Measuring the Quality of Post-secondary Education: Concepts, Current Practices and a Strategic Plan

Ross Finnie
Alex Usher

Research Report W|28
Work Network

April 2005
Measuring the Quality of Post-secondary Education: Concepts, Current Practices and a Strategic Plan

Ross Finnie
Queen's University*
and
Alex Usher
Educational Policy Institute**

April 2005

* The School of Policy Studies, Queen's University
  Kingston, ON, K7L 3N6
  Tel.: 613-533-6000, ext.74219, E-mail: ref@post.queensu.ca

** Educational Policy Institute (Canadian Office)
  1701-77 Bloor St. W, Toronto, ON, M5S 1M2
  Tel: 416-848-0237, E-mail: ausher@educationalpolicy.org

© 2005 Canadian Policy Research Networks Inc.
# Table of Contents

Foreword........................................................................................................................................... i  
Executive Summary ............................................................................................................................ iii  
Acknowledgements.............................................................................................................................. v  
Introduction........................................................................................................................................ 1  

   1.1 An Overview.......................................................................................................................... 5  
   1.2 The Minimum Standards Approach...................................................................................... 6  
   1.3 The “Rankings/Indicators” Approach................................................................................... 8  
   1.4 The “Learning Impacts” Approach....................................................................................... 12  
   1.5 The “Continual Improvement” Approach ............................................................................ 15  
   1.6 Assessing the Assessments ................................................................................................... 17  

2. The Conceptual Framework............................................................................................................. 19  
   2.1 The General Framework........................................................................................................ 19  
   2.2 Using the Framework as a General Starting Point for Measuring Quality ......................... 21  
   2.3 Thinking in Modelling Terms............................................................................................... 22  
   2.4 Estimating the Relevant Relationships................................................................................. 25  
   2.5 General Approaches to “Measuring Quality” in the Context of the Framework................ 27  
   2.6 Summarising the Role of the Framework............................................................................. 28  

3. Next Steps: Measuring Quality in the Short- and Longer-Run....................................................... 29  
   3.1 Empirical-Type Analyses in the Short-Run ......................................................................... 29  
       The General Nature of Such Analyses.................................................................................... 29  
       Potential Data Sources for Short-Run Analyses................................................................... 31  
   3.2 Short-term Measure No. 2 – Quality Self-Monitoring............................................................ 34  
   3.3 Longer-Run Measures ......................................................................................................... 34  

Conclusion............................................................................................................................................ 37  

Appendix A – Measuring Institutional Research Quality .................................................................... 39  

Bibliography........................................................................................................................................ 43  

Our Support......................................................................................................................................... 47
Foreword

Universities and colleges have experienced a radical shift in circumstances in the past 20 years. The demand for advanced education has soared as the new skill requirements for a good job escalated. Yet, at the same time, governments have been cutting transfers to cover operations. This revenue has been replaced in part by higher tuition fees and more active fundraising programs. But there have also been notable changes in the way university programs are organized such as the increase in class size and greater use of part-time lecturers and sessional teachers. Nevertheless, educational attainment of the Canadian population has increased dramatically as young people and mature workers have expanded their participation in higher education.

Canadians are vitally aware of the importance of advanced education. They know that Canada’s competitiveness and their own future living standards will depend on the knowledge and skills of the work force. The main worry, as tuition costs increased, has been whether all qualified students have been able to access the appropriate university or college program. Both governments and philanthropists have therefore made significant investments in student aid through loans, grants, and bursaries. And CPRN has contributed several major papers examining their success in ensuring access (available at www.cprn.org).

So far, however, few questions have been asked about trends in the quality of the learning experience of students. Are Canada’s universities and colleges able to match the quality of teaching and research provided by the best schools in other countries? What policy changes might be needed to allow our post-secondary institutions to realize their potential for excellence without compromising access?

With the help of funding from the Wilson Foundation, CPRN has launched a series of studies of quality issues in our post-secondary system. Our goal is to launch a national conversation about this issue. This first report, by Ross Finnie of Queen’s University and Alex Usher of the Educational Policy Institute sets the stage. It provides a broad conceptual framework for thinking about the quality of the learning experience. It also surveys of the current state of the measurement of the quality of post-secondary education and identifies strategies for improvement. The authors are optimistic about Canada’s capacity to build on existing sources of data to become effective in tracking quality.

The next two studies in the Post-secondary Education Quality Series will look at the incentive structure (funding and regulation) facing post-secondary institutions in Canada, and the scope for (and examples of) differentiation and innovation within the current structure. A third and final paper will synthesize the key findings and explore the policy implications.

I would like to thank Ross Finnie and Alex Usher for their important contribution to our understanding of how we can move forward to develop sound quality measures. I would also like to thank The Wilson Foundation, the Ontario Ministry of Training Colleges and Universities (Post-secondary Review Secretariat), Power Corporation of Canada and First Plazas Inc. for their
financial support for this study. A draft of the paper was thoroughly reviewed at a national Roundtable involving over 40 participants who provided stimulus to the final version. Finally, a special thanks to Red Wilson, who not only posed the question in the first place, but also invested half of the money required to come up with some answers.

Judith Maxwell
April 2005
Executive Summary

This paper addresses the issue of the measurement of “quality” in post-secondary education. It begins by providing a critical examination of current practices in quality measurement in Canada and around the world. It then proposes a conceptual framework to guide thinking about the relevant issues. This is followed by a discussion of the general kinds of data needed to estimate the relationships in question. We then suggest what might be done with existing data sources in the short-run to undertake various quality measurement exercises, and what a longer-run plan might look like. We also suggest alternative approaches to measuring and promoting quality which are less founded on rigorous empirical exercises and more focussed on qualitatively-oriented evaluation procedures which could be implemented almost immediately.

We consciously choose not to join the debate over what the specific purposes of the PSE system should be, or what particular factors are associated with a better or lower quality education. Instead, we concentrate on developing a general approach for addressing these issues and proposing a broad analytical tool to help identify the factors that limit or contribute to better PSE outcomes. This analytical tool can be used in different ways by different users.

The conceptual framework is simple, capturing the PSE experience as a story of inputs and outputs whose narrative flows like this:

- **Beginning Characteristics** – the characteristics and abilities of incoming students that affect the quality of their educational experience and the outcomes.
- **Learning Inputs** – the institutional financial resources, material inputs and the organization of those resources which thus comprise the determinants and characteristics of individuals’ learning experiences; all are at least potentially controllable and thus amenable to change, improvement, and policy initiatives at various levels.
- **Learning Outputs** – the “skill sets” or any other attributes of graduates arising from their educational experiences that help determine final outcomes.
- **Final Outcomes** – the more specific “ultimate ends” to which the educational system may contribute – everything from employment, income and job satisfaction, to civic participation and continued education.

The notion of quality we propose thus focuses on the value-added of the educational experience. The ‘higher quality experiences’ are those that result in superior learning outcomes, and better final outcomes. And in the end, identifying the relevant value-added relationships of interest is an empirical exercise, since sitting and thinking about what quality is will not tell us what really matters in practice – and what matters in different ways in different situations.

We discuss, with the help of this framework, the advantages and limitations of current “quality measurement” exercises, from the annual Maclean’s university issue to the self-evaluation exercises undertaken by institutions. Each of these carves out a limited part of the overall quality assessment framework and generally undertakes the ‘estimation’ in a less-than-ideal manner in terms of the data
employed (many of them ignore students’ beginning characteristics, for example), how those data are manipulated, the spin put on the findings, and in other respects.

With this framework in hand, we then review the data sources currently available and find significant gaps in terms of the requirements pointed to by our framework. Even so, enough data are currently available to estimate many of the empirical relationships the model suggests and thereby advance the cause of measuring educational quality.

For the longer term, we recommend further data collection to fill the gaps. The goal is a database that is longitudinal, with full sets of information on beginning characteristics, inputs, learning outcomes and final outcomes – these (not surprisingly) matching the elements of our conceptual model.

There is no ‘silver bullet’ in looking at educational quality. There are no simple measures you can point to and say “Yes, there is quality. Let’s have some more of it.” What we hope we have produced here is, instead, a framework for thinking about this important and difficult issue in an intelligent fashion, and some specific, practical suggestions for how we can move towards doing a better job of measuring quality in both the short- and longer-run. With this, the cause of making for a better, more accountable post-secondary education system should be advanced more than with all the smoke, mirrors, and hot-headed debates that have typically characterised ‘the measurement of quality’ to date.
Acknowledgements

The authors would like to gratefully acknowledge the following people who assisted in the development of this paper: Kelly Gallagher-MacKay for her assistance in developing a literature review on world-wide trends in quality measurement and accreditation; Amy Cervenan for doing the same with respect to developments in Canada; Judith Maxwell, Ron Saunders and two anonymous reviewers who provided insightful comments on a near final draft of this paper; the many participants in the December 7, 2004 CPRN roundtable on measuring educational quality who provided invaluable guidance in the development of the final draft; the CPRN project management and publications team – Heather Fulsom, Gisèle Lacelle and Trish Adams – for seeing the manuscript through the various stages of production; and (again) Ron Saunders, whose guidance, encouragement and patience throughout this project has been crucial to its completion.
Introduction

The history of attempts to measure and control quality in post-secondary education is not a particularly happy one. Many of these attempts have been initiated by governments as part of efforts to introduce accountability and transparency in their relations with universities and colleges. Not unreasonably, they have tried to find metrics for “quality” that would allow them to understand and evaluate the return they are receiving on their rather substantial investments in post-secondary education. Many of these attempts to measure “quality” have been met by suspicion and hostility by the educational institutions being measured; with some justification, they have pointed out that most of the very simple output indicators used by governments to measure quality tend not to take into account factors such as differences in locality, academic selectivity and the choice of courses offered at the institution. Debates on quality therefore tend to end up in mutual suspicion and recrimination: governments tend to think of institutions as unwilling to make themselves accountable, and institutions (especially universities) tend to think of governments as dangerously simple-minded if not actually Philistine.¹

Not all of the anger in this debate is due to disagreements over metrics. In many instances, the problem is that while everybody wants a “quality” post-secondary education system, institutions, governments and other stakeholders do not always agree at the even more fundamental level of how quality should be defined. Where there is no stated agreement as to the purpose of universities and colleges, either individually or collectively, there is no obvious way to even begin to measure each institution’s – or a complete system’s – degree of excellence in meeting its purpose. In such a context, measuring quality is an ill-defined (and perhaps ill-fated) exercise from the very start. These might seem to be trivial and arcane matters were there not material issues at stake.

The quality measurement issue has also been plagued with the range of competing agendas in play. In many instances, for example, attempts to measure quality have been undertaken as part of a larger agenda, driven by diminishing trust and confidence in large institutions, to make educational institutions more “accountable” to governments. While there is no inherent contradiction between institutional autonomy on the one hand, and transparency and providing value for money on the other, attempts to measure quality have often foundered on these rocks. In related debates, quality measurement is frequently the subject of charges that it leads to standardization. It is sometimes argued, for example, that if quality measurements target a certain set of inputs (e.g., professor-to-student ratios, the number of books in the library, the availability of high speed internet connections in residence rooms), or outputs (e.g., employability, post-graduation income, etc.), then all institutions will begin to target these outcomes and diversity or “non-standard” or hard-to-measure approaches to achieving quality may suffer.

Clearly, then, discussions of the measurement of quality have the potential to be very controversial, entwined as they frequently are with issues of standardization, accountability, power struggles over the meaning of quality – and many other such divisive debates. Yet the politicization of quality measurement issues need not be inevitable. This paper, for instance, will look at issues of quality

¹ See Skillbeck (2001) for a balanced and thoughtful discussion of the entanglement of the quality and accountability issues
measurement, but it will remain resolutely neutral with respect to the relationship between quality measurement and accountability measures. It attempts to arrive at principles for measuring quality that can apply regardless of the formal accountability procedures attached to them.

