

## **The Unbearable Lightness of Being: Universities as Performers**

by Janice Gross Stein  
Belzberg Professor of Conflict Management  
Director, Munk Centre for International Studies

### **I. Introduction**

It is no accident that as soon as we begin to talk about public reinvestment in higher education, we move quickly to a parallel discussion of accountability and performance measures. The political bargain is clear: in return for the investment of public funds, governments acting on behalf of the public demand “results” from their partner institutions. The conversation has moved beyond “transparency” about the **way** public money is spent to a discussion of **what** public money is buying. We are now in the Auditor-General’s world of “value for money.” Is the public getting “value” for the “money” it is spending on higher education?

In this paper, I make the following assumptions and arguments:

- Neither governments nor auditors can determine whether universities are providing “value for money.” It is a serious mistake to try, and universities will come to grief as have other public institutions if we try seriously. We should resist this particular conversation strenuously.
- Public evaluation of performance is here to stay. It is embedded in the political vocabulary.
- Evaluation of our performance is important to the academic community. If it is done well, it can provide stimulate self-reflection, learning, and improvement.
- Which measures we choose matters. We need to understand the limits of performance measures as measures and their power as incentives to drive strategic behaviour.

This line of reasoning leads to three broad guidelines to the construction of performance measures for higher education. We need to choose the “lightest” possible basket of measures which:

- minimizes the incentives to reduce the quality of educational programs.
- increases the opportunities for universities to gather valuable information about their performance at reasonable cost and to learn.
- rewards experimentation and innovation.

### **II. Procedural and Substantive Accountability**

Performance measures are the tangible manifestation of the far more complex concept of accountability. Indeed we cannot understand performance measures without embedding them in the larger context of accountability. At the

core of accountability is the demand by one who is entitled to do so for an “account” from another. Accountability presumes a relationship of entitlement, information, and the power to sanction those who provide the information. The essence of accountability is that the initiative is held by the questioners who have the power to punish if they are dissatisfied with the answers.

We can think about accountability in public institutions in two ways. There is accountability about rules and procedures – the way I do things – and accountability about substance – what I do. Procedural accountability is fairly easy. Over time we have agreed more or less on a set of procedures in public life, on fair rules, which constitute reasonable due process. When leaders break these rules, when they violate conflict-of-interest guidelines, when they give contracts to their friends outside of normal channels, when they fail to follow appropriate procedures, when they profit privately from public money, they can be held accountable and sanctioned.

It is entirely legitimate that governments demand that universities be accountable for their procedures. It is also legitimate for the auditor to audit the procedures that universities use to disburse public funds. Undoubtedly, the auditor will find some procedural violations and the universities will quickly fix these violations. Procedural accountability is a “no-brainer.”

Substantive accountability is far more challenging. The issue is not “How have I done it?” but “What have I accomplished?” What “results” have I delivered? It is here that we begin the slide down the treacherous slope of “value for money.” The pitfalls along the way are many. The fundamental trap is, I think, obvious: we are asked to quantify and monetize the value of a university education that students receive, of the research that we do, and of the public education that we provide. In anything but a trivial sense, this is not only a technically impossible series of tasks, it is seriously wrong-headed.

How does substantive accountability work now, at least in theory? In public institutions, we report, or give an account of what we have done, to a superior. In hierarchical organizations, with clearly drawn reporting lines neatly represented in an organizational chart, and clear command and control structures, accountability seems fairly obvious. These neat diagrams, however, often hide very large contradictions that usually escape our notice. At the core of accountability are five questions.

The first, often surprisingly difficult question is – who is accountable and to whom? In my university, the lines of accountability seem to be quite clear. I report annually to the chair of my department, who in turn reports to the dean, who reports to the chief academic officer, who reports to the president who reports to the Board of Governors. Accountability seems to flow neatly upward on the organizational chart. If we looked only at the formal procedures, however, we would miss most of what is happening.

