Higher Education Productivity Metrics

An essential guide for users and developers

First edition published in Japan
by the Asian Productivity Organization
1-24-1 Hongo, Bunkyo-ku
Tokyo 113-0033, Japan
www.apo-tokyo.org

Contents

ACKNOWLEDGMENTS

The APO is grateful to Professor Emeritus and former Vice President for Business and Finance, William F. Massy, Stanford University, who was responsible for writing this guide titled *Higher Education Productivity Metrics.* The guide was prepared as a result of the recommendations made by participants in the Conference on Raising Productivity in Higher Education held 29 August–1 September 2016 in Jakarta, Indonesia.

The APO wishes to thank the conference participants for their advice on helping higher educational institutions analyze and understand their own productivity performance using the measurements and formulas explained in this guide.

INTRODUCTION

The importance of productivity metrics in higher education is hard to dispute. They are essential for priority setting, quality management and cost control: as goes the saying, "If you can't measure something you can't improve it." They are important for accountability: both within colleges and universities and for external agencies. Finally, they are essential for public and political understanding of the higher education sector and the institutions within it. A large number of disparate metrics have been proposed, and many of them are in use at the present time.

The plethora of disparate metrics has generated considerable confusion. This point was brought home during the Asian Productivity Organization (APO) conference on productivity in higher education (2016). At the conference, teams from nineteen countries reported on APO-sponsored projects to develop measures of university productivity for use in their countries. Much good work was reported, but there was no consensus about what to measure or even the concepts underlying the various alternatives. This Guide grew out of a mutual desire for common concepts and vocabulary to inform future work by the APO and elsewhere in the world.

The confusion about productivity can lead to suboptimal decision-making by universities and oversight bodies, sometimes pushing institutions and their faculty in regrettable directions. Perhaps worst of all, the confusion has inhibited the kinds of coherent conversation needed for shared understanding. Institutions and stakeholders select their favorite metrics, often based on a combination of hunch, convenience, and ideology, and then defend them tenaciously.

This Guide presents a paradigm for organizing the different kinds of productivity metrics, applies it to a representative sample of the metrics currently in use, and suggests some criteria for gauging the efficacy of particular metrics. Hopefully, this will help unpack the current tangle of productivity metrics in higher education and, also, prove useful in starting a conversation about the way forward.

THE MULTIPLE MEANINGS OF "PRODUCTIVITY"

The new paradigm reflects the fact that not one, but two of the dictionary definitions for "Productivity" must be taken into account. The two definitions can be paraphrased using the fanciful example of a company that manufactures "widgets."

1. The quality or fact of being productive; the capacity to produce. e.g., "High-quality widgets are valuable and the company produces lots of them."

2. The physical quantity of output per unit of useful effort. e.g., "The company produces its high-quality widgets with the least possible expenditure of resources."

Colleges don't manufacture widgets, of course, but exactly the same principles apply.

It appears that the first definition entered common usage earlier than the second, which probably accounts for the fact that many university faculty describe "productivity" as producing lots of good works [1]. The second, generally newer, definition is what economists and most oversight bodies mean when they speak of productivity. They focus on the processes of production, not simply the amount of output produced. Both definitions are important, and both are widely used in higher education.

This Guide considers metrics for the academic outputs, produced mainly by faculty, that are utilized by students, research sponsors, and the other non-employee stakeholders of colleges and universities. Basically, these are education and research. (Administrative and support services also produce outputs, but they mostly are used internally.) Examples of educational outputs include student learning and degree attainment, as well as many other benefits associated with a college experience. Employers realize value from well-educated students, and the nation and world benefit from a well-educated citizenry. Research results push back the frontiers of knowledge and, hopefully, lead to solutions for current problems. Universities produce a third set of outputs, which is public service (e.g., cultural and sporting experiences, facilitation of civic advancement, and many more), but they are too varied to consider here.

Figure 1 presents a paradigm for analyzing productivity in higher education. The "Output" variables reflect the first productivity definition given above, and those for "Outputs Relative to Inputs" reflect the second. The "Financial Returns" variables reflect the degree to which the outputs are "productive" in the marketplace. Outputs that meet the third criterion generate revenues that exceed their costs (positive direct margins). Financial returns are not included in the dictionary definition but it completes the collection of things universities should consider when addressing questions of productivity.