Perhaps even more importantly, this paper will not project a definition of purpose onto the Canadian post-secondary education system, nor (therefore) will it define “quality”. The authors believe that while the country may sorely need some greater certainty concerning the definition of quality, the job of determining the ultimate goals of post-secondary education, from which quality measures ought to derive, properly belongs to governments, institutions and stakeholders. However, we do provide examples of what some of these goals might be.

What this paper will attempt to do is propose a conceptual framework that represents a neutral tool for thinking about these issues, while our survey of the current state of quality measurement and our identification of some strategies for improving quality measurement should be applicable regardless of the definition of quality adopted. Alternatively, the conceptual framework and broad set of empirical approaches it proposes may help policy makers and others arrive at definitions of quality in a more inductive fashion by identifying the aspects of an education experience that generate “better” outcomes as we decide to numerate these.

The purposes of this paper are therefore as follows:

- to provide a critical examination of current practices in higher education quality measurement in Canada and around the world;
- to propose a comprehensive framework for discussions pertaining to the measurement of quality which can help establish a set of principles for measuring it;
- to examine the data requirements of various approaches to measuring quality and to compare these to existing sources and those that might be developed for these purposes in the future;
- to suggest what might be done with the identified existing data sources in the short-run to undertake various quality measurement exercises;
- to suggest alternative approaches to measuring and promoting quality which are less founded on rigorous empirical exercises and more focussed on more qualitatively-oriented evaluation procedures which in a sense define quality in terms of the evaluation process;
- to identify other structures or practices that might be used to enhance the quality of higher education in Canada.
It should be noted very clearly that this paper focuses primarily on the quality of the education provided by universities to its students, and does not touch directly on how universities perform in terms of their research mission. While the latter is of course a fundamental and vital function of the post-secondary education system, the issues are substantially different. With respect to the measurement of research quality, we confine ourselves to a short overview of the issues and current practices in this sector in an appendix to this paper.

1.1 An Overview

In many OECD countries, there has been a growing interest in exploring the notion of “quality” in post-secondary education over the past twenty years. This growing interest has been co-incident with a number of other developments that also spanned the OECD: increases in enrolment, increases in the returns to education, restraint in public expenditures, and moves to increase transparency and accountability in public and quasi-public institutions. As a result, discussions about quality have been entwined with some of these issues. In some countries, including Canada, this mixing of the issues has tended to stifle discussions about quality because some see them as a slippery slope towards other, often unwanted agendas about accountability.

Broadly speaking, there are four approaches to quality measurement across OECD countries. The first approach is a largely qualitative, process-oriented one that seeks to ensure minimum standards. As a result, it is usually tied to an agenda of accountability to government, although Canadian institutions have used the minimum standards approach without any reference to governments. A second approach is largely quantitative and competitive and generally seeks to mark progress over time and to rank institutions against one another. Sometimes, the quantitative approach is linked to an accountability agenda as in the case of Key Performance Indicators (KPIs) in Ontario and Alberta. In other cases, as with rankings tables produced by private-sector publishers in a dozen countries, (e.g., Canada, the U.S., Germany), it is linked with a transparency agenda. A third approach – which has arisen largely in opposition to the second – is based on what might be termed learning assessments, such as the National Survey of Student Engagement (NSSE) in Canada and the United States and the Graduate Skills Assessment in Australia. Though not generally linked with the ranking agenda, the two approaches do share a quantitative bent. Finally, though little used, there is a fourth tradition, which lies very much within the Japanese tradition of kaizen, or continual improvement. This approach has a certain fraternity with the ISO 9000 approach to quality in that what is measured is not quality per se but adherence to a set of procedures designed to monitor and promote quality.
Table 1 – Approaches to Quality

<table>
<thead>
<tr>
<th>Approach</th>
<th>Agenda</th>
<th>Methodology</th>
<th>Unit of Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Standards</td>
<td>Accountability to Government</td>
<td>Mostly Qualitative</td>
<td>Usually departments</td>
</tr>
<tr>
<td>Rankings/Indicators</td>
<td>Accountability to Government, Transparency</td>
<td>Quantitative</td>
<td>Usually institutions</td>
</tr>
<tr>
<td>Learning Impacts</td>
<td>Internal Accountability</td>
<td>Quantitative</td>
<td>Institutions</td>
</tr>
<tr>
<td>Continual Improvement</td>
<td>Internal Accountability, Accountability to Government</td>
<td>Qualitative</td>
<td>Usually institutions</td>
</tr>
</tbody>
</table>

Though these four approaches differ considerably in character, they are not mutually exclusive. Each will now be discussed in turn.

1.2 The Minimum Standards Approach

The minimum standards approach is the most common throughout the OECD. It is the foundation of the periodic review process (often called “cyclical reviews” or “program reviews”) process that all Canadian universities use at a departmental level, and indeed most countries use some form of it at the discipline level. The United States is somewhat unique in using the same minimum standards approach at the institutional level through its process of institutional accreditation.

Perhaps the most important reason that the minimum standards approach is used so broadly is that monitoring and licensing of institutions is in most countries understood to be a legitimate function of government in a way that ranking institutions is not. To use a policy analogy: the regulation of food services is accepted as a legitimate government responsibility, but the preparation of “Best Restaurant” guides is not. As a result, in European OECD countries, the minimum standards approach has always gone hand-on-hand with some form of government participation or oversight. In the United States (where oversight is done by regional accreditation commissions) and in Canada, however, the process occurs without much in the way of such government oversight.

The minimum standards process typically starts with a self-audit. This process may vary somewhat from place to place, but the principle behind it is common enough. During the self-audit, the unit under review gathers information on the quality of its program from a number of different sources – comments from students, alumni, and employers; faculty publication records, etc. In jurisdictions where regional or national organizations have put in place certain recommended standards, the institution or unit must use these data to self-assess whether or not it is meeting these standards. American regional accrediting boards require these kinds of self-audits, as do the many professional program accrediting boards in Canada. In many countries, the self-audit is then followed up with

---

2 Jongbloed (2005).
3 Despite the vast size of the enterprise, there is very little in the way of useful literature on accreditation in the U.S.. For one of the better articles on this subject, see American Council of Trustees and Alumni (2003).
some form of external examination. In Canadian universities, this is done simply by inviting colleagues from other institutions to participate in the self-audit.

In some jurisdictions (such as Japan and Australia) there is a formal second step following the self-audit in which an external body in effect confirms the results of the self-audit through its own site visit.

Table 2 – Features of “Minimum Standards” Approach in Selected Countries

<table>
<thead>
<tr>
<th>Country</th>
<th>Unit</th>
<th>Self-Audit</th>
<th>Expert Panel</th>
<th>Government Role?</th>
<th>Results Public?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canada</td>
<td>Department</td>
<td>Yes</td>
<td>Usually academics</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>United States</td>
<td>Institution and Department</td>
<td>Yes</td>
<td>Academics, trained accreditors</td>
<td>No</td>
<td>Varies</td>
</tr>
<tr>
<td>France</td>
<td>Institution</td>
<td>Yes, in accordance with govt. guideline</td>
<td>Academics, economists</td>
<td>Drafts a final report based on self-audit and expert panel</td>
<td>Yes</td>
</tr>
<tr>
<td>Germany</td>
<td>Departments</td>
<td>Yes</td>
<td>Yes</td>
<td>Länder certify accreditation agencies responsible for different subject areas</td>
<td>Yes</td>
</tr>
<tr>
<td>Japan</td>
<td>Institution and Department</td>
<td>Yes</td>
<td>Government-certified evaluation agencies, includes academics and business</td>
<td>Indirectly, by certifying evaluation agencies</td>
<td>Yes</td>
</tr>
<tr>
<td>Ireland</td>
<td>Institution</td>
<td>Yes</td>
<td>Usually academics</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

There are several critiques of the minimum standards approach which are worth examining. The first is that the approach’s reliance at base on a self-examination is less than objective and that outside experts, particularly when they come from within the same institution or within the same discipline, may be open to implicit “log-rolling” (experts avoid making tough statements at other universities in order to avoid having tough statements made about them at their own institutions).

---

4 A description of the French system of academic evaluation may be found at: www.cne-evaluation.fr/fr/present/som_met.htm
5 A description of the German system of academic evaluation may be found at the website of the lander-created Akkreditierungsrat at www.accreditation-council.de/
6 A partial description of the Japanese system may be found in Japanese Ministry of Education, Culture, Sports, Science and Technology (2002).
7 Sarbak (2004).
The minimum standards approach tends to generate reports which may be useful as an internal management tool but are of little use for outsiders in trying to assess the relative strength or weakness of a department or institution. Often phrased in academic and management jargon, they do little to foster “transparency”. This is not always a hindrance to effectiveness, and where governments participate in the process the lack of transparency is offset by the fact that at least one major external stakeholder understands the process and has some “ownership” of the results. In North America, however, where government is not involved in these processes, the “minimum standards” process is both opaque and exclusive. As a result, it has been in North America that the desire for more transparent and easily understandable measures of quality has been greatest. Hence it was here that the greatest interest has been shown in quantitative indicators and easy-to-understand rankings.

1.3 The “Rankings/Indicators” Approach

The “Rankings/Indicators” approach to quality measurement was an outgrowth of several phenomena that coalesced during the 1980s through early 1990s. Following the proliferation of both students and institutions that occurred at different times in different OECD countries between the 1960s and 1980s, educational consumers (mostly parents) began to search for more comparative information regarding educational institutions. At the same time, as many OECD governments began to engage in some form of fiscal restraint, some began to look increasingly at “value-for-money” auditing techniques to evaluate the effectiveness of their spending on post-secondary education. Both phenomena generated a great interest in quantitative data regarding institutions, which the dominant “minimum standards” approach of quality assurance was ill-suited to provide.

In many different countries, private-sector publishers took the lead in providing quantitative data on institutions. Generally speaking, these took the form of institutional “rankings” or “league tables”. U.S. News & World Report was the first to do a major one at the undergraduate level; Maclean’s in Canada, the Times Higher Education Supplement in the UK, Asiaweek for East and South-east Asia are other examples of this phenomenon. Other publications focus on graduate education; others still developed rankings for particular disciplines (no fewer than four international publications currently dispense worldwide MBA school rankings).  

Generally speaking, these institutional rankings focus on input measures. In both Maclean’s and U.S. News & World Report, over 80% of the total score is generated directly or indirectly by the amount of resources a school has (be these government grants, tuition or endowment wealth/income) or by the average secondary school marks of its entering class. Much of the balance is made up from survey data measuring conceptually fuzzy concepts such as reputation.

---

8 A considerable literature is now developing on rankings, in no small measure thanks to the work of UNESCO-CEPES, which, in conjunction with the Institute for Higher Education Policy has undertaken an international study of higher education ranking system. The first results of this collaboration may be seen in the special issue of Higher Education In Europe, vol. XXVII, no.4, 2002, which is entirely devoted to the topic.

9 In Macleans, 17% of the weight is given to class size, another 17% to faculty quality, and 12% each is given to “finances” and “library quality”. Thus, 58% of the weight is split among four categories, all of which, in the end, are functions of the amount of resources at an institution’s disposal. Another 23% is given over to measures of the student body (effectively measuring academic selectivity), with the remaining 19% being awarded on the basis of “reputation surveys”
While there is nothing wrong, per se, with students and parents choosing institutions on the basis of the amount of resources available per student, it is quite wrong from an empirical point of view to conclude that institutions with more resources are necessarily “better”. Certainly, better funded institutions are in theory capable of providing their students with a better learning environment, but it is not clear that this occurs in practice. Thus, while these rankings may be an excellent exercise in transparency (although there are certainly arguments to the contrary, as stories surface from time to time about institutions devoting considerable human resources to manipulating the data provided to the body doing the ranking) – above all else, they provide handy, easy-to-understand quantitative capsules that may be used by non-specialists. Indeed, it is precisely the ease with which rankings systems are understood that is their greatest strength; above all else, they bring transparency to institutions.