The chart tells me that I am accountable to my chair, but I feel that I am – and indeed I am – also accountable to the students that I teach and work with every year. Two problems immediately become obvious. First, I am simultaneously accountable to at least two groups – my students and the university administration through my chair – that at times may have compatible expectations but, at times, may differ in what they expect. And second, although my students evaluate my teaching and forward the information annually to the chair of my department, they have at best only very limited capacity to punish me for what they may consider my poor performance. When a student once told me that he deserved a medal for staying awake through my course, I could only commend him for his perseverance.

It gets worse. I also feel accountable to my colleagues for the work that I do in my department, and to a wide network of scholars in other universities and research institutes who work in the same field for the quality of the research that I do. Finally, I work in a university which receives significant public funding, so I feel that I am accountable to the public for what I do. I have multiple accountabilities – to institutions, to networks, to peers, and to my students – with very different expectations and standards. These do not always sit comfortably together. When they do not, to whom do I give priority? The language of accountability, the rendering of accounts, will not help me very much here. Nor does a discussion about “value for money” that reduces a complex problem to a single dimension of measurement.

I have to look outside the formal structure for help in solving that problem of competing obligations. The danger here is obvious: to resolve the problem of multiple and at times competing accountabilities, I will look to measures of performance to provide guidance about what really matters. In short, in a context of multiple accountabilities, performance measures act as silent incentives for some kinds of behaviour rather than others.

The next two questions – for what am I accountable and by what standards? – are even more difficult. The “what” of accountability focuses on results. Am I succeeding in what I hope to accomplish? Am I meeting the standards? Another way of putting these questions is to ask what “numbers” do I provide to “measure” my “success”? Do I report on the number of published articles, or on their quality? Who judges quality and how? Do I report on whether I encouraged someone to think more deeply about some piece of conventional wisdom? If I spend two years writing a book, and publish nothing else while I am doing so, am I “unproductive?” Would one “good” book be the equivalent of two, three, five shorter publications? How do I know whether what I write “matters”? Peer review by my colleagues helps to answer these questions, but not completely. The opinions of my peers often range quite widely and review processes push to compromise, often leaching out the sharp edges of arguments. Often, after an elaborate process of peer review and revision, no one is satisfied.

The problem becomes even more complicated when we turn to my students. They evaluate my teaching every year, but are their opinions decisive? Is it my “job” to please my students, or to push them, to make them so uncomfortable that they begin to challenge, to create uncertainty so that they begin to wrestle with difficult questions? Do I spend more time teaching my students what we do know, or what we don’t know? And how do I report on whether my students are more uncertain, more open, more critical, more independent-minded, and more troubled and troublesome at the end of the year than they were at the beginning? Is the opinion of every student of equal weight? What I am accountable for is often a subject of deep controversy, a controversy that engages important values, different visions, and quite a broad spectrum of possible “results.”

Finally, how am I sanctioned for poor performance? My students can do so only very indirectly, by reporting my failures to my chair. My colleagues can comment on the poor quality of my research and my failure to advance knowledge. My performance is reviewed annually and I can be deprived of promotion or of a modest increase in salary as a signal of displeasure. But the university, like other institutions in public life, has quite deliberately built in protection against summary dismissal, to protect members of faculty who have passed the tests of tenure against those who would sanction them for what they say. Universities did so to provide safe space for dissent and criticism, for unpopular challenges to conventional wisdom. They long ago made a choice to give highest priority to the free and safe exchange of ideas, even at the sacrifice of some accountability. That choice invokes larger public values and a commitment to a public good.

This kind of choice is not unique to universities. Constitutional systems often create institutions that are deliberately not held accountable.<sup>1</sup> The Supreme Courts in the United States and in Canada, when they interpret the constitution and rule legislation unconstitutional, are not acting as agents of the public. The public cannot hold judges accountable in the usual ways by measuring their performance. How, in any event, would they conceivably do so? These two Courts are acting as trustees of the public good, free to make decisions, which may be widely unpopular, at least for a time.