OUTPUT (value-creation productivity)

- Quantity Produced (amount of the valued output made available to users)
- Design Quality (as specified for inputs & processes)
- Implementation Quality (as actually delivered to users)

OUTPUT RELATIVE TO INPUT (economists' productivity)

- Process Efficiency (optimizing the current production function)
- Process Innovation (improving the production function)

FINANCIAL RETURNS (marketplace productivity)

- Earned Revenue (direct, derived from the outputs' market prices)
- Production Cost (direct, at the process level; plus overhead allocations)
- Margin (gross, direct revenue minus direct cost; net, after overheads)

Figure 1. Productivity Paradigm for Colleges and Universities

The following paragraphs further define the variables in Figure 1. Most of these concepts are described in Massy [2].

Output (value-creation productivity)

Universities are well regarded because they create value for students and other external stakeholders. The values usually include "intrinsic" and "extrinsic" elements: the former deriving from the school's education and research mission, and the latter from decisions by customers, etc., in the marketplace. This corresponds to the first dictionary definition of productivity.

- *Quantity Produced* refers to the amount of valued output made available by the university. As noted, the output bundle reflects a balancing of mission and market-related priorities. The balance often varies dramatically across fields, degree levels, etc., and may be difficult for outsiders to understand.

- *Design Quality* refers to the outputs' intended attributes: "specifications" for short. For example, schools like Stanford University seek to provide education at a certain level for certain types of students, and have the capacity in terms of faculty, staff, facilities, and money to do so. Other schools may have different aspirations and capacities, and set the specifications for their programs accordingly. Metrics that describe these differences (e.g., those used in college ranking and some accreditation systems) are not difficult to develop, but no one metric is universally applicable. Users need to decide what kinds of attributes they desire in relation to their own goals, capacities, and resources.

- *Implementation Quality* refers to how the design specifications are reduced to practice. For example, a well-designed degree program may be delivered in a lackluster way by people who, though they may be fine scholars, lack teaching skill or motivation, or are overwhelmed by other duties. Maintaining implementation quality requires consistency over time and across students: e.g., good teaching in every encounter with students as opposed to periodic flashes of brilliance interspersed with mediocrity.

  Maintaining implementation quality requires a vigorous ongoing program of quality assurance. Many universities work hard on the assurance of design quality (e.g., through faculty curriculum committees), but they have much to learn about implementation quality [3]. Recent developments in computer-mediated learning, which includes built-in learning assessments, and the use of aggregate outcomes measures like graduation rates are encouraging.

Output Relative to Input (economists' productivity)

Universities operate in environments where resources are scarce. Hence, they must concern themselves with the economists' definition of "productivity," as well as the one centered on value. This corresponds to the second dictionary definition of productivity.

- *Process Efficiency* means exploiting current knowledge about how to produce the product or service with the least possible expenditure of resources, provided, importantly, that neither design nor implementation quality is compromised. Economists call this "optimizing the current production function." Efficiency is important in not-for-profit enterprises because a dollar saved is a dollar that can be applied to producing some other valued output.

Optimization involves allocation (using the right resources for each required task, given their capacities and relative costs), scale effects (the operation is neither too large nor too small), and technical efficiency (applying the right know-how to get the best possible outputs given the resource mix and scale). It also implies minimization of waste, which can occur when tasks are performed badly (leading to expensive rework) or purchased resources are neglected or diverted to other uses.

- *Process Innovation* aims to improve the production function as opposed to optimizing the current one. Often this involves use of information technology, though there are many other possibilities. Such changes include the "reengineering" of teaching and learning or other processes in order to do things in new ways or with new kinds of resources. The metrics used for process efficiency also can track the long term effects of process improvement as will be discussed later, and schools may supplement them with measures that describe the innovation activities themselves (e.g., the fractions of courses that involve computer-mediated instruction or that have been redesigned in other significant ways).

Financial Returns (marketplace productivity)

Universities sell their education and research services in order to generate revenue to supplement whatever they may receive from governments, donors, endowments, etc. Cost is incurred during the process of production, and margin represents the difference between cost and revenue. Calling out revenue, cost, and margin as a kind of productivity stretches the dictionary definition, but it emphasizes that even not-for-profit entities like universities must perform well in the marketplace.

- *Earned Revenue* refers to money generated directly by the product or service. Examples include student tuition and fees, government capitation grants that vary with enrollment, sums obtained from research grants and contracts, and income from public service events. Revenues like those from government institutional grants, gifts, and endowments are not included because, at least in the short run, they do not vary with output. In summary, earned revenue reflects success (or the lack of it) in the marketplace, which depends on "value" as assessed by users.

- *Production Cost* is the amount spent on resources used directly in the production process: i.e., academic and other staff, materials and purchased services, and the utilization of facilities and equipment. The focus here is on cost, *per se*, not on quality and efficiency measures like cost per enrollment. Production cost is important for resource allocation: in particular, decisions about what and how much to produce. It also is essential for the calculation of gross and net margin.

