ACCESS: CONTEMPORARY ISSUES IN EDUCATION
2008, VOL. 27, NO. 1 & 2, 8–23

ACCESS

The fundamental characteristics of international models and mechanisms for evaluating government-funded research

Chris L. S. Coryn

Western Michigan University, USA

ABSTRACT
In the last few decades the evaluation of scientific research has become a high-stakes enterprise. With increasing political governance and federal budgets often in the billions, the livelihood of individual researchers, research groups, departments, programs, and entire institutions often swing in the balance. With its foundations in the traditional peer review system, many nations throughout the world now have large-scale systems in place for prospectively (ex ante) and/or retrospectively (ex post) evaluating their government-financed research. This paper begins by presenting an overview of the research evaluation mechanisms in sixteen countries in terms of their primary reasons and motives for evaluating government-funded research, their basic units of assessment and core methods, and their key indicators and criteria. The paper concludes by classifying these models and mechanisms along dimensions of: (1) their basic approach to allocating or distributing research funding; (2) their general research evaluation approach or strategy; and (3) their overall quality.

Large areas of scholarly research are publicly funded and in most parts of the world government funding for research is highly contested. Under demands for greater accountability, due to constraints of diminished funding, and in the pursuit of general quality improvements, many countries have initiated systems for evaluating publicly-funded research at the national level. The evaluation of publicly-funded research now has a substantial tradition (Cozzens & Turpin, 2000), particularly in the European countries, dating to the early 1970s. Some of the earliest efforts were undertaken in the Nordic countries, of which Sweden was the first country to carry out systematic evaluations of its research in the 1970s, followed in the mid-1980s by Finland, Norway, and Denmark (Luukkonen, 2002). Although there are vast differences in the way governments fund research around the world (Campbell, 2002; Campbell & Felderer, 1997; Cozzens & Turpin, 2000; Laudel, 2006), and a diversity of approaches to evaluating publicly-funded research (ab Iorwerth, 2005; Geuna & Martin, 2001, 2003; OECD, 1987, 1997, 2003; Orr & Paetzold, 2006; von Tunzelmann & Mbula, 2003), almost all now share the common purpose of relating funding to performance (ab Iorwerth, 2005; Campbell, 2002; COSEPUP, 1999, 2001; Cozzens & Turpin, 2000; Geuna & Martin, 2003; Luukkonen, 2002; OECD 1987, 1997; RAE 2008, 2005; Scriven, 2006). Moreover, as the OECD noted in its 1997 report *The Evaluation of Scientific Research*:

> … research evaluation has emerged as a 'rapid growth industry'… [and] … there is an increasing emphasis on accountability, as well as on the effectiveness and efficiency of government-supported research … governments need such evaluations for different purposes: optimizing their research allocations at a time of budget stringencies; re-orienting their research support;

> rationalizing or downsizing research organizations; augmenting research productivity. To this end, governments have developed or stimulated research evaluation activities in an attempt to get 'more value for the money' they spend on research support. (OECD, 1997: 5)

This paper begins by presenting a summary of the primary purposes for evaluating publicly-funded research, the basic units of assessment, the core methods, the key indicators used to assess publicly-funded research, researchers, and institutions, and classifications of the research evaluation system and funding mechanisms in sixteen countries. Of the 272 nations, dependent areas, and other entities in the world, these countries represent more than two thirds of the world's top purchasing power parities, as well as a large majority of the world's "research superpowers" in terms of their scientific productivity and government monies dedicated to research. These nations are: Australia; Belgium; the Czech Republic; Finland; France; Germany; Hong Kong; Hungary; Ireland; Japan; The Netherlands; New Zealand; Poland; Sweden; the United Kingdom; and the United States (see Figure 1). The paper concludes with an overview of a study conducted to determine the merits of these national-level systems.

The fundamental characteristics of international research evaluation models and mechanisms

Herein, the fundamental characteristics of the sixteen countries' research evaluation models are presented in terms of their: (1) primary reasons and motives; (2) basic units of assessment; (3) core methods; (4) key indicators and criteria; (5) systemisation and consistency; and (6) funding model/archetype.