Governments, on the other hand, went in the other direction when developing Key Performance Indicators (KPIs). These became prevalent in the early 1990s as a means for U.S. states to tie institutional funding to some form of output measurement. By 1996, 36 states had tied some institutional funding (usually a very small portion) to Key Performance Indicators. Not all states that used performance indicators, however, used them as a funding tool and not all of them applied performance indicators at the institutional level. California, New Jersey and Ohio, for instance, collect performance indicator data but publish it only at the system level, not the institutional level. Texas publishes an astonishing amount of comparative institutional data, but not all of it is technically related to “performance”. Most performance indicators, though, have had some kind of funding attached to them, even if “performance-related pay” was usually a very small fraction of total institutional income. In the United Kingdom, the Higher Education Funding Council for England also collects and publishes performance indicator data but does not attach funds to “good” performance in particular areas.

In Canada, performance indicators have had a shorter history. A number of institutions (e.g., University of Calgary, University of New Brunswick, Conestoga College, Malaspina College) have

10 This gives rise to a topic which is important to mention even if it cannot be sufficiently tackled within the confines of this paper; which is that different audiences may have different definitions of quality. Students and parents may care more about $/student; researchers may care more about the quality of research infrastructure, faculty may care about prestige, governments may care about “value-added”. None of these is a priori “better” than another as a definition of quality – each may be appropriate depending on the audience and the context. This is yet another reason to use a quality assessment methodology which is flexible as to “final outcomes”

11 Heather Sokoloff, National Post January 31, 2004

12 See Bruneau and Savage (2002) for a detailed, if highly jaundiced, history of performance indicators and an overview of their use in different jurisdictions. A more balanced view – although one still highly sensitive to the problems of using quantitative measurements in measuring education – may be found in Grosjean et. Al (2000)


17 See the Texas Public Universities Data and Performance Report: www.thecb.state.tx.us/reports/pdf/0759A.pdf

informally used different types of performance indicators internally.\textsuperscript{19} Although British Columbia, Quebec and Newfoundland, to varying extents, collect and publish performance indicator data, Alberta and Ontario are the only provinces where institutional performance indicators are actually tied to funding, and even here they are for very small portions of the overall funding envelope (two percent or less).

Among those jurisdictions that publish institutional performance indicators (whether as part of a funding formula or not, institutional graduation rates (additional variations on this theme are at retention and persistence) appear to be the only key indicator that has universal acceptance as a measure of institutional performance. Beyond this core indicator, different jurisdictions use very different measures depending on local policy contexts. Employment indicators are widely used, as are measures of graduates’ satisfaction with their education. Financial performance appears just about everywhere, and is usually measured by having low administrative overhead (Massachusetts actually uses no less than six different financial performance measures). Less frequent are items such as faculty diversity, student default rates, fundraising performance, and quality data reporting. Table 3 shows the indicators in use in selected North American jurisdictions.

KPIs have avoided the methodological trap of the “rankings” approach by not relying on input measures as a proxy for quality. However, since the quality of inputs has an effect on the quality of outputs, equating outputs with “quality” (as KPIs implicitly do) merely re-creates the same problem in mirror image. Imagine, for instance, that hospitals were ranked based on the mortality rates of their patients. Before rushing to conclusions about the “quality” of each institution based on such rankings, one might want to know what kind of patients each hospital admitted. So, too, with educational institutions; judging open-access institutions by the same standards as ultra-selective ones without making adjustments for the nature of the student body guarantees biased results.

\textsuperscript{19} An excellent survey can be found in the appendix to Koch (2003).
Table 3 – KPI variables in selected jurisdictions

<table>
<thead>
<tr>
<th>KPI variable</th>
<th>ON</th>
<th>AB</th>
<th>IL 20</th>
<th>WA 21 **</th>
<th>MA 22 †</th>
<th>MD 23</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employment Rates</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employment rate plus Enrolled in School</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Graduation rates</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Persistence/retention rates</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintained/increased enrolment</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increased diversity of enrolments</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diversity of Faculty</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student Loan Defaults</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Net Cost</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graduate Satisfaction</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Administration Costs</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Cost of instruction</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of teachers produced</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X††</td>
<td></td>
</tr>
<tr>
<td>Avg GMAT/LSAT/GRE scores</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class size</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faculty Workload</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fundraising</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collaboration with K-12 system</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quality Data Reporting</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research Indicators</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

** In addition to having six state-wide indicators, each public campus is permitted to set and monitor two of its own performance indicators.

*** Maryland’s indicators actually focus more on graduation rates of minority students – effectively, diversity of graduates rather than diversity of enrolments.

† In addition to having general targets on certain issues, the Maryland Higher Education Commission also sets campus-specific targets in very specific sub-fields (e.g., African-American graduation rate in Computer Science).

†† Pass rate at teaching licensures.

---


21 See Quality Indicator System Report


Measuring the Quality of PSE: Concepts, Current Practices and a Strategic Plan 11
The UK Quality Assurance experiment

As part of a general attempt to raise quality and customer service standards across government departments in the early 1990s, the UK government created the Quality Assurance Agency (QAA). The QAA, to put it mildly, was not well-liked. Its intrusive bureaucracy and widely-derided evaluation tools have become the bogeyman of quality measurement activities the world over. It is therefore worth a brief examination of what, precisely, went wrong with QAA.

In theory, the QAA approach fits somewhere between the “minimum standards” approach (albeit one with a heavy external presence) and the rankings/KPI approach. For each of the 62 subject areas defined by the QAA, evidence-based specialist reviews had to take place. This took place in the first instance through self-audit, although the requirement for “evidence-based” self-audits generated a tremendous amount of paper. This would be followed by a 3-4 day visit from a group of at least four outside experts who would give scores of between 1-4 on the department’s performance in six areas (curriculum design, teaching, student achievement, student support/guidance, learning resources and quality assurance procedures). One site visit practice that generated no small amount of controversy was the assessments of teaching in which outside evaluators would sit in classrooms and make notes about teaching styles. In theory, this was a progressive attempt to introduce sensitive, qualitative observations about learning into the assessment process. In practice, these class visits were the subject of bitter criticism and helped to foster a perception that the QAA was over-intrusive. Despite all this effort, QAA reports rarely showed any serious gaps in departmental quality. Over the course of 2004 such reviews between 1993 and 2001 (all of which were published), virtually all departments under consideration were considered to have scored between 19 and 23 out of a possible 24 points.

Since 2001, a “lighter touch” administration has been in place at the QAA. It no longer attempts to evaluate quality at the subject level, though it has continued to try to develop nation-wide benchmarks and standards for graduation in various subject fields. Though this effort has been a collaborative one with institutions, the idea of making definitive “meta-statements” about subject-level competencies (i.e., making absolute statements about the required competencies to be a graduate in a specific field, such as sociology) is still met with considerable scepticism. The QAA continues, however, to look at institutional quality by focusing on the institution’s internal means of quality assurance (see below, “Continuous Improvement”, for more details).24

1.4 The “Learning Impacts” Approach

After nearly a decade and a half where the public policy emphasis was on KPIs and ranking schemes, a backlash emerged from within the educational community itself. Educational institutions – not without reason – felt that the existing ranking systems were more of a hindrance than a help because they spoke to inputs or outputs but not to the actual process of learning that occurred within an institution. Since institutions consider themselves to be in business precisely to help people learn,

24 See the National Committee of Inquiry into Higher Education (1997) for a detailed critique of the “old” regime. For a review of the newer procedures, see Quality Assurance Agency for Higher Education (2003).
it seemed deeply unfair that “quality” was being judged on measurements which effectively ignored
the educational process. One observer likened the process of measuring educational quality through
such measures with the drunk who loses his keys in the street but goes to look for them under a
streetlight because “the light is better over there”.  But the simple fact was that inputs and outputs
were easier to measure and describe than the learning process: hence their attraction, especially to
policymakers.

Thus was set in motion a search for a set of indicators that would actually describe the effectiveness
of the learning process within institutions in a simple, easy-to-understand manner. With funding
from the Pew Foundation, George Kuh and a small group of researchers worked with a consortium
of educational institutions to develop such an instrument. The result was the National Survey of
Student Engagement (NSSE – pronounced “Nessie”), which was piloted in 75 institutions in 2000.
The program grew quickly to several hundred institutions and spawned a sister-survey known as the
Community College Survey of Student Engagement (CCSSE or “Sessie”). Nearly a dozen
institutions in Canada have now used the NSSE.

NSSE is a fairly simple survey, just four pages in length, asking students about their learning
experiences at institutions, such as average frequency and duration of homework, frequency of
contact with faculty or other advisors, number of books read for courses and for pleasure, etc.
Drawing on about three decades of research on the effectiveness of educational practice among
college students (in the American sense of the word), these results are then turned into a series of
institutional “scores” which describe how well the university does at creating a “learning
environment”.  

As its rapid spread through the North American educational community attests, NSSE has been very
popular among institutions. Each participating institution receives its own scores as well as those of
other institutions within its “peer group” (based on the institution’s Carnegie classification). The
dominant sentiment among those who use it is that it is a superior management tool – it provides
precise, quantitative data regarding aspects of the learning experience which can be used to modify
policy and practice within an institution. In this sense, it is simply a superior instrument which an
institution can integrate into its existing “self-audit” regime.

While the NSSE does in some sense represent an advance over the earlier input/output techniques, it
is not without its flaws. While each institution receives its own results and, for comparative
purposes, those of its peer institutions, these are not made public. Indeed, it is doubtful whether
NSSE could ever have gained such wide acceptance so quickly if there had been a hint that the
results would in fact be made public. Also important from an empirical point of view is that the
NSSE does not, strictly speaking, measure learning outcomes; instead, it measures the correlates of
good learning outcomes and assumes learning is taking place on the basis of it. Even if one accepts
this assumption, one must bear in mind that the NSSE is essentially content-free; it can determine

26 See Kuh (2001, 2003) for further details on NSSE. Details on the CSSE may be found at the website
www.ccsse.org
27 The Carnegie classification is the standard typology used to classify American post-secondary institutions;
effectively it is a much more sophisticated version of the Maclean’s institutional typology, with 18 categories
instead of 3. A full description may be found at www.carnegiefoundation.org/Classification/
whether “learning” is taking place, but says nothing about what is being learned. Methodologists may also question the accuracy of a survey that relies on students self-reporting on questions such as “how often have you worked harder than you thought you could to meet an instructor’s standards or expectations?”. Finally, although the relationship exists intuitively, there does not appear to be any substantial literature linking good “learning” results to future career and life outcomes.

NSSE is not the only example of attempts to measure learning impacts; in fact, the general approach was pioneered several years earlier in Australia. After the 1997 West Report on universities made general recommendations on the skills and attributes graduates should acquire, attempts were made to measure some of these attributes. As a result, the Australian Council for Educational Research developed the Graduate Skills Assessment (GSA) instrument in order to test for critical thinking, problem solving, interpersonal understanding and written communication. The GSA was not originally intended to look at institutional effectiveness; instead it was meant to be a reliable, objective tool for graduates and employers that could measure “soft skills”. More recently, however, some in government have advocated that it be used as a test for measuring progress over time in universities. In theory, a GSA approach remedies some of NSSE’s deficiencies in that it measures demonstrated learning outcomes, as opposed to implied ones. However, since there is no compulsory entry-level test (one has been developed but it does not appear to be widely in use), the GSA is open to the same critique levelled at KPIs: namely, that output measurements without complementary input measurements are essentially meaningless.

In the United States, the Council for Aid to Education (a subsidiary of RAND Corporation), in collaboration with about two dozen institutions, has recently piloted an approach to learning measurement that corrects for this problem. Like the GSA, the CLA is meant to test general skills such as ability to communicate and critical thinking; unlike the GSA, the Collegiate Learning Assessment (CLA) has both an entry and exit component, allowing a look at precisely what progress each individual student has made over the course of his or her studies. It could be argued that this approach is superior to NSSE; it measures learning directly instead of inferring it and it looks at

---

28 See West (1997). The report suggested that, ideally, every graduate with a first degree should have acquired the following attributes:
- the capacity for critical, conceptual and reflective thinking in all aspects of intellectual and practical activity;
- technical competence and an understanding of the broad conceptual and theoretical elements of his or her fields of specialisation;
- intellectual openness and curiosity, and an appreciation of the interconnectedness, and areas of uncertainty, in current human knowledge;
- effective communication skills in all domains (reading, writing, speaking and listening);
- research, discovery, and information retrieval skills and a general capacity to use information;
- multifaceted problem solving skills and the capacity for team work; and
- high ethical standards in personal and professional life, underpinned by a capacity for self-directed activity.