These direct production costs are not the only ones that are important, however. Indirect costs (aka "overheads") also must be incurred for the provision of administrative and support services, without which the university cannot function. Thus, the direct costs are "burdened" with shares of overheads as determined by accounting rules. Speaking very generally, the direct costs are more relevant for resource allocation at the department and program levels and the burdened ones are more pertinent for price setting and resource allocation at higher levels.

- *Margin* equals the difference between revenues and costs. "Gross Margin" refers to earned revenues minus direct costs; "Net Margin" indicates earned revenues minus burdened costs. Both types of margin can be calculated for the university as a whole, for schools or faculties within it, and, with suitable data systems and software, for individual academic programs, departments, and even courses. Margins play a key role in resource allocation and decisions about expansion, contraction, or elimination of programs.

Contrary to the views of many faculty members, the inclusion of data about revenues, costs, and margins as a central consideration in academic decision-making does *not* mean the university is prioritizing money over academic value. Pursuit of a school's academic mission may lead it to cross subsidize programs that have high intrinsic value by using surpluses from ones that have high marketplace value but lower intrinsic value. Having data on net margins is the only way to determine these cross subsidies, and thus make good decisions about how to further academic mission.

METRICS FOR ASSESSING PRODUCTIVITY

Higher education is not lacking in metrics. The problems arise mainly from the near-impossibility of systematically adjusting the measured quantities of educational outputs for variations in quality. This adjustment is made routinely when assessing productivity in manufacturing. (The problem is not so bad for research due to the prevalence of peer review.) The quality measurement problem has led to a proliferation of surrogate metrics, some of which are not well understood and/or have questionable fitness for purpose. Probably the most notorious problem lies in the use of input variables as surrogates for teaching and learning quality, a practice that has been likened to, "Looking for your keys under the lamppost, where the light is good, when you know you lost them down the street." Such usage has been largely debunked, though inputs still appear in some popular higher education ranking systems.

In this section, a set of widely used metrics are listed and described on the basis of how they relate to the productivity paradigm set out in Figure 1. Though based on considerable experience, the proposed associations should not be regarded as definitive. They did not come from a rigorous research process, but rather are intended to start new conversations about productivity and its measurement. The goal is to develop suites of good metrics that, collectively, span all the dimensions in the paradigm. Established institutional ranking systems like those from U.S. News & World Report, Times Higher Education and Shanghai Jiao Tong University (all of which consist of suites of metrics) fall short on the completeness criterion, as well as suffer from data and weighting problems.

Metrics for Educational Programs

Table 1 associates productivity metrics for education with the variables called out in the productivity paradigm. A circle with a dot in the middle indicates that the metric provides a lot of information about where the measured entity stands on the productivity dimension; a circle without a dot indicates a weaker but still significant association. The Table has been simplified by combining Earned Revenue, Production Cost, and Margin under the single rubric of Financial Results.

The following paragraphs describe the reasoning behind the associations in the Table. The discussion is arranged by block number because the metrics called out in each block are at least roughly similar to one another. The presentations are brief and sometimes telegraphic because more metrics are included rather than describing a small number in more detail. However, there is enough clarity for readers who are familiar with the metrics to develop their own judgments about where they fit into the paradigm.

Block 1. Quantity of Educational Output Produced

Enrollments and credits earned are the primary measures colleges and universities use to quantify their teaching output. They produce value because students learn something in each course they take, whether they finish the degree or not. Degrees and certificates awarded add the value of the credential to the value of the credits, and Adjusted Credit Hours combines enrollments and degrees into a single quantity [4]. The last two metrics normalize the number of degrees, etc., to the size of the local labor force and to the population lacking a post secondary credential, which are two concepts that can help inform government funding decisions [5]. All the metrics provide good information about the quantity of output. Degree and certificate production also reflect design and implementation quality to some extent, because curricula that are well designed and delivered tend to enhance the prospects for timely graduation.