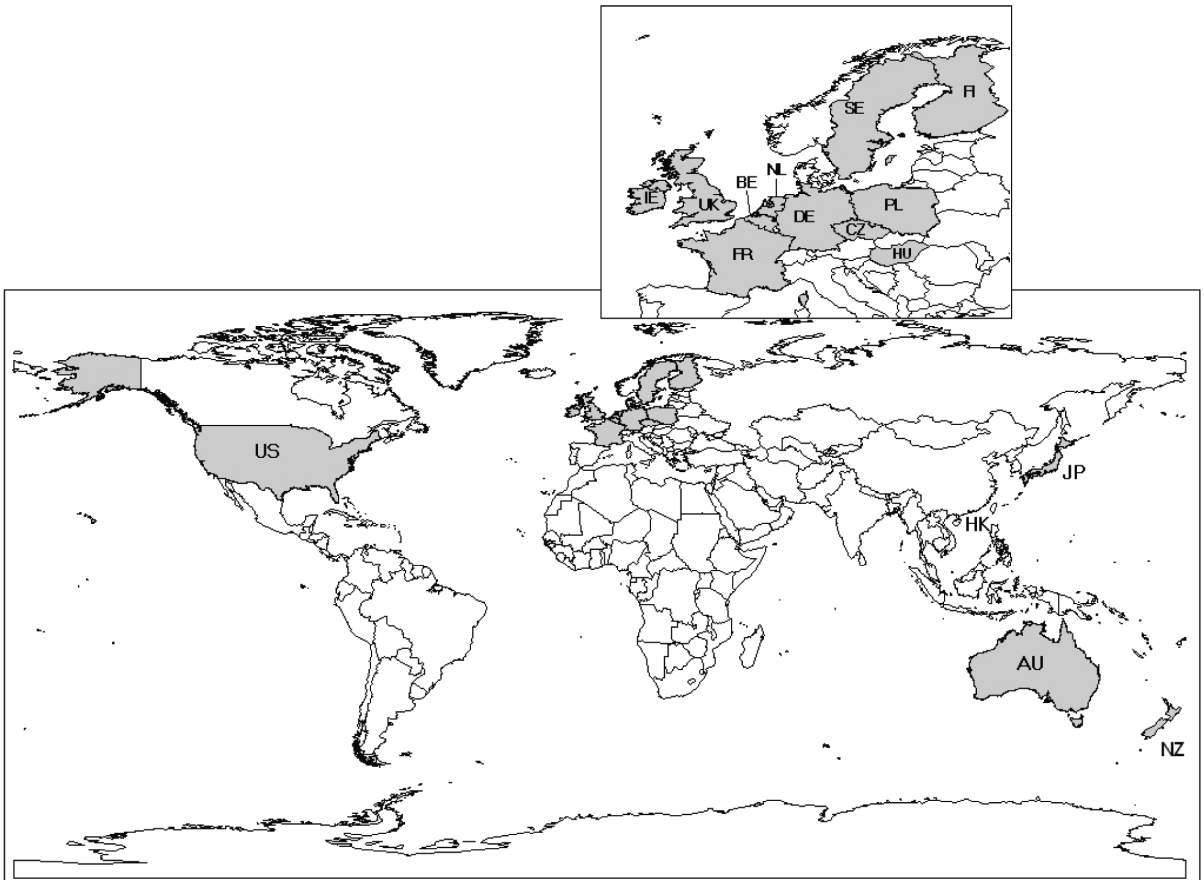

Figure 1 Sample countries
(AU = Australia; BE = Belgium; CZ = Czech Republic; FI = Finland; FR = France; DE = Germany; HK = Hong Kong; HU = Hungary; IE = Ireland; JP = Japan; NL = Netherlands; NZ = New Zealand; PL = Poland; SE = Sweden; UK = United Kingdom; and US = United States)

Primary reasons and motives

The evaluation of research serves numerous purposes, although there have been extensive debates, and in general, an overwhelming lack of consensus as to what these reasons and motives are or should be (e.g., Aksnes, 2005; Moed, 2005; Rousseau, 2004; Russell & Rousseau, 2002; van Raan, 2005). In part, this disagreement can be attributed to the larger context in which the evaluation of research takes place. In most cases the evaluation of a nation's research serves vastly different purposes than evaluation conducted by a department or research group considering candidates for a research position, tenure, promotion, or demotion, or than the evaluations conducted by a journal editor or peer reviewer assessing a paper's merits for publication.

There are essentially five fundamental purposes for evaluating research, although there is some overlap, which can be broadly classified as: accountability and efficiency; resource allocation; improvement; synthesis; and decision making. With the exception of improvement, most research evaluations are summative, and in some cases synthesis is done for ascriptive rather than summative purposes. Excluding synthesis, and as mentioned previously, if there is a single word to describe these purposes it is "governance" (Frederiksen, Hannson & Wennberg, 2003). Governance is a somewhat ambiguous term for social regulatory processes that implicate the political system directly or indirectly; it is analogous to the psychologists' and sociologists' term "social control" (Hannson, 2006).

In any case and whatever the purpose, the evaluation of research has been called *a priori* or *a posteriori* (Weinberg, 1963, 1989). In the first instance research is evaluated prospectively, often referred to as *ex ante* evaluation (Meyer-Krahmer & Reiss, 1992), for predicting future performance, normally on the basis of prior performance. In the second instance research is evaluated retrospectively, often referred to as *ex post* evaluation (Campbell & Felderer, 1997), after it has been completed. *Ex ante* evaluation of research is normally used for awarding research funding for proposed research, whereas *ex post* evaluation of research is applied for determining the merits or significance of completed research, for instance, in awarding Nobel Prizes.

In evaluating researchers and their research, accountability and efficiency, priority setting, resource allocation, synthesis, and decision making are primarily summative endeavors, although in some cases they can be done for formative, ascriptive, or less frequently, proformative (Coryn, 2007a) purposes. Improvement, however, is an entirely formative procedure in most cases, although it often occurs as a result of summative evaluation.

Accountability and efficiency

As a purpose for evaluating research, particularly publicly-funded research, accountability and efficiency is the responsibility for the justification of expenditures, decisions, or the results of research efforts. Accountability often requires some measure of cost-effectiveness, where cost-effectiveness is taken to be more than explanations of how financial resources were spent, but also justifications in the results produced from these expenditures. There is considerable variation in who is required to answer to whom, concerning what, through what means, and with what consequences. Economic, social, and other benefits, often referred to as impacts, are normally subsumed under accountability.