30 Neither the GSA nor the CLA are meant to examine narrow subject-level competencies; See the Collegiate Learning Assessment Conceptual Framework Document at: www.cae.org/content/pdf/CLA_ConceptualFramework.pdf and the Summary Technical Report at www.cae.org/content/pdf/technical_report.pdf. For a broader discussion of the CLA see Benjamin and Chun (2003).
individual students’ results as opposed to simply measuring the learning environment. However, the CLA is still in its infancy and has no track record to speak of; it remains a potential tool rather than an actual one. Still, the basic approach of testing general skills at more than one point in time to measure educational effectiveness is widely understood and accepted (if not always liked). It is, after all, the basic foundation on which K-12 measurement efforts such as Ontario’s Education Quality and Accountability Office were built.

1.5 The “Continual Improvement” Approach

While the “Rankings/Indicators” and “Learning Impacts” trends were gaining steam in North America in the 1990s, developments in Europe took a rather different turn. Although there was discontent with the “minimum standards” approach, there was no real discontent with the basic approach of self-audit followed by some external oversight. There were many innovations – such as choosing to do cyclical reviews of all departments nationwide in a particular field of study in a single year in order to get a picture of the nation’s ability to produce quality graduates in a given field (e.g., Chemistry) – but few of these entailed anything resembling a paradigm shift for quality oversight.

Yet there was some discontent with the minimum standards approach if for no other reason than that it provided no incentive for improvement. This thought led towards certain managerial concepts, common in for-profit enterprises, related to quality monitoring and improvement. These concepts come in a number of different guises; one well-known example is “performance benchmarking”; that is, an approach which required institutions to meet performance targets based on results at other organizations one wishes to emulate. Some U.S. states, notably Maryland, employ this type of strategy in their universities, as, increasingly, do the Australians. But perhaps the best known is the Japanese concept of *kaizen* (promoted in North America by Tom Peters), or “continual improvement.” This concept has in practice been enshrined through a set of concepts known as “ISO 9000”.

ISO 9000 (and its environmentally-oriented twin, ISO 14000) is a generic management system standard for quality management, which, according to the International Organization for Standardization, means what the organization does to fulfil:

- the customer’s quality requirements; and
- applicable regulatory requirements; while
- aiming to enhance customer satisfaction; and
- achieve continual improvement of its performance in pursuit of these objectives.

---

31 Indeed, during the beta-test phase, all students who took the CLA also took the NSSE in order to provide external validation.
32 This is in fact the case in France and Finland. It is interesting to note that the innovation of making certain reviews nationwide occurred in countries with reasonably centralized governments; it is difficult to imagine this kind of national exercise occurring in a federal system of government such as Canada’s.
33 See McKinnon and Davis (2000).
34 See www.iso.org for a description of the ISO 9000 certification process.
In other words, the ISO 9000 does not define quality as an objective product standard or benchmark. Rather, it defines quality as a process – a way of conducting one’s operations with a view to continual improvement. Virtually any organization can apply to be certified as being ISO-9000 compliant simply by documenting its quality monitoring and improvement techniques and demonstrating fidelity to them when audited by an accredited ISO 9000 certification agency.

Applied to education, this is an intriguing idea. Putting government in the role of accreditation agent, it allows institutions to set their own goals (albeit presumably with some form of government input). Each institution develops its own methods of defining and monitoring quality while the government’s role is to certify that institutions are in fact doing an adequate job of it. While this type of quality assessment may sound less rigorous than the benchmarking approach seen in KPIs, it is in fact a great deal more labour intensive at the institutional level as institutions have to devote significantly more resources to creating and analysing data on their own processes.

Sweden moved explicitly to this model of quality assurance in 1995 when the Hogkoleverket (National Agency for Higher Education) chose to define quality as “processes” (emphasis added) leading to enhanced quality in institutional operations. In 1998, Australia also moved some way toward this model when it created the Australian Universities Quality Agency (AQUA). Although the AQUA is in some ways a traditional American-style accreditation agency, the AQUA encourages institutions to develop data in support of their own planning processes and is increasingly concerned with auditing the effectiveness of the universities’ quality management process. In 2001, the UK’s QAA began to orient itself in this direction as well. Although there have been no moves towards using continuous improvement techniques at the university level in Canada, some community colleges have embraced the concept wholeheartedly: Columbia College in Calgary, Holland College in Charlottetown and Conestoga College in Southwestern Ontario have all received ISO 9000 certification.

ISO 9000 is not the only way to use self-study as a way to actively improve quality. Some American and Japanese universities have experimented with Total Quality Management while “Baldrige” criteria – really an extension of ISO to cover issues such as strategic planning, product development and “teamwork” - have also found adherents. These initiatives tend to be undertaken at an institutional rather than a systemic level, and even then they are more likely to be applied to non-academic units than academic ones. None of these have ever been explicitly applied in a jurisdiction-wide manner.

One clear drawback of this general approach is transparency. Because each institution measures its own metrics in its own way, the inter-institutional comparability of certain indicators is lost. Moreover, the oversight reports of institutional quality procedures – effectively a form of process audit – tend not to be easy reading, thus further harming a transparency agenda.

36 Emmanuel and Reekie (2004).
37 For an example of the first results of the new regimen, see Institutional Audits: Key Features and Findings of the First Audits, www.qaa.ac.uk/revreps/inst_audit/evalreport_main.htm
38 See Yudof and Busch-Vishniac (1996).
1.6 Assessing the Assessments

Having now done the global tour of quality measures, we summarise some of the pros and cons of these existing approaches to quality in Table 4.

Table 4 – Some pros and cons of various quality measurement techniques

<table>
<thead>
<tr>
<th>Approach</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Standards</td>
<td>Collegiality, sensitivity to context</td>
<td>Insufficiently transparent process and results</td>
</tr>
<tr>
<td>Rankings/Indicators</td>
<td>Quantitative results intuitively easy to grasp</td>
<td>Indicators frequently meaningless; insensitivity to context</td>
</tr>
<tr>
<td>Learning Impacts</td>
<td>Measures what happens “inside the box”, potential to unlock the true “value-added” of each institution</td>
<td>NSSE an indirect indicator lacking transparency, Direct indicators not yet well developed</td>
</tr>
<tr>
<td>Continual Improvement</td>
<td>Sensitivity to context, development of “quality communities” at each institution</td>
<td>Insufficiently transparent results</td>
</tr>
</tbody>
</table>

We now turn to present a general conceptual framework for thinking about quality and its measurement – wherever it might exist or however it might be measured.
2. The Conceptual Framework

2.1 The General Framework

We now propose a general conceptual framework, or model, for thinking about the various issues relating to the measurement of quality of post-secondary education. It is predicated on the idea that the notion of quality in which we are interested is focussed on the identification of i) inputs which determine and comprise the schooling environment and individuals’ educational experience and thus, affect ii) the learning outputs which result, which in turn influence iii) the final outcomes which we think ought to be enhanced by the education process, keeping in mind throughout that individuals enter the system with a set of iv) beginning characteristics which will affect their educational experiences, the learning outcomes they achieve, and the final outcomes they realise. The framework can be made more complex than this simple summary description suggests, but its basic structure remains the same: “inputs” determine the outcomes in which we are interested, and we evaluate quality according to the effects of different kinds of inputs while controlling for other relevant factors.

This notion of quality is thus, in broad terms, focussed on what may be principally thought of as the value added of the educational experience, whereby “higher quality experiences” are those that result in superior learning outcomes, and better final outcomes. These sets of relationships can be represented by the following very simple schematic (a slightly more formal representation of these relationships is shown and discussed in Section 2.3):

\[
\text{Beginning Characteristics} \rightarrow \text{Learning Inputs} \rightarrow \text{Learning Outputs} \rightarrow \text{Final Outcomes}
\]

The arrows do not represent purely causal flows in all cases. For example, although individuals’ beginning characteristics come before their exposure to learning inputs in the model, these characteristics will not generally affect those inputs, since individuals do not control the resources available to the institutions they attend or how those resources are organised to create the person’s learning experience. Individuals’ beginning characteristics can, however, be correlated with the learning inputs to which they are exposed (e.g., better students are more likely to go to better schools) or might otherwise interact with the inputs available and will in any event likely be related to learning outcomes and final outcomes and must, therefore, be taken into account in the general set of flows. In the other cases, the flows do indeed represent causal relationships.

The four sets of factors comprising this framework are as follows:

- **Beginning Characteristics**, represent the characteristics and abilities of incoming students as they start their programs, including critical thinking, analytic reasoning, communication skills, desire for learning, learning strategies, specific knowledge, or any other attribute which could affect their learning experiences, learning outcomes, or final outcomes. We might also be interested in attributes such as gender, family background, high school experiences, and other characteristics which either provide proxy measures of beginning characteristics or represent dimensions along which we might want to
breakdown any element of the general analysis of the quality of individuals’ educational experiences or the outcomes to which they contribute.

- **Learning Inputs** – as we define the term here – embody all the factors that reflect or determine students’ educational experiences and are at least theoretically controllable by the educational system (at the institutional or some other level) and therefore represent potential policy levers for affecting the quality of education. These “inputs” may be thought of as consisting of various different main types, ranging from i) financial resources available and dollars spent, through ii) material inputs, such as the facilities available to students (library books, IT services, etc.), to iii) how these resources are organised into the actual educational experiences to which students are exposed, such as the amount of contact time students have with their teachers, the kinds of exams they face, and so on (sometimes generally referred to as “pedagogies”). While these comprise quite diverse components to what we are giving a single term (“inputs”), we feel that the central idea of their comprising the determinants and characteristics of individuals’ learning experiences are sufficiently unifying and capture the general set of factors in which we are interested in evaluating “quality” and are therefore usefully grouped together in this way, even if in some places we separate them out along the lines just described.

- **Learning Outputs** represent the “skill sets” or any other attributes of graduates which culminate from their educational experiences (as well as their beginning characteristics). These include the sorts of general skills such as critical thinking, analytic reasoning, and communication skills mentioned above under beginning characteristics; softer skills such as being able to work with others; and technical knowledge such as how to bake a cake or build a bridge, how to market a product, or how to conduct a literature review. In short, “learning outputs”, as we use the term here, represent any skills or characteristics of graduates which the educational system potentially produces which, in turn, affect final outcomes.

- **Final outcomes** represent the “ultimate ends” to which the educational system may contribute, including not only such traditional measures as employment rates and incomes, but also any other outcome deemed to be important to individuals and society, such as job satisfaction, an appreciation of the finer things in life, “being a good citizen”, or continued participation in education through lifelong learning. (In this final sense, there can be recursive loops in the model; if a desired outcome is to create more lifelong learners, these learners will later return to post-secondary education with improved predispositions to learn (which is one of the beginning characteristics noted above). “Learning characteristics”, broadly defined, may therefore be both a beginning characteristic and an output). It should be noted that we assume, axiomatically, that the entire effects of the education system on final outcomes come through their effects on learning outcomes. We could allow for more direct effects, but this would not add to the model in any meaningful manner. We can also allow for the nature of the educational experience itself, (e.g., “Is the learning enjoyable”?, which in standard economics terms is known as its “consumption value”), to be a final output itself.
2.2 Using the Framework as a General Starting Point for Measuring Quality

Given this general framework, educational quality can be measured in a variety of different ways (e.g., using aggregate or micro data or studying the relationships in question in a more qualitative fashion), at different levels (e.g., a department, an institution, a system), taking different specific factors into account (e.g., focusing on overall differences across institutions and thus essentially treating them as black boxes or using institutions as simply a source of variation in specific inputs in order to identify which of these matter). But all these specific approaches can be placed in the general framework offered above, and consist in one way or another of identifying the relationships represented in that simple schematic. In short, we want to know the extent to which various inputs – such as dollars spent, pedagogies adopted, etc. – affect learning outcomes, and how these in turn affect final outcomes. The “better” inputs are our guide to “quality”.