Table 1. Mapping of Educational Metrics to the Productivity Dimensions

Block No.	Performance Metrics	Outputs			Outputs / Inputs		Financial Returns (all)
		Amount Produced	Design Quality	Implement-ation Quality	Process Efficiency	Process Innovation	
1	**Quantity of Educational Output Produced**						
	Number of enrollments, credits taught	●					
	Number of degrees, certificates, etc. awarded	●					
	Adjusted Credit Hours (enrollments & degrees—NRC)	●					
	Undergraduate awards per in-state employed adult	●					
	Undergraduate awards relative to adults with no degree	●					
2	**Measures of Learning Quality**						
	Scores on general “college learning attainment” tests		●	●			

(continued on next page)

(continued from previous page)

Block No.	Performance Metrics	Outputs			Outputs / Inputs		Financial Returns (all)
		Amount Produced	Design Quality	Implement-ation Quality	Process Efficiency	Process Innovation	
	Scores on tailored attainment tests (where available)		🟢	🟢			
	Self-reported scores on learning outcomes		🟡	🟢			
	Scores on student engagement surveys		🟡	🟢			
	Scores on student satisfaction surveys		🟡	🟢			
3	**Internal Measures of Student Outcomes**						
	Number of transfers & dropouts per 100 students		🟢	🟢			
	Graduation, transfer & dropout rates for cohorts		🟢	🟢			
	Average times to degree		🟢	🟢			
4	**External Measures of Student Outcomes**						
	Success with graduate school & external certifications		🟢	🟢			
	Employment uptake of graduates		🟢	🟢			
	Starting & mid-career salaries of graduates		🟢	🟢			
	Scores on employer satisfaction surveys		🟢	🟢			
5	**Qualitative Assessments by Outside Experts**						
	Adherence to Degree Qualification Frameworks		🟢				
	Institutional and program accreditation		🟢	🟢			
	Program reviews		🟢	🟡			
	Academic audits		🟡	🟢			

(continued on next page)

(continued from previous page)

Block No.	Performance Metrics	Outputs			Outputs / Inputs		Financial Returns (all)
		Amount Produced	Design Quality	Implement-ation Quality	Process Efficiency	Process Innovation	
6	**Characteristics of Admitted Students**						
	Fraction of applicants admitted		🟡				
	Fraction of admittees matriculating		🟡				
	Test scores & GPAs for matriculating students		🟡				
7	**Quantity and Quality of Inputs to Educational Processes**						
	Student faculty ratios		🟢	🟡	🟡		
	Qualifications of faculty		🟡	🟡			
	Size of library collections		🟡	🟡			
	Quality of facilities		🟡	🟡			
	Size of endowment		🟡				
8	**Cost per Unit of Educational Output**						
	Credit hours per faculty FTE		🟡	🟡	🟡	🟡	
	Cost per credit hour		🟡	🟡	🟡	🟡	
	Cost per degree or certificate awarded		🟡	🟡	🟡	🟡	
	Academic awards per dollar of subsidy & net tuition		🟡	🟡	🟡	🟡	
	NRC Multifactor Productivity Index		🟡	🟡	🟡	🟡	
9	**Direct Measures of Educational Process Innovation**						
	Percent of courses offered online					🟢	
	Percent of courses using advanced IT, etc.					🟢	

(continued on next page)

(continued from previous page)

Block No.	**Performance Metrics**	**Outputs**			**Outputs / Inputs**		**Financial Returns (all)**
		Amount Produced	Design Quality	Implement-ation Quality	Process Efficiency	Process Innovation	
	Percent of courses using “flipped classrooms,” etc.					●	
10	**Course-Level Activity-Based Costing (ABC)**						
	Activity-Based Costing						●
	Enhanced Activity-Based Costing		○	●	●	●	●

Key: ● means “strongly associated”; ○ means “weakly associated.”

FTE, full time equivalent; NRC, National Research Council.

Block 2. Measures of Learning Quality

The first metric, scores on general “college learning attainment” tests have been the holy grail for the assessment of college learning. To the extent they are successful, they reflect both design and implementation quality. These tests are rooted in a hoped-for consensus about what students should learn in college, but such has proven to be elusive and there are also technical problems (e.g., which students will take the tests and whether they are sufficiently motivated to produce valid results)[4]. The next three metrics provide indirect evidence about implementation quality, but not so much on design quality. This is because students can judge whether they are engaged and satisfied, but they are less able to discern which course materials are important or meaningful.

Block 3. Internal Measures of Student Outcomes

These three metrics refer to success in pursuing the degree, which depends on both design and implementation quality as noted above. Cohort measures are obtained by tracking groups of entering students independently until all or most group members have graduated or left the university due to other reasons. They are more meaningful than aggregate annual numbers for graduations and dropouts, but it takes more effort and elapsed time to obtain them.

Block 4. External Measures of Student Outcomes

These are typical metrics for how graduate schools, certification agencies, and employers respond to graduates’ educational attainment. Universities track these results by surveying graduates, or in some cases entities that are known to employ significant numbers of graduates. Australia and some other countries require such surveys, and the idea seems to be catching on in many parts of the world. All four metrics tap elements of program design and implementation quality. External stakeholders often are in an excellent position to gauge these factors because they have direct experience with what graduates can and cannot do. However, the metrics are subject to certain practical limitations, as will be discussed later.