While accountability is most often considered a purpose for the evaluation of a nation's research or its expenditures of taxpayer monies on research initiatives or agendas, it is equally applicable to research institutions, groups, departments, or individuals; that is, they are equally accountable for justifying expenditures, decisions, or the results of their research efforts. This can also be extended to include accountability for who is tenured, promoted, demoted, hired, or fired by a research institution, group, or department, for example. At the personnel level, accountability serves to justify costs to students, taxpayers, colleagues, and others in the selection of researchers. In practice, however, many systems of accountability are subject to several forms of corruption and

"hence are likely to reduce the sense of responsibility for and quality of performance" (Rogers, 2005: 2).

Resource allocation

Resource allocation, or apportionment, is often an explicit, and in some cases implicit, purpose for the evaluation of researchers and their research. Conceptually, resource allocation involves matters such as national priority setting, which normally includes the distribution of research funding (Coryn, 2007b).

Resource allocation may be one of the most important purposes underlying the evaluation of research, and not entirely unrelated to accountability. Ultimately, investments in research are like other types of investments, more uncertain, but conceptually similar (Scherer, 1967). However, these allocations frequently involve a great deal of trial and error. In strategic planning, a resource-allocation decision is a plan for using available resources, especially human resources in the near term, to achieve goals for the future. It is the process of allocating resources among various projects, units, or alternatives.

A typical allocation plan has two parts. First, there is the basic allocation decision and second there are contingency mechanisms. The basic allocation decision is the choice of which items to fund in the allocation plan, and what level of funding each should receive, and which to leave unfunded. That is, resources are allocated to some, not to others. There are two contingency mechanisms. There is a priority ranking of items excluded from the plan, showing which items to fund if more resources should become available and there is a priority ranking of some items included in the plan, showing which items should be sacrificed if total funding must be reduced.

All decision makers have to work within a world where resources are scarce in comparison with alternatives for their use. Those responsible for the allocation of funds to competing lines of research are no exception to this rule of constrained decision making, and certain characteristics of research make it particularly difficult to decide on the best distribution of resources. The most salient of these characteristics is that the net benefit from any line of research is, by its very nature, uncertain, since there is no sure way of predicting whether a particular researcher or group of researchers will be able to produce research of a significant value.

Improvement

Since Scriven's introduction of the term "formative evaluation" in 1967, improvement has been recognized as a fundamental purpose for many evaluative endeavors. As an explicit enterprise, however, evaluation for improvement is a relatively new and often ignored purpose for the evaluation of researchers and their research.

In some countries, the intended purpose of research evaluation is to invoke an intra-regional or inter-researcher competitive spirit (Saegusa, 1999a, 1999b), in order to produce general quality improvements in its researchers and their research (Swinbanks, Nathan & Triendl, 1997), and ultimately their place in the world's research spectrum (e.g., their international ranking). Normally, improvement is a secondary function of national-level evaluation of publicly-funded research, expected to occur as a result of competition for research monies.

However, these efforts do not always invoke an inter-regional competitive spirit, but rather encourages game playing in some cases. While improvement of a nation's researchers or their research is a long way from improving the quality of a manuscript submitted to a journal for publication, it is nevertheless an essential function of the evaluative endeavor as it is currently understood. In some parts of the world, however, evaluating the research of one's peers or

colleagues is still viewed as an incursion upon longstanding cultural traditions, despite the potential for general quality improvements in their research.

> … research assessment is an alien concept that runs directly against the grain. This is a region, after all, in which deep-rooted traditions demand respect for elders and the promotion of harmony and co-operation at the expense of individuality and competition … openly judging the quality of scientists and firing those who do not come up to mark is hard … in cultures built on Confucian and Buddhist values of respect and group harmony. (Swinbanks, Nathan & Triendl, 1997: 113)

Nevertheless, many Eastern governments and research institutions are recognizing that more creativity and innovation in their research systems may be essential to the future success of their economies, and are rapidly adopting and adapting Western techniques of research assessment in an attempt to improve the productivity and the quality of their research output (Campbell, 1997; Coryn, 2006a, 2006b; Frankel & Cave, 1997; Swinbanks, Nathan & Triendl, 1997).

Synthesis

There are some (e.g., Campbell Collaboration, 2006; Cochrane Collaboration, 2006) who view the purpose of research evaluation as a synthesis activity, much along the lines of modern meta-analysis or systematic review (Glass, 1976; Pawson, 2006), which is primarily a summative undertaking, but also a special case of ascriptive evaluation. Essentially this view sees scientific knowledge as an accumulative endeavor and uses statistical techniques to combine the results of several studies that address a set of related research hypotheses for computing an average effect size across all relevant studies is computed using a weighted mean, whereby the weights are equal to the inverse variance of each study's effect estimator (e.g., Cohen's *d*, Hedge's *g*, Glass' Δ).

Meta-analytic studies have grown in number over the last few decades and its popularity in the social sciences and education is nothing compared to its influence in medicine, where literally hundreds of meta-analyses have been published in the past twenty years. Moreover, the increasing use of meta-analysis has encouraged some researchers to view their studies as making contributions to previous research and to report their results so that they can easily be incorporated (e.g., effect sizes and confidence intervals) into future meta-analysis.

These types of evaluations of research are useful evaluative endeavors, for example, for getting to the bottom line, identifying critical competitors, and possible side effects, among others, and are often considered the gold standard for evidence-based policy and practice, particularly in the health disciplines. More recently, large-scale synthesis of this type can be observed by the establishment of the United States Department of Education's (USDOE) What Works Clearinghouse (WWC), which collects, screens, and identifies studies of effectiveness of educational interventions, including programs, products, practices, and policies.