Much of the discussion of accountability uses the convenient fiction of the individual in order to grapple with these issues. It is far easier to think about procedural and substantive accountability when we think about individuals. Who gave the order? Who broke the rule? Who got poor results? In my office at the university, when I ask – how could this have happened? – I intuitively think someone who must have been the weak link in a high-performing chain. But most decisions in public institutions and their results are the result of chains of collective action. When we look at substance – whether it is the improvement of security or the creation of new knowledge – it is collaborative and collective

action, flowing through complex causal chains, that produces or fails to produce results. Individual accountability, conclude two seasoned observers of public institutions, is nothing more than a useful myth to affirm control and to motivate people to do their best.<sup>2</sup>

### **III. The Costs of Accountability and Performance Measures**

The language of accountability, so pervasive now in our societies, does not take us far enough. It does not because of the difficulties in holding collectivities rather than individuals to account. It does not because so much of what matters to us is difficult to measure, to translate into the numbers that a culture of accountants, comptrollers, regulators, and auditors require. I had an astonishing conversation recently with a public auditor who was preparing to audit the Department of Foreign Affairs in Canada. He wanted to know whether the fact that the government had appeared to change its policy about the war in Iraq bespoke a lack of accountability. Were we, he asked me, getting “value for money” from the Department of Foreign Affairs? He could not understand why I was utterly confounded by the question. This is the kind of risk we run when we monetize our public culture.

Accountability favors measurable and comparable accounts – numbers that we can add and subtract. We only see what we measure and miss what we don't, and much of what is important in public institutions cannot be measured and compared. A focus on accounting, on the ledger, channels our public conversation into the concrete, the tangible, and leaves little room for the intangible, for what we cannot measure. Taken to extremes, a culture of accountability transforms the conversation about the public good into a discussion of the public's business. It impoverishes our conversation, not only about our universities but about all our public institutions.

That is not the only cost of accountability. Accountability has been most successful when it focuses on procedure, on the rules, and those who break them. The scandals that have riveted public attention in the last few years have all focused on those who break the law – in corporations, in government, in the civil service, in the voluntary sector. With each wave of scandals, governments and regulators enact a new set of rules, designed to close the loopholes that allowed the violations to happen. Some of these rules are wise and necessary – the restrictions on conflict of interest, peer review of candidates for senior appointments, legislation to protect whistle blowers. Many of the rules, however, only add to the cumbersome burden of regulation, and stifle initiative and creativity.

The same kind of problem exists in universities. Universities in Britain spent vast amounts of time, energy, and money collecting the data that were needed for agreed-upon performance measures. Measuring and reporting consumed an ever-increasing amount of resources. Universities raided each

other's faculty before every review to increase their own research profile with "productive" faculty members. An examination of the report by the British Auditor-General and Comptroller as early as 1994 tells the tale of what was to unfold. "On the basis of financial criteria alone," the Auditor-General wrote, "an institution can be judged as healthy but may be engaged in activities which are not viable and which can eventually impact on overall health, for example, *courses which do not attract sufficient students, research work which fails to attract financial support, and an infrastructure which is inefficient or under invested.*"<sup>3</sup> The language, I think, speaks for itself. To compensate for the monetization of measures, the Funding Council reviews the quality of higher education, but virtually all the subjects reviewed were rated as satisfactory or better. Despite the obligatory rhetoric about quality, the yardstick for measurement was clear.

One unexpected result was the creation of a market for highly visible faculty within the British university system and a sharp increase in salaries for the select few. The other – expected – consequence was that university leaders became even less willing to take risks, to innovate, to create, and to experiment. They focused their attention on improving their relative standing on agreed-upon measures and diverted resources from other sectors of the university so they would "rank" higher. Their behaviour is not unlike our own universities when it comes to the annual rankings in MacLean's magazine. Teachers almost always teach the test when the test results of their students are made public. In other words, we "game" the system and performance measures become the game.