Block 5. *Qualitative Assessments by Outside Experts*

Adherence to a Degree Qualification Framework reflects recent efforts by the Lumina Foundation and others to establish general criteria for appropriate learning content and level for degrees in a particular field. This speaks directly to design quality, but not at all to implementation quality. Accreditation teams ask whether institutions, or in some cases programs, have the necessary inputs for delivering quality education, and (importantly for our purposes) whether they have processes in place to deliver both design and implementation quality. Program Review teams typically investigate the details of design quality but not implementation quality. Academic Audit looks at the management of design quality decisions, and the effectiveness and consistency of the teaching processes (implementation quality). Audit is arguably the gold standard for evaluating quality assurance and improvement processes [3].

Block 6. *Characteristics of Admitted Students*

Metrics describing the characteristics of a school's students are important determinants of applicant choice and institutional prestige. Students' decisions to apply are influenced by published program descriptions (design quality) and the choices of students that applicants view as peers, among many other things. However, the lack of public information about schools' teaching quality and consistency limits applicants' ability to gauge implementation quality, and thus the measures' effectiveness for that purpose.

Block 7. *Quantity and Quality of Inputs to Educational Processes*

Measures of the quantity and quality of inputs are sometimes used as surrogates for quality. All these metrics have some relation to design quality because "richer" designs often cost more than bare-bones ones. Student-faculty ratios also can be associated with implementation quality: e.g., when lower ratios lead to smaller class sizes and allow faculty to spend more time with students.

The associations are shown as weak, however, because technology and other process innovations may produce excellent low-cost designs, or money may be used for non-teaching purposes like research and student amenities rather than for small classes.

Block 8. *Cost per Unit of Educational Output*

These metrics are used for assessing process efficiency and the effects of process innovation. Short-run changes may reflect process efficiency changes and longer-term ones the effect of process innovation, which takes longer to accomplish. All the metrics are heavily confounded with design quality. This can cause serious problems for the unwary user, who may confuse productivity improvement with quality erosion, as entities “race to the bottom” in order to look good on the metric [2]. The NRC (National Research Council) Multifactor Productivity Index, which measures the change in outputs divided by the change in inputs, mitigates these problems to some extent, but it does not eliminate them [4].

Block 9. *Direct Measures of Educational Process Innovation*

Direct measures of process innovation are proving to be worthwhile supplements to the indicators in Block 7. Three examples are listed, all of which are strongly associated with particular instances of innovation. Other possibilities include the use of e-portfolios, learning commons, and other facilitators of active learning. Such data may be collected from institutional records (e.g., timetabling or room utilization), or by institutional or third-party surveys of faculty and IT staff.

Block 10. *Course-Level Activity-Based Costing (ABC)*

The need for metrics related to the Financial Returns from academic programs has generated a growing interest in ABC at the level of individual courses and degree programs. Models for accurately tracking revenue, cost, and margin at these levels have been developed over the last ten years and are now beginning to catch on [6]. (The key breakthrough was to use student registration, timetabling, and other university systems data to capture the durations of various teaching activities, the resources used, and the costs thereof, in all courses for every semester.) Enhanced Activity-Based Costing adds reports about the configuration of teaching activities for each course [2]. These reports mitigate the confounding problems referred to in Block 7 by providing detailed information about class sizes, use of auxiliary faculty, and other quality-influencing factors, right along with the information about costs, and thus builds firewalls against a race to the bottom. These methods shall be discussed in more detail shortly.

Metrics for Research Programs

Table 2 associates some representative research metrics with the paradigm. The layout is the same as in Table 1, except for simplifications due to omitting the distinctions between design and implementation quality, and process efficiency and process innovation. These distinctions are difficult to judge at an institutional or departmental level because value creation takes place continuously at the level of the principal investigator. Note that the table does not include composite indicators like the Shanghai Jiao Tong index, which are constructed from individual metrics of the type shown. As with Table 1, the reasoning behind the associations appears in the paragraphs following the Table.

