comparative technical work is required (global benchmarks in this area have been set by ACER through OECD PISA: www.pisa.oecd.org).

Can change in practice improve future assessment results?

As the previous point emphasised, the psychometric sensitivity of any assessment instrument is critical to the utility of the measurement process. Such sensitivity underpins the capacity of the assessment to detect and ultimately inspire improvements in practice.

It is relevant to ask, therefore, if there is a sufficiently clear link between educational activity and assessment results. Put simply, when teachers and learners do better, do assessment results go up? When they do worse, do they go down? Aside from intrinsic motives (which do not require external assessment), motives to improve will surely be dampened if progress is not reflected in improved assessment results. Ultimately, lack of a good connection could prompt learned helplessness and apathy.

Does the assessment foster innovation and diversity?

With a few exceptions, such as training for licensing examinations, outcomes assessments should not encourage test-centric curriculum, teaching or learning. It would be a retrograde step for educational excellence if an outcomes assessment constrained rather than promoted innovation in educational practice. While knowledge and skills in non-core areas may be of little use in some situations they are invaluable in others, and any loss of flexibility to design targeted curricula has the potential to reduce the diversity of graduate skills.

Indeed, the production of objective data on student learning outcomes must be seen as essentially useful for educational leadership and management rather than as a form of unnecessary regulation. Shifting the primary focus of quality assurance away from input and process factors must lead to creative designs and applications.
Qualities of productive outcomes assessment

Testing of learning outcomes is still relatively uncommon in Australia, as in many systems around the world, and as such can prompt understandable fears and concerns. Yet measurement is a precondition for improvement. Hence assessing higher education learning outcomes is essential in so far as universities have an intrinsic interest in improving learning. There is a caveat, however, that higher education outcomes assessment is only worth doing if it can be done well.

So to help Australia build capacity in this area this briefing has advanced a series of questions that critical consumers should ask of proposed outcomes assessments. In summary:

1. Is the data collection well positioned and does it add value to existing sources?
2. Is there clarity around what is being measured?
3. Is the data collection instrument technically sound?
4. Is the assessment instrument culturally appropriate?
5. Are data collected in a reliable manner?
6. Are procedures for scoring data transparent and verifiable?
7. Are data being reported in valid and relevant ways?
8. Can results be linked to helpful benchmarks?
9. Will assessment results prompt useful practice change?
10. Can change in practice improve future assessment results?
11. Does the assessment foster innovation and diversity?

This briefing suggests that the next few years will see increasing research, discussion and activity around developing metrics for measuring the outcomes of university education. Done well, the assessment of learning outcomes has the potential to work with other initiatives to support the next wave of educational improvement in universities. Done badly, assessment efforts risk causing organisational and even systemic harm, or, more likely, simply wasting time and money. The practical guide sketched in this briefing is offered to help frame the development of policy and practice – to help systems, institutions and students yield improved returns from their higher learning investments.