The encouragement of innovation and creativity in research is one of the most important – and neglected – challenges in public universities. Leaders of the Howard Hughes Medical Institute recently built a new laboratory, the Janelia Farm Research Campus, explicitly designed to enable scientific creativity. Its operations are modeled on two research centres with outstanding records of scientific discovery – Bell Labs and the Laboratory of Molecular Biology funded by the Medical Research Council of Britain. The formula at the new lab, explained Dr. Gerald M. Rubin, vice-president of the Hughes Institute, is designed to free the scientists to do their work without constant accounting for results. Scientists work in small groups, led by people who do hands-on research. They receive ample financial and technical support so that they are freed from the constant distraction of grant-writing and grant-management. The only directive is that people work on an interesting problem, however long it might take. This model is completely *inconsistent* with any set of measures used to measure performance in universities on an annual basis. Why would Hughes create and support this kind of laboratory?

Funding agencies, leaders at Hughes explained, demand safe research and predictable results. A demanding system of peer review screens out high levels of risk and "mainstreams" most research projects. The Hughes Institute encourages young scientists to commit to long-term research projects, to projects that may be high-risk, or to projects that are unlikely to be supported by granting

agencies. The vision at Janelia Form – long-term goals, research that makes a critical difference, no pressure to generate practical results, no annual performance review, no grant writing, no administration, and genuine multidisciplinary work – is very different from established university practice. More to the point, it self-consciously acknowledges that some of its bets on young scholars will fail, but accepts these carefully thought through failures as the cost of innovation. It frees scientists to do their work, recognizing the serious costs of constant reviewing and measuring to innovation, experimentation, and risk-taking.<sup>4</sup>

This kind of model is not a substitute for established structures of research and teaching in public universities. Indeed, Gerald Rubin explicitly insists on the complementarity between the two kinds of structures. Nevertheless, universities need to make much greater space for long-term experimentation, and to think seriously about performance measures which encourage and reward risk-taking in research and teaching. This may mean moving beyond established standards of peer review and citations in “high-impact” journals which often serve as gatekeepers of conventional wisdom.

This overview of performance measures focuses explicitly on their costs and identifies several that merit serious consideration. Perverse strategic behavior and the expenditure of large amounts of scarce resources – time, energy and money – on data collection and manipulation are two of the most obvious. Performance measures tend to stifle innovation and risk-taking as they channel activity into improving performance on agreed-upon standards.

Performance measures also tend to focus on the tangible and the quantifiable, in part because quality is so difficult to measure in so many areas. The University of Toronto, for example, publishes an annual report on its performance.<sup>5</sup> It uses measures such as student retention and degree completion, class size, research council funding, research yields, and research revenue, all quantitative measures of performance. Only faculty honours and citation counts try to get at the quality of research, and citation counts raise some special issues, principally in the social sciences and humanities, but even in the sciences. This reliance on quantitative measures drives the university system in perverse ways. “Once a phenomenon has been converted into a quantitative measure,” argues Deborah Stone, “it can be added, multiplied, divided, or subtracted, even though these operations have no meaning in reality. Numbers provide the comforting illusion that incommensurables can be weighed against each other, because arithmetic always ‘works.’ .... Numbers force a common denominator where there is none.”<sup>6</sup> Stone is right.

Any analysis of performance measures needs to concentrate not only on what is measured and how it is measured, but also on the incentives these measures create within the university system and on the opportunities they provide for institutions to learn about quality.

#### IV. Measuring Performance: the OECD Experience

How do other societies fund their universities as performers? The closest comparators are likely within the OECD countries. Most fund universities through a combination of input factors – or resources universities use – and outputs – teaching and research. Typically these outputs are measured quantitatively: the number of credits that students accumulate, the number of students that receive degrees, the number of research publications, and the number of patents and licenses researchers receive. In addition, some governments include measures of the relative success of graduates on the labour market, the number of graduates working in jobs related to their training, and the success of universities in attracting outside funding.<sup>7</sup>

Three problems are immediately obvious. First, is the relative inattention to quality, largely because of the inherent difficulties in measuring quality across different domains. Consequently, as Jongbloed and Vossensteyn observe, “Every output indicator...will have its shortcomings. The main reason for this is that the services of a university are not sold in a kind of market where supply is meeting demand and prices reflect costs, quality, and scarcity.”<sup>8</sup>

Second, some of the performance measures clearly reflect the political agendas of funding governments. This should come as no surprise. Measures of effectiveness - the political terrain on which the battles of accountability are fought - are constructed in society. Although the polemics revolve around measures that is often not what the debates are about. Hidden beneath the polemics are arguments about goals and values, where measures serve as surrogates. The fundamental issues of any conflict over policy come alive in how we choose to count the dimensions of the problem. The choice of measures becomes deeply political, reflecting these often hidden differences.