Table 2. Mapping of Research Metrics to the Productivity Dimensions

Block No.	Performance Metrics	Outputs		Outputs / Inputs (all)	Financial Returns (all)
		Amount Produced	Quality (all)		
11	**Quantity of Research Output Produced**				
	Books, papers & other creative materials	🟢	🟢		
	Research student completions & job placements	🟢	🟢		
	Citations		🟢		
	Downloads of digital materials		🟢		
	Sponsored research funding	🟢	🟢		🟡
	Collaborative projects with industry	🟢	🟢		🟡
	Patents & other intellectual property	🟢	🟢		🟡
12	**Research Output Produced per Faculty FTE**				
	Above measures, per faculty FTE			🟡	
13	**Quantity and Quality of Inputs to Research Processes**				
	Institutional funds spent on research	🟢	🟢		
	Non-faculty professional research staff FTE	🟢	🟢		

(continued on next page)

(continued from previous page)

Block No.	Performance Metrics	Outputs		Outputs / Inputs (all)	Financial Returns (all)
		Amount Produced	Quality (all)		
	Amount of research space, specialized labs, etc.		●		
	Research-active faculty		●		
	Average teaching loads	●	●		○

Key: ● means "strongly associated"; ○ means "weakly associated."
FTE, full time equivalent.

Block 11. Quantity of Research Output Produced

Universities count books, papers, and other creative materials (sometimes weighted by type and publication vehicle) as primary measures of their faculty's research and scholarship. These counts reflect quality as well as quantity due to the effects of editorial evaluation and peer review. Research student completions and job placement reflect the same kinds of factors. Citation counts, now readily available for many fields, often are taken as measures of quality. Downloads of digital materials (e.g., reports, software, and databases) also are used as quality measures in some areas. Sponsored research funding and collaborative projects with industry reflect the quantity and quality of past research and serve as enablers of future research. The revenue they provide also affects financial returns. Patents and other intellectual property reflect past research outputs and generate financial returns.

Block 12. Research Output Produced per Faculty FTE

Reporting the outputs described above on a per-faculty (or per-research-active faculty) basis provides a rough measure of process efficiency, but the association is shown to be weak because so many other variables are involved.

Block 13. Quantity and Quality of Inputs to Research Processes

The discipline of peer review link the quantity and quality of research outputs more closely to resource availability than was the case for teaching undergraduates, where there is no peer review for quality. This is reflected in the first four metrics. The number of "research-active faculty," which counts people who regularly produce more than a threshold number of publications, etc., provides another indicator of research output. Finally, average teaching load (for all faculty or only the research-active ones) affects the amount of time that faculty spend on their research during the academic year.

EVALUATION OF THE PRODUCTIVITY METRICS

All the metrics in Tables 1 and 2 have arguable associations with one or more of the productivity-paradigm elements presented earlier. Indeed, having such "face validity" was a major consideration in constructing the tables. But that does not mean they are of equal utility. In part, the confusion about higher education productivity arises because some of the metrics in wide use are not fit for purpose, even though they do seem relevant on their face.

What are the attributes of a good metric? A new book on research evaluation by Professor and Canada Research Chair in History and Sociology of Science, Yves Gingras, Université du Québec à Montréal, explores this question in some depth [7]. The basic requirements are that the indicator "can be measured and that it aims at faithfully representing" the underlying concept and "how the reality behind the concept changes over time or place." Gingras lists three crucially important criteria for a good metric. To paraphrase his argument, a good metric should possess the attributes of:

1. *Adequacy.* The metric should correspond to the object or concept ("target object") being evaluated: in our case, an element of the paradigm in Figure 1. It should have logical or empirical linkages to the target object, or at least face validity rooted in our intuition about it.

2. *Sensitivity.* The metric should vary in a manner consistent with the "inertia" of the object being measured. Different objects change with more or less difficulty (and rapidity) over time. The metric's volatility should be consistent with the volatility of its target object, lest it send erroneous signals about changes (or lack of them) over time or space.

3. *Homogeneity.* The metric should be homogeneous in its composition, so that it will not send ambiguous signals about which objects are changing. In particular, a good metric will not be associated with more than one target object unless the multiple objects are so closely linked that the confusion as to which is varying does not matter for the purpose at hand.

Gingras uses these criteria to criticize some widely used indices of research performance.

Table 3 reports conclusions about how each group of metrics in Tables 1 and 2 fares with respect to Gingras's three criteria. The first column contains the block numbers used in the previous tables, and the second one provides a shorthand description of the metrics in the block. The third column shows what is believed to be the Target Object for each block. This is based on the preponderance of associations in Tables 1 and 2, though there are some exceptions. The remaining columns classify the metrics in each block as good, mixed, or weak as described below. Together, these classifications call out the metrics' "fitness for purpose." Again, the reasoning behind the scoring is found immediately following the Table, and the earlier caveat about brevity applies here as well. These ideas should be viewed as preliminary suggestions only, but they do demonstrate how Gingras's criteria can be used for educational as well as research metrics.