Decision making

Decision making, although summative in purpose, has been classified as a separate purpose since there are other aspects of the decision making function involved in the evaluation of researchers and their research than the usual summative issues: whether or not one has been accountable for research spending; if research is worthy of synthesis to inform policy or practice; or if research resources have been distributed justly. It also involves matters such as selection, prioritisation, and prediction. For selective purposes, decision making involves the evaluation of proposals, whether or not for funding, research submitted for publication, research products, and research personnel. That is, "which research proposals receive funding, which articles get published, and which researchers … get appointed and promoted" (Frankel & Cave, 1997: 1).

Priority setting in research, usually at the national level, serves the purpose of answering questions such as "now what?" "how much?" and "to whom?" For example, "whether or not to go to

the moon, and how much should go for the support of high energy physics" (Weinberg, 1989: 4–5). Priority setting, while a purpose for evaluating research, is sometimes a precursor to other purposes, namely the aforementioned process of selecting from amongst research proposals. Given that the results of most research are largely unknown, this selecting includes prioritising (e.g., which research projects are most important?) – which brings one to the fact that decision making often requires making predictions.

Prediction, though not a fundamental purpose of evaluation, is almost unavoidable in the evaluation of research and researchers. As Salmon (1998) points out, there are at least three - probably more – legitimate reasons for making predictions. First, predictions are made on the basis of simple curiosity about future events, without waiting for the events to transpire. Second, predictions are often made for the sake of testing a theory or hypothesis. Third, there are situations in which some practical action is required, and the choice of optimal actions depends upon predicting future occurrences. It is the third case which is of interest in the evaluation of research, particularly in regards to researchers. However, this is not the type of prediction which deals with the predictive aspect of scientific knowledge embodied in the predictive content or power of a scientific theory, for instance. It is the prediction of future performance on the basis of past performance.

As shown in Table 1, 94% of the national systems evaluate their publicly-funded research for reasons of accountability and efficiency, 63% for resource allocation, 50% for improvement, and 31% for other types of decision making (e.g., setting research policies or priorities). A large majority (81%) evaluate their publicly-funded research for two or three of these reasons.

Table 1. International research evaluation models' primary reasons and motives for evaluating research

	AU	BE	CZ	DE	FI	FR	HK	HU	IE	JP	NL	NZ	PL	SE	UK	US
Accountability and efficiency	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Resource allocation	✓	✓					✓	✓			✓	✓	✓	✓	✓	✓
Improvement			✓		✓		✓		✓	✓	✓	✓			✓	
Decision making		✓	✓						✓	✓						✓

Basic units of assessment

Typically, national research evaluation models emphasize one or more of the following eight units of assessment:

Research products. Research products are normally confined to scholarly publications, but may also include patents, computer programs, and other technologies and innovations.

Individual researchers. Individual researcher's performance; usually includes research products.

Research groups. Researchers from different institutions or universities active in the same specialty or discipline.

Programs or projects. Programs and projects usually in relation to national priority areas (e.g. renewable energy research); includes large- and small-scale government-financed research programs and projects.

Departments. Departments are usually discipline-specific units (e.g., chemistry, education, physics, mathematics, psychology, sociology) within an institution.

Institutions. In most countries, institutions are typically places of higher learning/ education (i.e. universities).

Disciplines. Entire scientific disciplines or research collectives.

Policies. National research or research evaluation policies; including research funding policies.

As shown in Table 2, the most common unit of assessment in the sampled countries is the institution (69%), followed by research products (50%). Only the United Kingdom uses departments as a unit of assessment; albeit, within institutions via assessment of research products.

Table 2. International research evaluation models' basic units of assessment

	AU	BE	CZ	DE	FI	FR	HK	HU	IE	JP	NL	NZ	PL	SE	UK	US
Research products	✓			✓			✓					✓		✓	✓	✓
Individual researchers	✓		✓		✓		✓					✓	✓	✓		
Research groups					✓			✓					✓			
Programs or projects		✓	✓	✓	✓	✓			✓					✓		✓
Departments															✓	
Institutions	✓	✓		✓		✓	✓	✓	✓	✓	✓		✓		✓	
Disciplines				✓	✓					✓				✓		
Research products		✓	✓		✓	✓		✓	✓							

Core methods

Methodologically, most national systems typically use one or more of the following 13 approaches or strategies to evaluate their publicly-financed research:

Bibliometrics. Typically, bibliometric methods are confined to scholarly publications (including patents) and citations to them; it also includes spatial mapping, data mining, data visualization (e.g., research networks), webometrics, and similar techniques.

Case studies. Gathering and analyzing data about one or a small number of examples as a way of studying a broader phenomenon; done on the assumption that the example (i.e., case) is in some way typical of the broader phenomenon.

Comparative studies. Studies of more than one event, group, or nation to isolate factors that explain patterns; most often cross-national comparisons.

Cost analysis. Most often, classical costs-benefits, cost-effectiveness, cost-utility, cost-feasibility, return on investment analyses, and financial ratio analyses; rarely considers non-monetary and other types of costs.

Expert panels (internal). Expert panel evaluations of research can be seen as the result of the meeting of traditional (micro-level) peer review with the growth of, and demand for evaluation

in public policy; in contrast to traditional peer review it aims at assessments of research on the meso-level (the institutional level) and the macro-level (the national level), whereas traditional peer review makes assessments at the micro-level (single manuscripts, applications or applicants); internal expert panels consists only of experts within the country/nation.