Finally, it is obvious how universities can treat these measures as incentives when their funding is conditional on their performance. Universities may accelerate their students’ progress by lowering standards, they may encourage their faculty to publish quantities of research rather than focus on quality, and they may skew their teaching programs to align closely with current employment opportunities. Any of these would be a serious mistake which could undermine the capacity of the university to deliver a “high quality performance” over the longer term.

How significant is performance-based funding as a component of general funding of universities within the OECD? Less than one might think. In 11 of the OECD countries, teaching and research budgets are “lumped” together and universities are free to use the funding in any way they see fit. In Germany, for example, the total funds allocated are based principally on last year’s budget, with appropriate adjustments for price changes and policy adjustments.<sup>9</sup> Only in

Belgium, New Zealand, and the United Kingdom is research funded on the basis of a formula which incorporates some performance measures.

In Belgium and New Zealand, funding is driven largely by student numbers or the number of graduates. Only in the UK does research funding depend significantly on measures that tries to build in some qualitative measures amidst the largely quantitative indicators. The funding councils rate departments on their quality once every four to five years, using peer reviews. Those who are ranked highest receive approximately four times as much funding for research infrastructure as do those in the middle, while those at the bottom receive no funding at all. Surprisingly, once universities receive the funds, they are free to allocate the funds across departments in any way they wish. Recently, the United Kingdom has begun to try to reduce the burden of data collection and analysis that the Research Assessment Exercise requires.

A review of the funding of higher education among OECD countries finds that the role of output measures in funding decisions is relatively small. "With a few exceptions," Jongbloed and Vossensteyn observe, "one cannot speak of a high degree of performance orientation" in the OECD countries.<sup>10</sup> Performance measures are more important as learning tools for universities and for students than they are as mechanisms of accountability. That is probably as it should be.

## **V. Constructing Performance Measures**

We need to be measured in our use of performance measures, mindful of their capacity to drive behaviour as well as their capacity to assess achievement. We need to understand as well that performance measures tend to focus on quantity, at the expense of the much more difficult challenge of quality.

First, we must minimize the incentives of a set of performance measures to reduce the quality of research and teaching in order to increase quantity. When we count the number of students that we graduate, for example, we must also examine student-staff ratios, the size of classes, and ask students to evaluate the quality of their experience. Are students able to meet with faculty during a semester, are they given opportunities to participate in research? Taken together, a composite of these measures is more likely to reflect how well a university is meeting its obligations to its students.

Graduate students cross the boundary between teaching and research and blur the two in helpful ways. When we recruit graduate students, are we getting our first choices? Where do our graduates go when they leave us? Do they go to other excellent universities, to leadership positions in the private or voluntary sector, or in the public sector? Do they work mainly in Canada or internationally? What have our graduates accomplished ten years after they leave us? We need to ask them whether they feel they were given the highest quality education to prepare them for the careers they have followed.

When we count the research dollars a university receives, and add up the number of faculty publications and honours, we also need to ask faculty whether they feel they have adequate time to do research. Are younger faculty mentored by more senior faculty? Do they get helpful advice and support? Is the research infrastructure adequate? Citation counts are helpful in assessing quality, but can be misleading as well. Studies often cite other studies that support their research findings, or attack a line of investigation that is critical. We need to explore reputational effects: where do other excellent universities send their students to do graduate work? Who is doing innovative and exciting work in this field and where are they? When we are recruiting faculty, are we getting our first choice most of the time, some of the time, or rarely? Why? How are we different from other departments working in the same field? These kinds of questions are likely to turn up answers that can be helpful to departments as they evaluate their record and seek to improve their performance. They are, in other words, useful learning tools.