Table 3. Evaluation of the Productivity Metrics

Block	Type of Metric	Target Object	Adequacy	Sensitivity	Homogeneity
Education					
1	Quantity produced	Quantity	good	good	good
2	Internal quality measures	Quality	mixed	mixed	mixed
3	Internal outcomes measures	Quality	good	good	good
4	External outcomes measures	Quality	good	mixed	good
5	Expert quality judgments	Quality	good	mixed	good
6	Admission statistics	Quality	weak	good	good
7	Inputs	Quality	weak	good	mixed
8	Unit cost	Output/Input	good	good	very weak
9	Direct process innovation	Output/Input	good	good	good
10	Activity-based Costing	Output/Input	good	good	good
Research					
11	Quantity produced	Quantity	good	good	good
12	Above items/faculty	Output/Input	good	good	good
13	Inputs	Output/Input	good	good	good

Most of the *Adequacy* scores are "good," which is not surprising since adequacy was the main criterion for selecting the metrics in the first place. The "mixed" score for Block 2 arises because, as noted earlier, student responses provide an incomplete picture of educational quality. The weak score for Block 6 is due to the questionable linkage, also noted earlier, between students' school choice behavior and the delivered quality of teaching, and the one for Block 7 is due to the same problem as it applies to inputs. The rating for Block 8 (unit cost) is "good" because all the metrics possess good face validity. (The ambiguity problem discussed earlier is handled under the homogeneity rubric.)

"Good" *Sensitivity* scores are given for metrics that provide direct measures of their target quantities, in the sense that the number of enrollments is a direct measure of teaching quantity. The "mixed" rating for Block 2 stems from the sampling errors inherent in self-reported learning attainment and student engagement and satisfaction surveys. The underlying variables that these surveys measure don't usually change quickly, but the survey results often do. Such problems also plague external outcomes measures and expert quality judgments (Blocks 4 and 5): samples often are small or infrequent, and the responses themselves are subject to error. Salary surveys suffer from additional problems due to the difficulties of representative samples of graduates and employers, not to mention the time lags that must be incurred before getting mid-career results.

The majority of metrics have reasonably good *Homogeneity*, but there are a few important exceptions. Block 2 gets a mixed rating because the most widely used metrics tend to short-change implementation quality, and thus provide an incomplete view of an institution's performance. Another exception occurs with respect to student faculty ratios (Block 7), which often are viewed as reflecting only process efficiency as opposed to quality. Once again, confounding these two interpretations has unfortunate policy implications. By far the worst problem occurs in Block 8 (cost per unit), where all the metrics inexorably confound quality and efficiency. This problem generates more mischief in the evaluation of university productivity than any of the others (see the example in Massy [2]).

DISCUSSION

In writing this Guide, an attempt has been made to bring a sense of order to the plethora of performance indicators that are being used in higher education. This is more than an academic exercise: people describe particular metrics as measuring things like "quality" and "productivity" without being at all precise about what is meant, leading to unnecessary confusion and conflict about the selection and use of metrics. The Guide begins (Figure 1) by defining a paradigm, made up of three dimensions and eight sub-dimensions, for productivity as applied to higher education. It continues (Tables 1 and 2) by suggesting how representative metrics for education and research productivity map onto the various dimensions. It concludes by evaluating the metrics' fitness for purpose in terms of adequacy, sensitivity, and homogeneity.

The purpose of measurement may vary by the type of entity that will use the results. Government agencies and oversight bodies need aggregate measures that can be used at least in part for accountability. For example, the U. S. National Governors Association included "Academic Awards per $100,000 of government subsidy payment and net tuition receipts" and "Undergraduate awards per employed in-state adult" in its list of metrics [5] because undergraduate awards are what many stakeholders feel they are buying with their funding. On the other hand, individual universities need disaggregated measures like cost per credit hour for particular courses and degrees because, in addition to the need for internal accountability, they seek information for informing their own improvement decisions. Many metrics can serve for both external oversight and internal improvement, but even here there will be differences in emphasis and interpretation. Thus we should neither seek nor expect consensus on a limited or standard set of metrics. What we should expect, though, is clarity about what each metric is intended to measure. Especially, we need suites of metrics that cover all the elements of the productivity paradigm.