For example, if the data (qualitative or quantitative), indicate that spending more money in a particular way (e.g., hiring more professors) or adopting certain kinds of pedagogies (e.g., more contact time) result in higher levels of certain kinds of learning outcomes (e.g., critical thinking, or the ability to design a house or strip down a car engine), and that those outcomes in fact generate the final outcomes deemed important (higher employment rates, higher incomes, happier graduates, better citizens), then we can say that a system (or institution) characterized by more of the inputs in question than another system (or institution) is of higher quality.

Of course this entire approach is predicated on identifying the “correct” inputs and outputs and being able to measure these. Can what constitutes the “quality” of an educational experience in fact be reduced to a specific set of inputs, however detailed that list may be? Can these inputs then actually be measured in a meaningful way and at reasonable cost? This is where the framework meets the challenges of the actual estimation exercise. But the practical problems that arise in this manner do not make the framework any less useful for at least thinking about the issues at hand and acting as a broad guide to whatever quality measurement exercise might be undertaken. The alternative is to undertake such an exercise without such an analytical framework, which might be thought of as essentially ignoring the problems the framework illuminates.

Alternatively, we could follow a different approach (discussed further below) where we essentially estimate the relationships between institutional indicators and the outputs, omitting any explicitly measured inputs. This approach would allow us to judge the quality of specific institutions or systems in terms of how they take students with a given set of beginning characteristics and generate a set of learning outcomes and subsequent final outcomes without regard to what exactly the institution or system was doing in terms of specific pedagogies or other inputs. (Although, see also the remarks below regarding the use of secondary analysis of such findings to relate them to specific inputs). We could also take a middle route and include institutions in the analysis, but condition on (in the statistical sense) any given set of inputs, such as including total spending and other measures of the total resources available to see how they compare in terms of the outputs they generate with those given levels of resources, thus presumably identifying “quality” in these terms.

The approach adopted in terms of the inclusion of institution-specific indicators versus detailed and essentially institution-neutral input measures will depend on the purposes of the exercise. Some may wish to rank institutions in some overall fashion without regard to how they achieve what they do.
Others may wish to identify specific inputs which seem to be associated with better quality educational experiences and, hence, perhaps encourage the system, or specific institutions, to adopt the better inputs so long as they are cost effective. There may be many other purposes to which this type of analysis can be put; the value of the conceptual framework offered here is that it can help us understand what we want to measure and the empirical exercise that would need to be undertaken to best achieve the desired measurement of quality.

These variants show how the conceptual framework may be applied in a general manner in order to help us understand what we want to do, what we should do, and what we can do in terms of measuring quality in various specific circumstances – depending on the particular exercise we have in mind. As another illustration of the general applicability of the framework, the relationships could be estimated at different system levels. We might imagine that the college-trade/vocational system has one general set of relationships, the university system another. Or we might think that different “kinds” of universities have different sets of relationships; for example, larger universities may be more research-oriented while smaller ones may focus more on teaching. Or, alternatively, that the different kinds of institutions should simply be taken account of by being factored into any general analysis (e.g., allowing for certain institution-specific effects). The framework can incorporate any of these considerations. A different set of situations (e.g., different institution types) might, for example, imply a different set of parameters, and quite possibly different inputs and outputs – and different rankings to the degree the exercise is focussed in that direction. The pertinent aims of any specific exercise need only be defined, and worked into the framework.

2.3 Thinking in Modelling Terms

In more formal statistical modelling terms, this conceptual framework can be thought of as a set of functional relationships, or a structural model, where the basic goal is to obtain empirical estimates of the relationships in question in order to understand which factors affect which outcomes to what degree (i.e., our desired measure of quality). This kind of approach is shown more specifically in schematic included below as Figure 1.
The beginning characteristics, represented by the “$B_i$” term (the $i$ subscript indicating there are as many of these characteristics as we wish – and are able – to measure), affect the learning outputs ($Z_i$), which in turn affect the final outcomes ($Y_i$), while the $B_i$ also affect the $Y_i$ directly (e.g., the effects of native intelligence on final outcomes will operate both through the learning outputs and directly). The inputs to the educational system ($X_i$) affect learning outcomes and, consequently, final outcomes. (As noted above, we can also allow for direct effects but this represents an unnecessary complication at this point).

The specific relationships in question are represented by the Greek letters shown in the appendix schematic, each letter corresponding to the indicated relationships among each of the different sets of variables. The relationships between the various educational inputs ($X_i$) and learning outcomes ($Z_i$) are, for example, represented by the $\alpha$ term, the subscript $i$ here representing the number of parameters that represent the influences of the inputs on the learning outputs, where each input will at least potentially have an effect on each outcome, and there may be further interactions between different sets of inputs (e.g., class size might interact with professor quality). The same is true for the $\beta$, $\gamma$ and $\delta$ terms. $F_i$, of course, represents all the “other factors” which may influence a result but which lie outside the model.

The inputs which have the greatest (positive) effects on the learning outputs will be identified as those which are most important to the “quality” of the educational experience being measured (or its component parts), while the learning outcomes that have the greatest effect on final outcomes are those that ultimately matter most. Hence, the inputs which ultimately have the greatest effects on the most important learning outcomes might comprise one useful way of thinking about the quality of an educational experience/system. (The reader is again reminded that the richness or pleasure of the education experience for its own sake can easily be incorporated into this framework.)
It is worth noting that the framework implies the identification of a potentially large number of parameters or effects representing how each of the inputs of interest affects each of the learning outcomes and then final outcomes, or the combined effects of these two. Quality measurement is thus potentially a detailed and complex undertaking. But this is unavoidable, and simply represents the reality that quality comes in many dimensions and can have diverse effects. From these detailed results, summary indexes could be constructed to compare institutions in an overall fashion, or to evaluate progress over time, but these would by their very nature be somewhat ad hoc (e.g., how should we compare the effects of an input on one kind of final outcome versus another?) and should be used even in the best of circumstances with advisement.

While this model approach might seem like unnecessarily sophistication (or complication) to some, it represents a conceptually fairly simple, and – we would argue – very powerful framework for thinking about the measurement of educational quality. Indeed, statistical models do not have to be estimated in any conventional sense, and “data” or “evidence” could come in any form, including purely qualitative analyses as discussed below). What the framework offers is a context into which any analysis can be placed, which is helpful for thinking about the related problems, issues, opportunities, and so on.

For example, this framework nicely illuminates the importance of taking beginning characteristics into account when estimating any of the quality relationships in question. The beginning characteristics of a student – or entire student body if an aggregate-level analysis is being undertaken – are likely to affect both learning outcomes and final outcomes, and be correlated with certain inputs. If the beginning characteristics are not taken into account, the relationships between the measured inputs, learning outputs, and final outcomes will be confounded with these beginning characteristics, leading to “biased” estimates of quality. Looking at employment rates of graduates of different institutions, for instance, will not necessarily represent any sort of useful indicator of the associated differences in the quality of the education offered as much as differences in the characteristics of incoming students. This is a simple, common-sensical idea, but the model makes it especially clear.

The value of the conceptual framework proposed here is thus its utility for identifying in a general, yet precise manner the notion of educational quality and how it can be measured. We can then use the framework to help us understand how quality has been measured to date in Canada and elsewhere, to think about what other approaches might be used, and what data or other research tools could be applied or developed to better identify the parameters of interest.

Note that the framework is effectively goal-neutral, in that it can incorporate any desired outcome or any relevant set of inputs or learning outcomes. Any of these could be defined by the user. So whether “employability” or “citizenship” or “happiness” is the desired final outcome, the model provides a framework for thinking about quality and effectiveness accordingly. If certain kinds of learning outcomes are thought to be the key elements of what the schooling experience is meant to lead to, these too can be placed in this framework. The same is also true for the choice of inputs. In this sense, the model is no more than a broad tool which can help us think about what we want to measure, why we want to measure those things, and how we should set about doing this.
Another way of thinking about this framework is that it allows us to go directly back and forth between concepts and measurement. If, for example, a different set of final outcomes, or learning inputs, or beginning characteristics, is deemed important (at least potentially), these are easily worked into the model, which then points us to the relevant measurement issues. For example, some might want to put more emphasis on taking account of students’ beginning characteristics out of concern that these are correlated with inputs (i.e., better students might go to schools with different resources available or different pedagogical traditions), with learning outputs (they might leave school with more skills at least partly due to their initial characteristics rather than the schooling experiences to which they have been exposed), and with final outcomes (i.e., those with more innate talent will presumably do better in the labour market and in terms of other post-schooling outcomes due at least in part to their early advantages). These concerns are easily incorporated into our conceptual framework, which in turn points to the relevant measurement issues at both theoretical and practical levels.

2.4 Estimating the Relevant Relationships

The empirical challenge is essentially to estimate the relationships identified above; more specifically, the $\alpha_i$ and $\beta_i$ parameters which summarize the relationships between inputs, learning outcomes, and final outcomes – however this might be done.

Many different approaches can be imagined, but it is perhaps useful to begin thinking in terms of an individual-level analysis aimed at identifying the educational inputs that matter most and then evaluating different institutions using these criteria.

This exercise might begin, at least conceptually, with the gathering of data on a sufficiently large and representative sample of graduates (we are ignoring dropouts and the associated issues for the moment), those data including all the relevant final outcomes, learning outcomes, inputs to which the individual was exposed, and their beginning characteristics as they entered the educational system. We could then estimate the parameters of the model, or framework, in question. Most importantly, different individuals would have been exposed to different inputs thanks to the variation in experiences across or within institutions, and this would be the key means of “identifying” the input effects (the $\alpha_i$) in question. Such data should also, in principle, allow us to identify the effects of the learning outcomes on the final outcomes (the $\beta_i$) which, when taken with the $\alpha_i$ would allow us to evaluate the overall impact of different inputs on final outcomes; essentially the identification of what inputs represent educational “quality” in a functional sense. The effects of the beginning characteristics would also be estimated, and might be interesting in themselves, but are presumably not interesting of themselves in any measurement of quality exercise – here they only act as “controls” so that the other parameters in which we are interested are estimated in an unbiased fashion.

In a multivariate regression framework, this set of procedures can be thought of in terms of estimating the structural model that corresponds to the framework that has been proposed above. There would be a set of variables corresponding to each of the learning and final outcomes in which we were interested and for which we had data; for the relevant inputs; and for the beginning characteristics. Note that we would not need any particular number of observations per institution to identify the inputs that matter; the institutions serve, in this framework (see an alternative approach
below), only as a source of variation in the inputs offered, and as long as there were sufficient variation in the input measures among the entire pool of “observations” (i.e., the sample of graduates), the parameters could be identified. Indeed, in a well-specified model in which all inputs, beginning characteristics, and outcomes were perfectly defined and measured, institutional effects would go to zero, since institutions can really be thought of as simply a specific bundling of inputs. As a result, if those inputs are completely captured by the model, there should be no residual “institution effects”. Indeed, the existence of residual institutional effects should make us think about what underlying differences they may reflect (e.g., commitment to teaching, accumulated assets, better quality students, etc.), and work them explicitly into the model if possible.

Assuming that sufficient data were available, one could use the estimates of the parameters of interest to assess institutions according to their characterisation in terms of the relevant input measures: do they have more or less of certain kinds of resources (money, staff, materials, etc.), do they use their resources to generate certain kinds of learning experiences (e.g., the kinds of contact time, class sizes, etc.) more than others, and so on. A similar exercise could, at least conceptually, be carried out at a departmental level, or more broadly at a system level (e.g., one province versus another) or to compare changes over time.