Performance measures, if they are to be useful in encouraging change, need to be dynamic rather than static. They need to help us understand processes rather than take snapshots at a given moment in time. We need to ask faculty who is doing innovative and experimental work in their field. Where are new approaches and discoveries likely to come from? What have we tried in the last five years that has succeeded? What have we tried that has failed? Answering that second question is just as important as answering the first. *If we cannot identify any failures, it is very unlikely that we have a culture of risk-taking at all.* What programs have we closed to make room for innovative, high-risk, ambitious research?

For performance measures to be valuable to universities and to those they serve, they must sit lightly. Collection and analysis of data can consume millions of scarce dollars and even scarcer faculty time and energy. The Research Assessment Exercise in Britain, which is widely regarded as onerous, takes place once every five years for research units and departments. Therein, I suspect, lay an important message. Annual performance measures do not tell us much; the variation can be the product of small shifts that are not terribly meaningful. We need to lengthen the time horizon of evaluation, and look for improvement that is built over several years. Short time horizons work against innovation and risk-taking, and in favour of continuing to do what we are already doing.

Designing performance measures that focus on quality requires careful and deliberate staging over time, so that the building blocs are put in place one after another without skewing the system of higher education. The measures must include subjective as well as objective indicators if we are to capture quality. They must explicitly build in indicators that help universities to learn and

to improve. They must have long enough time lines so that risk-taking and intelligent failures can be recognized.

Above all, we must avoid the trap of monetizing the value of a university education. No Auditor-General can reach a judgment about whether universities are providing “value for money.” In the choice of a light basket of measures, phased in over time, we can help to inform the judgment of whether we are doing the very best that we are capable of doing. Which measures we put into the basket of accountability will shape the system of higher education over the next several decades. Measures and numbers are contested because they are surrogates for political conflict over values. Our world, G.K. Chesterton wrote a century ago, “looks a little more mathematical and regular than it is; its exactitude is obvious, but its inexactitude is hidden; its wildness lies in wait.”<sup>11</sup> As we construct a more demanding architecture of accountability, we need to understand clearly why we are counting what we are counting, who chooses the measures and how they are chosen, how the measurers and the measures are politically connected, and what incentives these measures will create.

## ENDNOTES

- 
- <sup>1</sup> Robert Keohane, *Political Accountability*, paper prepared for Conference on Delegation to International Organizations, p.8.
- <sup>2</sup> James G. March and Johan Olsen, *Democratic Governance*, New York: Free Press, 1995, 157-8, 161.
- <sup>3</sup> Report by the Comptroller and Auditor General, *The Financial Health of Higher Education Institutions in England*, London: House of Commons, 1994.
- <sup>4</sup> See Nicholas Wade, "New Hughes Haven for Science Reaches for the Stars," *New York Times*, October 19, 2004, D2.
- <sup>5</sup> University of Toronto: Office of the Vice-President and Provost, *Performance Indicators for Governance*, Annual Report, September 2004.
- <sup>6</sup> Deborah Stone, *Policy Paradox and Political Reason* (New York: Harper Collins, 1988), p.136.
- <sup>7</sup> Ben Jongbloed and Hans Vossensteyn, "Keeping up Performances: an international survey of performance-based funding in higher education. *Journal of Higher Education Policy and Management* 23 (2, 2001): 127-145.
- <sup>8</sup> Jongbloed and Vossensteyn, "Keeping up Performances," 129.
- <sup>9</sup> Jongbloed and Vossensteyn, "Keeping up Performances," 132.
- <sup>10</sup> Jongbloed and Vossensteyn, "Keeping up Performances," 135.
- <sup>11</sup> G.K. Chesterton, *Orthodoxy* (New York: Dodd, Mead, 1908, republished Image Books, Doubleday, 1990), p. 81.