Expert panels (external). Expert panel evaluations of research can be seen as the result of the meeting of traditional (micro-level) peer review with the growth of, and demand for evaluation in public policy; in contrast to traditional peer review it aims at assessments of research on the meso-level (the institutional level) and the macro-level (the national level), whereas traditional peer review makes assessments at the micro-level (single manuscripts, applications or applicants); external expert panels consists only of experts outside the country/nation.

Expert panels (mixed). Expert panel evaluations of research can be seen as the result of the meeting of traditional (micro-level) peer review with the growth of, and demand for evaluation in public policy; in contrast to traditional peer review it aims at assessments of research on the meso-level (the institutional level) and the macro-level (the national level), whereas traditional peer review makes assessments at the micro-level (single manuscripts, applications or applicants); mixed expert panels consists of both internal and external experts.

Interviews. A conversation between two or more people where questions are asked by the interviewer to obtain information from the interviewee; interviews can be divided into two general types, interviews of assessment and interviews for information.

Observations. Observations are usually conducted by auditors or expert panels; observers do not normally interact with those being observed; usage varies; often a supplement to other methods.

Self-evaluations. Evaluating and reporting on the quality or value of one's own work; often a supplement to other methods.

Site visits. Site visits are usually conducted by auditors or expert panels; unlike observations, observers interact with those being observed; usage varies; often a supplement to other methods.

Strategic plans. Analysis of an individual's, group's, project or program's, or institution's strategic research plans; sometimes used to set performance targets or standards; often a supplement to other methods.

Surveying. Sampling from a population in order to make inferences about the population; usually in the form of questionnaires, less often in the form of interviews; sometimes a census of an entire population; usage varies; often a supplement to other methods.

Clearly, the most commonly employed methodology is the expert panel (see Table 3). Every country in the sample uses at least one of the varieties of expert peers; 31% using primarily internal peers; 19% using primarily external peers; and 50% using primarily mixed-peer panels. Nearly half (44%) also use some form of self-evaluation. Other emerging techniques include (social) network analysis, spillover analysis, and data mining and visualisation, for example (Coryn & Scriven, 2007b).

Table 3. International research evaluation models' core methods

	AU	BE	CZ	DE	FI	FR	HK	HU	IE	JP	NL	NZ	PL	SE	UK	US
Biometrics	✓	✓						✓			✓					
Case studies				✓	✓									✓		✓
Comparative studies		✓	✓	✓	✓											✓
Cost analyses		✓				✓			✓							✓
Expert panels (internal)	✓		✓	✓											✓	✓
Expert panels (external)					✓	✓					✓					
Expert panel (mixed)		✓					✓	✓	✓	✓		✓	✓	✓		
Interviews					✓					✓	✓					
Observations																
Self-evaluations				✓	✓	✓		✓		✓	✓					✓
Site visits				✓	✓						✓			✓		
Strategic plans			✓	✓	✓		✓									✓
Surveying														✓		

Key indicators and criteria

A large majority of national systems are driven by a cluster of quality indicators and criteria, which usually include one or more of the following:

Patents. Patent applications and patents granted by EPO, USPTO, and JPO; frequently viewed as indicators of innovation.

Local or regional impact. Impact of research at a local or regional level; estimated using bibliometric techniques or peer or expert panel assessment, in most cases.

National impact. Impact of research at a national level; estimated using bibliometric techniques or peer or expert panel assessment, in most cases.

International impact. Impact of research at an international level; estimated using bibliometric techniques or peer or expert panel assessments, in most cases.

Researchers. Professionals engaged in the conception or creation of new knowledge, products, processes, methods and systems, and in the management of the projects concerned.

Students. Students enrolled in research-related programs; sometimes students enrolled in any program of study.

Degrees awarded. Students completing research-related programs of study; usually at the doctoral level.

External research funding. Research funding received from non-governmental sources (e.g. private sector).

Esteem. Awards, keynote speeches and addresses, journal editorships, and similar indicators.

Research inputs. Equipment, staff, funding, and other relevant inputs.

Research outputs. All varieties of research outputs, including, but not limited to scholarly publication, products, and patents.

Research process. Everything that occurs prior to research outputs; includes for example, vision, design, planning, operation, justification (e.g. of goals), fidelity, management, activities, and procedures.

By far, most national systems place the greatest emphasis on the impacts of research (see Table 4); in particular international impact (by 100% of the sampled countries). The way in which these impacts are estimated, however, varies widely (e.g. bibliometrics, peer review). Research outputs are also commonly used as quality indicators (by 81% of the sampled countries); yet, sometimes in reference to quantity rather than quality. Economic indicators, such as GERD, BERD, and GBAORD, have not been included here as most countries typically monitor these data for policy decisions regarding research expenditures.