An alternative approach alluded to above would be to make institutions, rather than inputs per se, the focus of the analysis. This could be done by turning the approach just described on its head, and conducting the analysis including (in the limit) only institution identifiers (and presumably beginning characteristics in order to control for those). Such an approach would generate direct measures of the “value added” of each institution without regard to what specific inputs are employed by the various institutions – although the results of such an institution-based analysis could then be inputs in a secondary analysis to try to relate “performance” to inputs. Note that such an approach would, unlike the preceding approach where institutions are simply a source of variation in the (“generic”) inputs in question, require a sufficient number of observations (individuals) per institution to directly identify the institution-specific differences/effects.

As mentioned earlier, an alternative approach would be a sort of middle ground, where a mix of measured inputs and institutions are included in the analysis, thereby estimating the effects of each kind of specified input while also measuring how institutions perform conditional on their measured inputs. The approach used will depend on the goals of the measurement exercise. Another twist on this general approach would be to use aggregated data rather than individual-level data, such as average final outcomes (employment rates, earnings, satisfaction, etc.), average inputs, average learning outcomes, and average beginning characteristics. The econometrics would be slightly different, but the basic nature of the analysis would be the same.
2.5 General Approaches to “Measuring Quality” in the Context of the Framework

What, then, is the state of the art in Canada (and elsewhere) of measuring quality, as surveyed in the previous section, placed in the context of the analytical framework proposed here? Effectively, what our framework points to is to incorporate the strengths of both the rankings/indicators approach and the learning impacts approach, while better taking into account student background characteristics. Objections regarding insensitivity to differences in institutional missions can be overcome simply by controlling for – or specifying a completely different set of – beginning characteristics, inputs, learning outputs, final outcomes and/or the associated relationships for each institution or type of institution.

Let us discuss a little further what is currently going on in Canada in order to help flesh out the framework and its usefulness. A selected list of exercises that are at least ostensibly or implicitly deemed to be related to “quality measurement” could include the following:

- the Maclean’s university issue;
- institutions’ surveys of their graduates;
- the NSSE survey of schooling experiences;
- Statistics Canada’s National Graduates Surveys;
- self-evaluation exercises undertaken by institutions themselves or by external examiners as part of general accreditation exercises.

What we see, in the context of our conceptual framework, is that each of these gets at a relatively limited part of the overall exercise with varying degrees of success. The Maclean’s survey is largely about the measurement of inputs – professor-to-student ratios, class sizes, the number of books in the library, and so on. There are also a few potential controls for beginning characteristics (e.g., average entrance marks of first year students) although these are not identified as such. Instead, the magazine seems to offer these characteristics as additional descriptors of the educational experience (which they are to the degree the educational experience is in fact directly determined by the quality of the student body), rather than as factors to be controlled for in assessing quality. Learning outcomes are scant, as are final outcomes; while the latest version reports the results of a satisfaction survey of graduates, it does not incorporate the data in its rankings. As a result, there is neither a real linking of inputs to learning outcomes or final outcomes nor anything in the way of a proper statistical exercise which attempts to identify the parameters which would be of most interest to us; including, perhaps most importantly, the true “quality” of education offered at different institutions.

Similarly, institutions’ own surveys of their graduates are typically based on selected (non-random) samples, focussed on a few final outcomes (e.g., employment rates), and do nothing to control for beginning characteristics. The NSSE survey of schooling experiences gets at certain sets of inputs that have been found to be correlated with learning outcomes, but generally does not link these directly to actual learning outputs or final outcomes. Statistics Canada’s National Graduates Surveys are built on a solid sample frame (of graduates) and provide a number of interesting final outcomes (employment, earnings, job satisfaction, etc.) but do not relate these to inputs or learning outcomes. The self-evaluation exercises undertaken by institutions themselves or by external examiners as part
of general accreditation exercises, which are in general particularly well developed at the college level, are – at least in the latter case – largely aimed at ensuring that graduates finish their studies with certain skill sets (i.e., learning outcomes) and attempt to link these to inputs (to make sure the programs do their skill-development job), as well as final outcomes (largely in order to monitor the skills that are in demand), but typically do so in a non-quantitative manner and are focussed on the specific situation of each institution.

Thus, each kind of exercise typically carves out a limited part of the overall quality assessment framework and generally undertakes the “estimation” in a less-than-ideal manner in terms of samples, the specific data collected, and so on. While each of these sorts of undertakings has some merit, they are generally limited, partial, and imperfect components of the general quality assessment exercise. In the next section we discuss what might be done to improve on these efforts in both the short- and long-run.

2.6 Summarising the Role of the Framework

The conceptual framework proposed here can help us define the notion of quality education and think about its measurement. While it is framed in a simple “model”, we believe it provides a quite powerful way of thinking about quality issues. This sets the stage for the rest of the paper, in which we review what has been attempted in terms of measuring quality to date in Canada and around the world, and to then think about the approaches that could be adopted here in both the short-term and in a longer time-frame to better achieve the desired goals with respect to quality measurement.
3. Next Steps: Measuring Quality in the Short- and Longer-Run

We have already described in very general terms the types of data that would be required to estimate the relationships needed to define and measure quality as implied by our conceptual framework. In an ideal world, we would have data that linked individuals’ beginning characteristics, inputs (principally institutional variables), learning outputs, and final (or “graduate”) outcomes. The problem, of course, is that no such data currently exist, and even with all the money and good will in the world, even a limited version of such a functioning database containing all these elements would by definition take years to develop. The creation of such a database would, however, be enormously valuable for many purposes beyond the measurement of educational quality and as such should be considered an urgent priority for any government wishing to understand more about the process of education.

We will return to the essential elements of such a longer-run project below. But first we discuss what could be done to further the measurement of quality in the meantime. We propose two courses of action. The first is to take advantage of existing data to estimate as many of the empirical relationships implied by our conceptual framework as possible, and the second is to foster “continuous improvement” measures of the type described above and increasingly in use in European countries.

3.1 Empirical-Type Analyses in the Short-Run

The General Nature of Such Analyses

While a full (longitudinal) dataset which followed individuals overtime time and included all the elements implied by our conceptual framework would constitute the (currently unavailable) ideal, there are many existing data sources which capture particular pieces of the fuller picture and could, if put together in a creative manner, go a long way towards identifying many of the relationships of interest and otherwise advance the cause of measuring education quality in Canada.

One could, for example, start with the excellent properties in terms of sample frame and other statistical properties of Statistics Canada’s National Graduates Surveys (NGS), a large (approximately 35,000 observations in each panel) and representative sample of trade-vocational, college, and university graduates currently undertaken every four years and based on telephone interviews carried out two and five years following graduation. Using this, individual graduates could be measured in terms of their employment outcomes, earnings levels, job satisfaction, and other measures of interest on the file. This would represent a relatively limited and job-oriented set of outcomes (which is perhaps normal for a “general purpose” database such as the NGS), but which could be expanded in future incarnations to include other outcomes of interest.

The NGS also contains a good deal of personal information such as sex, age, language, immigration status, place of residence (current, during school, before school), disability status, current marital status, and so on. In addition, the survey gathers a fair amount of information, either directly from the respondent or through the institutional files of the school attended, relating to the educational
experience such as type of program, major discipline, normal length of program, time taken to completion, participation in a coop program.

While this basic structure and the information directly available on the NGS makes it a good starting point, the institution identifiers would allow much more information to be pulled in to individuals’ records. Such information would represent, in the terms of our model, “beginning characteristics” (potentially at both the individual and institution-average level) and “inputs” and possibly even a limited set of learning outcomes.

With this information, one could run a regression of individuals’ final outcomes on their inputs while controlling for beginning characteristics. Coefficients could thus be obtained for the effects of the different inputs on the final outcomes in question. Those coefficients could then be used to infer the “quality” of the various inputs, which would be interesting of themselves, answering the question “what matters?”. If one so desired, institutions could then be compared according to the degree they possessed the better inputs – thus implying the “quality” of the education offered in the sense implied by all the caveats offered here and above.

Conversely, as discussed earlier, institutions could be evaluated more directly by regressing the final outcomes on institution identifiers alone (although note the earlier comments regarding the sample sizes required for such an exercise) while controlling for beginning characteristics – but, in particular, not including inputs. Or, some combination of inputs and institution identifiers could be used to both identify inputs of importance and assess institutions’ performance conditional on whatever inputs were included.

Alternatively, estimation could be based entirely on aggregate data (i.e., institutions’ averages), rather than the micro-data nature of the NGS data. By such an approach, average final outcomes would be regressed on average beginning characteristics and average inputs. As discussed in Section 2, the econometrics would change a bit, but the fundamental properties of the exercise (conceptually and statistically), would remain the same.

Finally, even if the entire model could not be estimated (or different elements were estimated with varying degrees of completeness), certain specific pieces could potentially be estimated. For example, if the NSSE data could be linked to some sort of final outcome measure (perhaps as part of a separate estimation exercise) at either the individual or aggregate level, this information could help us identify how inputs had their effects on final outcomes through the learning outcomes in question and thus provide us with one of the important specific elements of any overall “quality assessment” exercise.

We cannot place too much weight, however, on saying that these are only a few examples of the sort of exercise that could be carried out. The actual undertakings would presumably depend on the objectives of the exercise and the precise data available. All we wish to communicate here is that potentially interesting options are available even in the short run.

**Potential Data Sources for Short-Run Analyses**
As just described, because no full database exists, short-run analyses will likely involve the cobbling-together of several different data sources. Potential candidates include a number of
repositories of institutional data on students and finances; some of this is already held by Statistics Canada while the rest is held at the institutional level. Other data are available from surveys such as the Canadian Undergraduate and College Survey Consortium surveys, NSSE, and other sources, although such data is obviously only available at those institutions which participate in the surveys. A working list of the data elements that might be required for quality measurement exercises and the potential data sources to meet them are shown below in Table 5. We now discuss each of the major kinds of elements – beginning characteristics, inputs, learning outcomes, and final outcomes in turn.