In what direction is best practice moving at the present time? With respect to outputs, the need for tracking both enrollments and degrees (as well as other awards) now appears to be well established, and the use of Adjusted Credit Hours as a composite measure appears to be particularly attractive. Likewise, there has long been agreement that the assessment of design quality is important: for example, as with the development of Degree Qualification Frameworks. There is far less agreement about the need for assessing implementation quality for teaching and learning, let alone the methodology for doing so. The expansion of Academic Audit and the development of Generally Accepted Audit Principles for use in teaching [2] were recommended, but those ideas have yet to be adopted. Measures like "Success with graduate school and external certifications," "employment uptake," "salaries," and “employer satisfaction surveys” are useful because they get at both design and implementation quality, but that does not mean the two need not be measured separately.

Probably the greatest challenges lie in the area of cost per unit of output (“outputs/inputs”). Cost per credit hour and similar metrics confound productivity with the value of output: for example, does a cost reduction signify better productivity or lower quality? These widely used measures can alert managers and analysts to the possibility of productivity gains or losses, but they should never, ever be regarded as determinative or used alone for purposes of accountability. The only solution to this problem is to derive such measures from detailed data on faculty usage, class sizes, etc., instead of calculating them from data for total departmental costs and total credit hours. The disaggregated data can be obtained from Enhanced ABC models. They enable one to drill down and determine whether cost reductions are due to better efficiency or process innovation as opposed to increasing already-large class sizes or other quality-reducing actions. Juxtaposing cost results with activity descriptions makes pressures for productivity improvement less likely to trigger a “race to the bottom” in terms of quality. Today’s ABC models cost out research as well as teaching, but the results do not yet include the “Enhanced” activity measures found in the teaching models. Exploratory work has begun on how to alleviate this difficulty, however.

Shifting to the aggregate level, the NRC Multifactor Productivity Index is a better productivity measure than single-factor metrics like “Credit hours per faculty FTE,” “Cost per credit hour,” and the like. This is because it combines credit hour and degree production in its output measure, and also because of certain technical advantages in the way it is constructed. The danger of triggering quality degradation remains, but the problem is not quite so pressing. It still is necessary, however, to maintain a robust quality assurance system (ideally for both design and implementation quality) to guarantee that what is viewed as “productivity” improvement does not in fact represent quality diminution. Hopefully, the NRC index will be adopted more widely as time goes on.

These developments kindle optimism that the “black box” of what goes on in academic production can be penetrated. Enhanced ABC models are available and affordable for most schools today, although their adoption is at an early stage. It is expected that the current pressures on universities, and funding agencies and students, will drive adoption of these tools and others (e.g., the so-called “academic analytics”) relating to “big data.” Such developments will benefit both the universities and their stakeholders.

REFERENCES

[1] Massy W.F. Honoring the Trust: Quality and Cost Containment in Higher Education. San Francisco: Jossey-Bass; 2003

[2] Massy W.F. Reengineering the University: How to Be Mission Centered, Market Smart, and Margin Conscious. Baltimore: Johns Hopkins University Press; 2016

[3] Massy W.F., Graham S.W., Short P.M. Academic Quality Work: A Handbook for Improvement. San Francisco: Jossey-Bass; 2007

[4] Sullivan T.A., Mackie C., Massy W.F., Sinha E., eds. Improving Measurement of Productivity in Higher Education. Panel on Measuring Higher Education Productivity: Conceptual Framework and Data Needs. Washington, DC: The National Academies Press; National Research Council (NRC), 2012.

[5] Travis R., Reyna R. Complete to Compete, from Information to Action: Revamping Higher Education Accountability Systems. Washington, DC: National Governors Association (NGA) Center for Best Practices; 2011.

[6] Anguiano M. Cost Structure of Post-Secondary Education: Guide to Making Activity-Based Costing Meaningful and Practical. Virginia Beach, VA: Pilbara Group; 2013. https://www.ucr.edu/about/admin/docs/Cost-Structure-of-Post-Secondary-Education-Guide-v2.pdfArchibald

[7] Gingras Y. Bibliometrics and Research Evaluation: Uses and Abuses. Cambridge, MA: The MIT Press; 2016.

[8] Mackie C. Conceptualizing and Measuring Productivity in US Higher Education. TIAA Institute and National Association of College and University Business Officers (NACUBO); 2016. https://www.tiaainstitute.org/public/pdf/institute_series_conceptualizing_measuring_productivity.pdf

[9] Massy W.F. Course-Level Activity-based Costing As an Academic and Financial Tool. TIAA Institute and National Association of College and University Business Officers (NACUBO); 2016. https://www.tiaainstitute.org/public/pdf/institute_series_course_level_activity_based_costing.pdf

[10] Massy W.F., Sullivan T.A., Mackie C. Data Needed for Improving Productivity Measurement in Higher Education. Research and Practice in Assessment; Winter 2012: 5-15.

APO