Table 4. International research evaluation models' key indicators and criteria

	AU	BE	CZ	DE	FI	FR	HK	HU	IE	JP	NL	NZ	PL	SE	UK	US
Patents		✓									✓					
Local impact	✓						✓				✓	✓			✓	
Regional or national impact	✓			✓	✓	✓	✓		✓		✓	✓		✓		✓
International impact	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Researchers	✓		✓			✓	✓				✓	✓			✓	
Students	✓	✓		✓	✓		✓			✓	✓		✓		✓	
Degrees awarded	✓			✓	✓		✓			✓	✓	✓	✓	✓	✓	
External research funding	✓			✓	✓						✓	✓			✓	
Esteem			✓			✓				✓	✓	✓			✓	
Research inputs	✓				✓			✓	✓		✓					✓
Research outputs	✓	✓	✓	✓		✓	✓	✓		✓	✓	✓	✓	✓		✓
Research process										✓		✓	✓	✓		✓

Model classification

Most national-level research evaluation models can be considered in terms of two general types, which are (Campbell, 2002):

Type A Type A research evaluation systems apply an approach which is systemic and consistent

Type B Type B research evaluation systems use pluralised approaches, and can be characterised by a high degree of situation-specific variability in terms of their conceptions and methods

As shown in Table 5, 37% (6 of 16) were classified as Type A systems versus 63% (10 of 16) being classified as Type B systems. However, many of these national systems are considered experimental, being reformed, or currently under development, making them difficult to correctly classify. In such cases, these models were placed in the Type B category as they cannot be considered either systematic or consistent.

Table 5. International research evaluation models' systemisation and consistency

	AU	BE	CZ	DE	FI	FR	HK	HU	IE	JP	NL	NZ	PL	SE	UK	US
Type A	✓						✓				✓	✓			✓	✓
Type B		✓	✓	✓	✓	✓		✓	✓	✓			✓	✓		

Funding system archetypes

Another useful way to conceptualize and classify the various international systems is by their research funding system models, or archetypes, of which there are three major categories (Coryn, 2007; Coryn, 2008; Coryn, Hattie, Scriven & Hartmann, 2007):

Type I Large scale performance exercises of various hues; future funding allocations are made on the basis of prior performance; sometimes used in conjunction with Type II and III models.

Type II Bulk funding models; generally block grant allocations of research funds; sometimes a mix of direct funding for public research institutions and universities and competitive grants programs offered by independent funding agencies.

Type III Indicator-driven mechanisms; research financing is distributed on the basis of student numbers, external funding, teaching volume, and other quantifiable measures via various funding formulas.

Not considered in this classification, however, is the centralised versus decentralised, or mixed systems for funding research. Most countries have centralised research funding mechanisms (i.e., research funding comes from one government agency). Belgium and the United States, however, are decentralised in that multiple agencies or government branches fund a large portion of the countries' research. In any case, 31% (5 of 16) were classified as Type I models, 44% (7 of 16) as Type II models, and the remaining 25% (4 of 16) as Type III models (see Table 6). Although the Netherlands' model was classified as Type I, this exercise presently has no connection with the level of funding received, but is in force to improve the public accountability of research activity.

Table 6. International funding system archetypes

	AU	BE	CZ	DE	FI	FR	HK	HU	IE	JP	NL	NZ	PL	SE	UK	US
Type I							✓				✓	✓			✓	✓
Type II		✓	✓			✓		✓	✓	✓				✓		
Type III	✓			✓	✓								✓			

Quality of the national models

In a study of the quality of the 16 countries' models for evaluating and financing research (Coryn, 2007; Coryn, 2008; Coryn, Hattie, Scriven, & Hartmann, 2007; Coryn & Scriven, 2007a), a descriptive discriminate function analysis was used to assess the degree to which dimensions of validity, credibility, utility, cost-effectiveness, and ethicality discriminated between Type I, Type II, and Type III funding systems. There was a large canonical correlation (R_c = .84) on Function 1 with an effect size of R_c^2 = 70.39% between the grouping variable (Type I, Type II, and Type III) and the composite predictor variables (validity, credibility, utility, cost-effectiveness, and ethicality). The full model test for the function was significant; ^ = .28, $X2(10, 32) = 34.49$, $p < .01$. However, as shown in Table 7, the test of Function 2 (i.e., discrimination between Type II and Type III models) was not significant and therefore excluded from subsequent analyses. The means and standard deviations for each of the three types of models on the five metadimensions are presented in Table 8.

Table 7. Wilk's lambda and canonical correlations for the three funding system archetypes

Function	**^**	***X²***	***df***	***p***	***Rc***	***Rc²***
1-2	.279	34.49	10	.00	.84	70.39%
2	.940	1.68	4	.79	.25	6.05%

Table 8. Means and standard deviations on the five metadimensions for the three funding system archetypes*

Metadimension	**Type I**		**Type II**		**Type III**	
	M	**SD**	**M**	**SD**	**M**	**SD**
Validity	73.00	11.11	41.35	8.12	36.50	21.21
Credibility	74.00	12.40	44.07	5.69	37.00	17.59
Utility	71.40	10.45	42.57	4.92	33.50	20.77
Cost-effectiveness	68.60	13.92	41.85	7.37	34.25	15.02
Ethnicity	63.80	14.97	38.42	8.98	33.25	11.94

* The possible range of weighted scores on any dimension was from 0-100, or 0%-100%

Standardised discriminant function coefficients and structure coefficients were examined to determine which of the dimensions contributed to the differences in the three types of models. As shown in Table 10, validity emerges as the dimension most correlated with the grouping variable (i.e., type of model) on Function 1, meaning that it contributes the most to separation of the models. The group centroids showed Type I models (group centroid = 2.13) being substantially higher on the composite dimensions than Type II (group centroid = -0.69) and Type III models (group centroid = -1.47). This and the structure coefficients indicate that the differences (i.e., separation) observed on Function 1 can be attributed mostly to validity, and to some extent credibility, utility, cost-effectiveness, and ethicality given that these were all positively correlated in the function. Therefore, Type I models have more of these traits (validity, credibility, utility, cost-effectiveness, and ethicality) than either Type II or III models in the linear equation.