**Table 5 – Data Elements and Potential Sources of Aggregate Data for Institutions**

<table>
<thead>
<tr>
<th>Beginning Characteristics</th>
<th>Aggregate Data for Institutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Admin data, CUSC, CCSC</td>
</tr>
<tr>
<td>Gender</td>
<td>Admin data, CUSC, CCSC</td>
</tr>
<tr>
<td>Ethnicity/Aboriginal Status</td>
<td>CUSC, CCSC</td>
</tr>
<tr>
<td>Immigration Status</td>
<td>New Survey Required</td>
</tr>
<tr>
<td>Mother Tongue/Ethnicity</td>
<td>New Survey Required</td>
</tr>
<tr>
<td>Primary Language</td>
<td>Admin data</td>
</tr>
<tr>
<td>Family Type</td>
<td>New Survey Required</td>
</tr>
<tr>
<td>Family Size</td>
<td>Admin data, CUSC, CCSC</td>
</tr>
<tr>
<td>Location of High School (Urban/Rural)</td>
<td>CUSC, CCSC</td>
</tr>
<tr>
<td>Disability Indicator</td>
<td>CUSC, CCSC</td>
</tr>
<tr>
<td>Aboriginal Indicator</td>
<td>CUSC, CCSC</td>
</tr>
<tr>
<td>Presence of children</td>
<td>CUSC, CCSC</td>
</tr>
<tr>
<td>Financial Savings</td>
<td>CUSC, CCSC</td>
</tr>
<tr>
<td>Student Income</td>
<td>CUSC, CCSC</td>
</tr>
<tr>
<td>Secondary School Marks</td>
<td>Admin data</td>
</tr>
<tr>
<td>Secondary School Literacy Scores</td>
<td>New Survey Required</td>
</tr>
<tr>
<td>Skill test scores at university entrance</td>
<td>New survey required (e.g., CLA)</td>
</tr>
<tr>
<td>Emotional Quotient scores at entrance</td>
<td>New survey required</td>
</tr>
<tr>
<td>Purpose for attending PSE</td>
<td>New Survey required</td>
</tr>
<tr>
<td><strong>Inputs</strong></td>
<td></td>
</tr>
<tr>
<td>$/student</td>
<td>SFIUC + Admin</td>
</tr>
<tr>
<td>$/student in salaries</td>
<td>SFIUC + Admin</td>
</tr>
<tr>
<td>$/student in libraries</td>
<td>SFIUC + Admin</td>
</tr>
<tr>
<td>$/student in IT</td>
<td>SFIUC + Admin</td>
</tr>
<tr>
<td>$/student in student services</td>
<td>SFIUC + Admin data</td>
</tr>
<tr>
<td>$/student in student aid</td>
<td>SFIUC, OSAP + Admin data</td>
</tr>
<tr>
<td>Governance indicators</td>
<td>New Metric Required</td>
</tr>
<tr>
<td>Physical infrastructure indicators</td>
<td>New Metric Required</td>
</tr>
<tr>
<td>“Learning Environment” indicators</td>
<td>New Metric Required, possibly NSSE</td>
</tr>
<tr>
<td>Student-staff ratio</td>
<td>SFIUC + Admin data</td>
</tr>
<tr>
<td><strong>Learning Outputs</strong></td>
<td>Aggregate Data for Institutions</td>
</tr>
<tr>
<td>Degree/diploma</td>
<td>Admin data</td>
</tr>
<tr>
<td>Time-to-completion</td>
<td>CUSC</td>
</tr>
<tr>
<td>Generic work skills</td>
<td>New Metric Required, possibly NSSE or CLA</td>
</tr>
<tr>
<td>Quantitative Literacy</td>
<td>New Metric Required, possibly NSSE or CLA</td>
</tr>
<tr>
<td>Prose Literacy</td>
<td>New Metric Required, possibly NSSE or CLA</td>
</tr>
<tr>
<td>Problem solving</td>
<td>New Metric Required, possibly NSSE or CLA</td>
</tr>
<tr>
<td>Potential Final Outcomes</td>
<td>Aggregate Data for Institutions</td>
</tr>
<tr>
<td>-------------------------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>Employment</td>
<td>Institutional Surveys/NGS</td>
</tr>
<tr>
<td>Earnings</td>
<td>Institutional Surveys/NGS</td>
</tr>
<tr>
<td>Satisfaction with Education</td>
<td>Institutional Surveys/NGS</td>
</tr>
<tr>
<td>Civic Engagement indicators</td>
<td>New Metric Required</td>
</tr>
<tr>
<td>“Happiness”/Life satisfaction indicators</td>
<td>New Metric Required</td>
</tr>
<tr>
<td>Lifelong learning</td>
<td>New Metric Required</td>
</tr>
</tbody>
</table>

Admin = Institutional administrative date; CUSC = Canadian Undergraduate Survey Consortium; CCSC = Canadian College Survey Consortium; SFIUC – Survey of Financial Information of Universities and Colleges; OSAP = Ontario Student Assistance Program; NGS = National Graduates Survey; CLA = Collegiate Learning Assessment.

Administrative datasets which capture students’ **beginning characteristics** at the point of entry to their programs would represent a potentially rich source of beginning characteristics – a single variable such as high school grade average would, for example, go a significant way in helping control for the “ability” of incoming students. Some of this is currently held centrally by Statistics Canada; much more is held at each institution. Other potential institutional/administrative-type data include (general) application files and tax files (e.g., individuals could be linked to their family, or at least neighbourhood, of origin).

No multi-institutional surveys currently exist to measure some of the “beginning characteristics” defined here (e.g., incoming students’ “skills”, “emotional quotients”, the purpose for attending PSE, and other factors deemed important to individuals’ learning and final outcomes), and as a result a new survey would need to be constructed in order to capture these characteristics. Work along these lines should be undertaken even if no longitudinal database is ever undertaken; more knowledge of the impact of these characteristics practices could, of course, potentially allow us to control for beginning characteristics at the individual level rather than use the measures to impute an average quality of incoming student as would be the purpose in the short run. To be clear, the degree such information could be linked to individuals’ records on an individual level as opposed to institutions’ average values (i.e., for the “NGS-based”-type analysis), such analyses would be further enhanced.

Many “raw” **input measures** – including those used by Maclean’s – could be obtained from a reworking of data held internally by institutions or given to Statistics Canada through its annual financial surveys. Further measures could presumably be obtained if institutions would agree to cooperate. Work would have to be done to weight institutional expenditures appropriately given their different sets of program offerings; this is necessary in order to ensure that $/student ratios in fact reflect the underlying cost structure of different institutions. Separating teaching and research functions, and spending, would represent a general issue to be faced. Statistics Canada’s ESIS/USIS/CSIS (administrative) files might also be useful here for capturing individuals’ (and institutions’ average) “education experiences”.

With respect to the **arrangement** of inputs, or “pedagogies” (as opposed to simply the amount of inputs), NSSE already gives excellent data about things such as frequency of staff-student contact...
and could be an extremely valuable source for these sorts of variables. Unfortunately, at present NSSE is neither in very broad use nor are its data publicly available. Broader acceptance of NSSE and greater openness regarding results would be required to make analyses that include this type of input measurement meaningful. (Creating a Canadian-ised version of CSSE would be useful in order to extend this model to community colleges.)

Learning outputs represent the area where perhaps the least information is currently available. Many “institution performance” indicators (e.g., graduation rates, time-to-completion) are already available at least in principle, and could be part of any “learning outputs” analysis. However, measures of true “learning outputs” as defined here – including both “hard” and “soft” skills – are not generally available, and would therefore presumably require new surveys (comparable to the Collegiate Learning Assessment in the U.S.), or even exams to the degree these were deemed suitable for capturing the relevant information and were otherwise feasible.

To the extent that the final outcomes one wished to study were related to employment, job satisfaction, and the like, the National Graduates Survey represents an excellent starting point. To the extent that one wished to study “softer” outcomes, such as levels of civic engagement or general happiness, a new survey – or new questions added to the NGS – would be required. Other methods of obtaining information on graduates would come up against significant sample frame problems in a context where the importance of obtaining data for representative samples cannot be overstated.

Having certain elements required by any complete quality analysis available in disparate places will of course not necessarily resolve all the related empirical needs, because ultimately the data on beginning characteristics, inputs, learning outcomes, and final outcomes need to be pieced into a single coherent analysis. But we can at least begin to think about using the various data sources currently available to carve off specific pieces of the overall story, to estimate the more complete model in a restricted manner, to consider what data improvements could be made in the short-run, and to ultimately help us move to longer-run solutions.

For example, the NSSE might help us identify certain “inputs” (using our definition – in the case of NSSE referring specifically to different kinds of pedagogies) associated with different institutions, while graduate surveys (including Statistics Canada’s National Graduates Survey) could then link final outcomes to those inputs. Perhaps administrative data (such as application data) could then be used to help control for beginning characteristics in any such exercise. Giving standardised exams at the point of entry into a program/institution or at finishing could help us identify beginning characteristics and at least some elements of students’ final learning outcomes. Linking each of these elements together could then help us identify the effects of different kinds of inputs on learning outcomes and final outcomes while controlling for beginning characteristics. In the absence of perfectly longitudinal data and full information on each of the kinds of variables required for a full and complete analysis, undertaking such exercises might represent useful steps in the short-run.

3.2 Short-term Measure No. 2 – Quality Self-Monitoring

In addition to embarking on the sort of quantitative approaches just described, it would also seem worthwhile to adopt – in the short-run but perhaps on a more permanent basis – in a more comprehensive manner the “continual improvement” model described above. Such procedures could be undertaken in parallel to, or possibly simply replace, the existing systems of cyclical reviews and...
other such evaluation procedures. Such an approach would, interestingly, be keeping in line with the current major thrust of what comes closest to comprising “quality measurement” procedures in most European countries.

With such an approach, accountability and transparency might be improved by adopting the practices of some other countries by inviting governments or government agencies into the process and making the results public. However, it is not clear that a public document would be read (or understood) or that government participation would even be welcomed in the Canadian context. In any event, the basic elements of this approach would be to require institutions to:

- develop a set of metrics to measure “institutional quality”;
- develop a set of procedures to improve quality based on the data provided;
- periodically publish data on institutional quality.

A more certain role for government, especially in the short-term, might be to provide technical assistance to institutions in developing their own quality procedures measures. Over the long-term, government (or an agency thereof) might continue to perform an oversight function, but probably one with a light touch, such as ensuring that institutional metrics were appropriate and that quality improvement procedures were in fact being followed. It would not, specifically, make any comment as to the relative quality of different institutions – as this is not in the spirit of these sorts of practices. In an ideal situation, government intervention would be almost unnecessary as universities and colleges performed their own quality assurance procedures and became recognised for doing so – while of course also benefiting from the quality enhancements that should generally emerge from such practices.

### 3.3 Longer-Run Measures

The longer-run measures we suggest are, in broad terms, simply those summarised above for developing data that would better facilitate the sort of quantitative analyses implied by our full conceptual framework. We have described above the general nature those data would ideally take (longitudinal, possessing full sets of information on beginning characteristics, inputs, learning outcomes, and final outcomes), and have also discussed how certain elements of these kinds of data are available already or could be obtained in a time frame shorter than that which concerns us here. But the principles are basically the same, and need not be repeated.

What is perhaps worth mentioning are some speculations on how such a database could – and should – perhaps be constructed.

First, there is probably little doubt but that the task should be undertaken by, and centred at, Statistics Canada. They have not only the technical expertise and the established reputation for building high quality datasets, but also have access to many of the ancillary datasets that could be used in the construction of the overall “quality measurement database” we have in mind, or would be the natural custodian for any others that might be collected.

Second, the database should probably build, as much as possible, on existing databases and instruments – thus saving resources and taking advantage of proven and known instruments. These
could include more obvious candidates such as the YITS, the National Graduates Survey, a plethora of institution-based measures, and so on. But it could also pull from less currently well-known datasets such as the “ESIS”, which consists of full institution-based administrative records of virtually all Canadian post-secondary students. While the full potential of this dataset remains uncertain, principally due to its under-utilisation to date, it could be a trove of useful data which could in turn possibly be linked to the other data used in the construction of the quality measurement file. Post-secondary application records represent another such example. Other potential candidates should be considered.

Third, there are already a great deal of resources going into data gathering and the construction of measures of various kinds. As a single example, individual institutions alone spend significant sums in attempting to follow up with their graduates. Yet despite these expenditures, the results are doomed to disappoint, as these bodies do not have the resources to do a truly excellent job of designing questionnaires, establishing proper sampling frames, following through on non-respondents, and so on. If these efforts and resources were better coordinated, there is little doubt that a much better job could be done for less money, with obvious benefits for all.

Fourth, the creation of such a new database would present a great opportunity to be bold and creative. A survey should be conducted of what has been done the world over, and we should build on that to best serve our particular needs using the potential resources available (including our rather uniquely rich YITS, NGS, and other existing databases).

Finally, additional data developments undertaken for such a project would likely have other benefits as well. Measuring educational quality would be but one use for such an extensive dataset, but once put together, the file would without question be of significant value to many other users, many of whom could probably not even be identified in advance. In short, such a file would likely be characterised by a strong “public good” element.
Conclusion

We began this paper by noting that while all educational stakeholders believe in improving the quality of post-secondary education, discussions about how to measure quality have become problematic due to mutual suspicions between governments and institutions over the quite separate issue of accountability. In our global tour of approaches to quality measurement, we have also seen how poorly quantitative measurement styles have often fared at extracting useful, contextual information and how equally poorly most qualitative measurements have fared both in terms of communicating educational success and failures to a wider public and in capturing the specific challenges facing each educational institution. Our challenge, therefore, was two-fold. The first was to find a means of measuring quality in a way that would permit a de-linking of the topic from that of the important but separate issue of the accountability relationship between post-secondary education institutions and government. The second was to find a means of examining quality that combined both the clarity and rigour of certain types of quantitative analysis while still controlling for the many important differences in institutional resources, missions and student bodies. We met this challenge by providing a structural model of educational quality which, while highly quantitative in nature, permits many equally acceptable “outcomes measures” for post-secondary education, and controls for many differences in student characteristics, financial inputs, the nature of the educational experiences, and various forms of learning outcomes.

Some readers may have approached this paper looking for simple answers to some basic questions, such as: “What is the definition of quality post-secondary education?” “How can we measure it?” “How can we rank or evaluate institutions or systems according to their quality?” “How will we recognise an improvement or deterioration in quality when we see it?” “How can we tell if resources are being well spent?” “How do we know if students are getting the education they should have?” We have, unfortunately, answered none of these questions directly and as a result we expect that some readers may find this paper to be frustrating and disappointing.

What we have done here is offered a general framework for thinking about these questions and both some general and specific suggestions for how the required empirical/practical exercises might be carried out in the shorter- and longer-run. Using these suggestions to actually answer the questions posed above is left to the reader, and the job won’t necessarily be simple, or easy. We would argue, though, that this is the nature of the beast. There is no magic solution or silver bullet in measuring educational quality. An appreciation of this point is probably a good starting and ending point for this paper.

But we would suggest that our conceptual model is in fact rather simple in its fundamental nature: for our purposes, education experiences can be defined in terms of the relevant “inputs” (resources, pedagogies, etc.) to which a student is exposed, and it is the effects of the different kinds of inputs in terms of how they affect first learning outcomes and then final outcomes, while controlling for individuals’ beginning characteristics, in which we are interested when we want to measure “quality”. Because once those inputs are identified and quantified, an institution or system can be evaluated in terms of the “quality” inputs it possesses. Or, in a related but different manner, we can assess institution-specific quality more directly by estimating the learning outcomes and final outcomes it generates (again controlling for beginning characteristics) conditioning on the resources
available to the institution (or otherwise controlling for given sets of inputs) to the degree we wish, thus generating a “conditional” notion of quality by which institutions can be (carefully) ranked. Or we can estimate “institution effects” and relate these to inputs, thus identifying quality in a reverse fashion than the first.

In any event, to be put into practice, the details need to be specified. What are the ultimate goals of the education system in terms of the outcomes of graduates to which their schooling should lead? What are the relevant “inputs” of the system? What are the learning outcomes that might link inputs to final outcomes? What are the sorts of beginning characteristics which need to be controlled for? What data are available to carry out such an exercise? A workable definition of “quality” depends upon the existence of a set of answers to these questions that commands a reasonable consensus from all higher education stakeholders. This, unfortunately, is sorely lacking in Canadian post-secondary education. If our paper helps in framing these issues and ultimately leading to exercises which answer some of those simple questions posed a couple of paragraphs back, it will have done its job. We hope it represents a step forward in this respect.
Appendix A – Measuring Institutional Research Quality

The research function at universities has always had a much stricter quality regime than the teaching function. Decisions on hiring, promotion and tenure of academic faculty are in no small measure determined by the number of publications one has to one’s name. Since all academic articles are subject to peer review (i.e., typically an assessment by at least two referees selected for their expertise in the subject matter at hand), one can say with some confidence that there is a tremendous amount of work that goes into assuring the quality (if not the quantity) of the research work being performed by academic staff.

Quantitative evaluation of research at institutions as a whole has a much shorter pedigree than the evaluation of individual pieces of research. In fact, serious efforts to measure university research are really only about a decade old. In order to look at university research indexes, we looked at eight different surveys either partially or entirely about research effectiveness: Alberta’s Key Performance Indicators40, Ohio State University’s Academic Scorecard41; Research indicators for the states of Illinois42 and Texas43, the Top American Research Universities Survey (produced annually by The Center at the University of Florida)44, the Deutsche Forschungsgemeinschaft (German Research Foundation)’s Funding Rankings45, the Times Higher Education Supplement’s (THES) World University Rankings46, the Academic Ranking of World Universities47 produced by the Institute for Higher Education at Shanghai Jiao Tong University48. An examination of these different sources shows that most measures of university research quality focus around some combination of the following five sets of indicators:

1) **Public Research Dollars Received.** Virtually all OECD countries have some form of agency that distributes public money for research according to a competitive, peer-reviewed process. An institution’s scholars’ ability to attract this kind of competitive money is used by all the research assessment exercises surveyed, with the exception of the two sets of world university rankings. In some cases, the raw total of dollars received is used as a measure; more commonly, the figure is expressed as dollars per faculty member or, in one instance, dollars per student. Often, controls for “type of institution” are used to distinguish between the performance of institutions based on breadth of program offerings at the graduate and undergraduate level.

---

41 From OSU’s Academic Plan; available at www.osu.edu/academicplan/scorecard.html.  
46 See World University Rankings, Times Higher Education Supplement, 5 November 2004 at www.thes.co.uk/worldrankings/.  
48 These last two are technically not research rankings, but they do contain elements which implicitly rank universities either by the quality of their research or their researchers.
2) Private Research Dollars Received. Most research rankings also stress the importance of attracting private research dollars. This can be expressed in a number of ways: raw dollars and dollars per faculty member are common; but it is also common to see private dollars expressed as a percentage of public dollars, indicating that the creators of many research indicators consider the balance of public-private to be as important as the actual amount of private funds attracted.

3) Publications. Some type of bibliometric analysis is often included in university research rankings, especially those with an international bent. This is probably the most controversial of the various measures of research quality as most bibliographic measures have certain inherent biases towards the sciences. The sciences, for instance have certain “gold standard” publications (i.e., *Science* and *Nature*) that the humanities and social sciences lack; conversely, the social sciences and humanities have spawned a profusion of journals that are of highly uneven quality. Social scientists are also much more likely to write books than scientists, and the contribution of these longer monographs is missed in bibliometric measurements. Finally, given the ubiquity of English as the language of Science, institutions with non-anglophone staff are seen as being at a competitive disadvantage in bibliometric analyses.

4) Faculty Awards. Another common measure of research “quality” is the number of prominent awards received by faculty members. The definition of what awards are significant enough to warrant inclusion varies somewhat from place to place. One rankings system only uses Nobel prizes, though most are considerably more inclusive. Canadian institutions, when they wish to demonstrate their research prowess, often cite the number of Killam Fellowships received by faculty as an indicator of faculty quality.

5) Technology Transfer. Some jurisdictions – notably U.S. states – use measures of technology transfer as a way to measure the “relevance” or “impact” of a university’s research effort. Common indicators here include numbers of patent applications and patent awards, number of “inventions”, number of spin-off companies produced and various measures of revenue from licenses, royalties, spin-off companies, etc.

The Top American Research Universities also uses two other measures of research intensiveness. The first relates to the quality of undergraduate students, as measured by the median entering GPA, while the second deals with the number of graduate students and postdoctoral students at each university. Technically, neither of these are measures of research quality per se – however, they are likely correlates of academic prestige and size, respectively, both of which are arguably important in determining which institution is “top”.

The Deutsche Forschungsgemeinschaft (DFG) also uses certain other measures not seen elsewhere. The German ranking system also puts a certain emphasis on research being performed in a networked manner and therefore attaches considerable attention to mapping research networks. Institutions that appear “central” to research networks receive higher scores than those that do not. The German government is also concerned about internationalization of research, and therefore institutions are ranked on the number of Alexander von Humboldt and German Academic Exchange Service (Deutsche Akademischer Austauschdienst, or DAAD) scholars, the number of DAAD

49 For an example see www.dfg.de/en/ranking/download/dfg_funding_ranking_2003.pdf
graduates students, and the amount of money received from EU research funds. The THES’ World University Rankings also looks at internationalization of faculty.

### Table 1A – Main Research Quality Indicators

<table>
<thead>
<tr>
<th></th>
<th>Public Research Dollars</th>
<th>Private Research Dollars</th>
<th>Publications/Bibliometrics</th>
<th>Faculty Awards</th>
<th>Technology Transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alberta KPIs</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ohio State</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Illinois(^{51,52})</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Texas</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>America’s Top Research Universities</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>THES’ World Rankings(^53)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Academic Rankings of World Universities (Shanghai)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>DFG (Germany)</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Generally speaking, there appears to be less controversy about quality measurement in research than there is about quality measurement in undergraduate education. In part, this is likely because there is greater agreement as to what constitutes good “final outcomes”. It likely also stems from the fact that the empirical foundations for research indicators are considerably stronger than those for undergraduate education. The two most notable differences are that none of the research indicators listed above require the deployment of any survey instruments (a stark contrast to undergraduate education), and that most of the above information can be obtained from third-party sources.

---

\(^{50}\) Foreign researchers who already have their PhD may apply for an Alexander von Humboldt Foundation to study at a particular university; Students who wish to study in Germany may apply to the DAAD for funding. In both cases, the application is competitive and the researcher or student chooses at which institution to study.

\(^{51}\) Technically, Illinois measures change in research dollars on a year-to-year basis rather than the amount of research dollars themselves.

\(^{52}\) Neither Illinois nor the DFG look at private research dollars directly; instead, they look at public dollars and total dollars. (which includes private funding).

\(^{53}\) The THES and DFG do not look at faculty awards, but they do look at % of faculty who are international – i.e., are originally from other countries.
Bibliography


Our Support

Funding for this project was provided by:

- The Wilson Foundation
- Ontario Ministry of Training Colleges and Universities (Post-secondary Review Secretariat)
- Power Corporation of Canada
- First Plazas Inc.

Donations:
BCE Inc.
BMO Financial Group
COGECO Inc.
Maclab Enterprises
Ontario Municipal Employees Retirement System
Power Corporation of Canada
Scotiabank

Members of the Board of Directors, Campaign Committee and management team
Many e-network subscribers and friends of CPRN

Project Funding:

Corporations:
Bell Canada
Business Development Bank of Canada
CIBC
DVA Navion
Ekos Research Associates Inc.
Holding O.C.B. Inc.
Home Depot Canada
Ketchum Canada Inc.
RBC Financial Group
SNC-Lavalin Group Inc.
TD Bank Financial Group

Federal Government Departments,
Agencies and Commissions:
Canada Mortgage and Housing Corporation
Canadian Heritage
Citizenship and Immigration Canada
Health Canada
Human Resources Development Canada
Human Resources Skills Development Canada
Indian and Northern Affairs Canada
Industry Canada
International Development Research Centre
Law Commission of Canada
Parks Canada
Privy Council Office
Social Development Canada
Statistics Canada
Treasury Board of Canada, Secretariat

**Provincial Governments:**

**Alberta**
- Human Resources and Employment

**British Columbia**
- Ministries of Health
- Ministry of Children and Family Development
- Office of the Deputy Minister to the Premier

**Manitoba**
- Department of Family Services and Housing

**Ontario**
- Association of Colleges of Applied Arts and Technology of Ontario (ACAATO)
- Ministry of Community and Social Services
- Ministry of Finance Pre-budget Consultation
- Ministry of Health
- Ministry of Training, College and Universities (MCTU)

**Saskatchewan**
- Department of Community Resources and Employment

**Municipal Governments:**

City of Toronto

**Foundations:**

The Atkinson Charitable Foundation
The Bertelsmann Foundation
Canada Millennium Scholarship Foundation
Canadian Health Services Research Foundation
The Change Foundation
The Hospital for Sick Children Foundation
R. Howard Webster Foundation
J. W. McConnell Family Foundation
The Muttart Foundation
The Neptis Foundation
RBC Foundation
Associations and Other Organizations:
AFP Foundation for Philanthropy – Canada
AFP Calgary Chapter
AFP International
AFP Toronto Chapter
Association of Fundraising Professionals (AFP)
Canadian Institute of Planners
Canadian Labour and Business Centre
Canadian Labour Congress
Canadian Medical Association
Canadian Population Health Initiative
Canadian Public Health Association
Coalition of National Voluntary Organizations
College of Physicians and Surgeons of Ontario
Conference Board of Canada
The Learning Partnership
National Voluntary Organizations
Nuclear Waste Management Organization
United Way of Canada
Université de Montréal
University of Toronto (Faculty of Law)

Other:
Liberal Caucus Research Bureau