Moreover, using the quality categories shown in Table 9, very few of the national models met the minimum threshold for being assigned to a quality category greater than F (see Table 10) as judged by two multidisciplinary panels of researchers and evaluators.[1]

Table 9. Quality category descriptions

Quality category	Description
A	Excellent; clear example of exemplary performance; no deficiencies
B	Very good; strong overall but not exemplary; no real deficiencies of consequence
C	Good; reasonably good overall; minor but non-fatal deficiencies
D	Satisfactory; barely adequate; several serious deficiencies
F	Absence of merit; clearly inadequate; fatal deficiencies

Table 10: Country quality categories and scores

(AU = Australia; BE = Belgium; CZ = Czech Republic; FI = Finland; FR = France; DE = Germany; HK =Hong Kong; HU = Hungary; IE = Irland; JP = Japan; NL = Netherlands; NZ = New Zealand; PL = Poland; SE = Sweden; UK = United Kingdom; and US = Untied States)*

	AU	BE	CZ	DE	FI	FR	HK	HU	IE	JP	NL	NZ	PL	SE	UK	US
Category	D/F	F	F	F	F	F	D	F	F	F	C/D	A/B	F	F	C	D
Score	48/52	40/40	38/43	46/47	36/31	11/11	58/51	36/40	42/38	43/42	72/66	85/90	46/44	46/42	73/70	68/69

* The possible range of total weighted scores was from 0-100, or 0%-100%

Concluding remarks

In most countries, the competition for government research monies is getting increasingly competitive, which is particularly evident in systems that operate on performance-based funding (Type I models). Methodologically, large-scale research evaluations of government-financed research are most often binary in nature. That is, they are normally either a variant of traditional peer review (e.g. expert panels of one type or another) or are driven by indicators (e.g. publications, external funding). Both approaches have strengths and weaknesses. The indicator method, however, encourages the "moral hazard" or undue focus on productivity or assessment benchmarks, diverting attention away from "more academically useful research into tactics for cultivating citations," for example (von Tunzelmann & Mbula, 2003: 15).

As illustrated by the national systems presented in this paper, research evaluation as conducted throughout the world can be characterised by increasing levels of size and complexity. However, most countries still regard their systems as experimental. Moreover, there is a near world-wide interest in the United Kingdom model, which has become a "benchmark for research evaluation" (von Tunzelmann & Mbula, 2003: 6) – although the RAE will soon move to a more cost-effective, metrics-based system. Conversely, there has been some suggestion that the United Kingdom's RAE does not itself lead to enhancements in the quality of research in the United Kingdom, but does encourage universities and departments to compete with one another, for example, "by [universities and departments] bidding to attract star researchers in order to improve their record of achievement" (Barker & Lloyd, 1997: 56).

In New Zealand, several concerns have been raised in reference to the Performance-Based Research Fund (PBRF), introduced in 2003 (Curtis & Matthewman, 2005). Among them is the real cost-benefit ratio of participation, with reports that many universities have spent more on the exercise than they will gain in funding increases (Nature, 2006). Questions have also arisen as to whether the quality of research has improved as a direct result of the assessment (Coryn, 2007c). The PBRF scoring system has received the most criticism and, after the latest assessment (2006), the controversial unit of assessment will be reviewed.

Note

1. A detailed discussion of the methodology used to arrive at these conclusions exceeds the scope of this paper. For a detailed presentation please see Coryn (2007a, b, c, 2008) and Coryn, Hattie, Scriven and Hartmann (2007).

References

ab Iorwerth, A. (2005). *Methods of Evaluating University Research Around the World*. Ottawa, Canada: Department of Finance.

Aksnes, D. W. (2005). *Citations and their Use as Indicators in Science Policy: Studies of validity and applicability issues with a particular focus on highly-cited papers*. Unpublished doctoral dissertation, University of Twente, Enschede, The Netherlands.

Barker, D., & Lloyd, P. (1997). Evaluation of Scientific Research in the United Kingdom. In OECD (Ed.), *The Evaluation of Scientific Research: Selected experiences* (pp. 47–58). Paris, France: Organization for Economic Co-Operation and Development.

Campbell Collaboration. (2006). *About the Campbell Collaboration*. Retrieved February 21, 2006, from http://www.campbellcollaboration.org/About.asp

Campbell, D. F. J. (2002). *Conceptual Framework for the Evaluation of University Research in Europe* (Working Paper). Washington, DC: Center for International Science and Technology Policy, The George Washington University.

Campbell, D. F. J., & Felderer, B. (1997). *Evaluating Academic Research in Germany: Patterns and policies* (Political Science Series, No. 48). Wein, Austria: Institut für Höhere Studien.

Campbell, P. (1997). Asian Tiger Claws its Way up the Ratings. *Nature*, *388*, 9–10.

Cochrane Collaboration. (2006). *The Cochrane Library: An introduction*. Retrieved February 21, 2006, from http://www.cochrane.org/reviews/clibintro.htm

Coryn, C. L. S. (2006a). The Validity, Utility, Credibility, Cost-effectiveness, and Ethicality of Research Evaluation Systems: A comparative analysis of seventeen national-level systems. Paper presented at the meeting of the European Evaluation Society/United Kingdom Evaluation Society, October, London, UK.

Coryn, C. L. S. (2006b). A Comparative Analysis of Seventeen National-Level Research Evaluation Systems. Paper presented at the meeting of the American Evaluation Association, November, Portland, Oregon.

Coryn, C. L. S. (2007a). *Evaluation of Researchers and their Research: Toward making the implicit explicit*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo.

Coryn, C. L. S. (2007b). I Think … Therefore, I Need Funding. Paper presented at The Evaluation Center's Evaluation Café series, Kalamazoo, Michigan.

Coryn, C. L. S. (2007c). Models and Mechanisms for Evaluating Publicly-Funded Research: An international comparison with an emphasis on New Zealand's PBRF. Paper presented to the Tertiary Education Commission, Wellington, New Zealand.

Coryn, C. L. S. (2008). *Models for Evaluating Scientific Research: A comparative analysis of national systems*.

Saarbrücken, Germany: VDM Verlag.

Coryn, C. L. S., Hattie, J. A., Scriven, M., & Hartmann, D. J. (2007). Models and Mechanisms for Evaluating Government-Funded Research: An international comparison. *American Journal of Evaluation*, *28*(4), 437–457.

Coryn, C. L. S., & Scriven, M. (2007a). Are National-Level Research Evaluation Models Valid, Credible, Useful, Cost-effective, and Ethical? *Journal of MultiDisciplinary Evaluation*, *4*(8).

Coryn, C. L. S., & Scriven, M. (Eds.). (2007b). *Reforming the Evaluation of Research: New Directions for Evaluation*. San Francisco, CA: Jossey-Bass.

COSEPUP. (1999). *Evaluating Federal Research Programs: Research and the Government Performance and Results act*. Washington, DC: National Academy Press.

COSEPUP. (2001). *Implementing the Government Performance and Results Act for Research: A status report*. Washington, DC: National Academy Press.

Cozzens, S. E., & Turpin, T. (2000). Processes and Mechanisms for Evaluating and Monitoring Research Outcomes from Higher Education: International comparisons. *Research Evaluation*, *8*(1), 3–4.

Curtis, B., & Matthewman, S. (2005). The Managed University: The PBRF, its impacts and staff attitudes. *New Zealand Journal of Employment Relations, 30*(2), 1–17.

DEST. (2003). *Australian Science and Technology at a Glance 2003*. Retrieved May 17, 2006, from http://www.dest.gov.au/NR/rdonlyres/74F1B2A0-68C8-4270-BEC7-DF9C92D28A5F/2073/at_a_Glance_2003.pdf

Frankel, M. S., & Cave, J. (Eds.). (1997). *Evaluating Science and Scientists: An East-West dialogue on research evaluation in post-communist Europe*. Budapest, Hungary: Central European University Press.

Frederiksen, L. F., Hannson, F., & Wennberg, S. B. (2003). The *Agora* and the Role of Research Evaluation. *Evaluation: The international journal of theory, research and practice*, *9*(2), 149–172.

Geuna, A., & Martin, B. R. (2001). *University Research Evaluation and Funding: An international comparison* (Electronic Working Paper Series, Paper No. 71). Sussex, England: University of Sussex, Science and Technology Policy Research.

Geuna, A., & Martin, B. R. (2003). University Research Evaluation and Funding: An international comparison. *Minerva*, *41*, 277–304.

Glass, G. V. (1976). Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher*, *5*, 3–8.

Hansson, F. (2006). Organizational Use of Evaluation: Governance and control in research evaluation. *Evaluation: The international journal of theory, research and practice*, *12*(2), 159–178.

Laudel, G. (2006). The 'Quality Myth': Promoting and hindering conditions for acquiring research funds. *Higher Education*, *52*, 375–403.

Luukkonen, T. (2002). Research Evaluation in Europe: State of the art. *Research Evaluation*, *11*(2), 81–84.

Meyer-Krahmer, F., & Reiss, T. (1992). Ex ante Evaluation and Technology Assessment: Two emerging elements of technology policy evaluation. *Research Evaluation*, *2*(1), 47–54.

Moed, H. F. (2005). *Citation Analysis in Research Evaluation*. Dordrecht, The Netherlands: Springer.

Nature. (2006). Editorial: Evaluate This: The objective evaluation of research isn't working as it should. *Nature*, *440*, 1–2.

OECD. (1987). *Evaluation of research: A selection of current practices*. Paris, France: Organisation for Economic Co-Operation and Development.

OECD. (1997). *The Evaluation of Scientific Research: Selected experiences*. Paris, France: Organisation for Economic Co-Operation and Development.

OECD. (2003). *Governance of Public Research: Country case studies*. Paris, France: Organisation for Economic Co-Operation and Development.

Orr, D., & Paetzold, M. (2006). Procedures for Research Evaluation in German Higher Education: Current fragmentation and future prospects. *ACCESS: Critical Perspectives on Communication, Cultural & Policy Studies*, *25*(2), 16–30.

Pawson, R. (2006). *Evidence-Based Policy: A realist perspective*. London, England: Sage. RAE 2008. (2005). *RAE 2008*. Retrieved January 22, 2006, from http://www.rae.ac.uk/

Rogers, P. J. (2005). Accountability. In S. Mathison (Ed.), *Encyclopedia of Evaluation* (pp. 2–4). Thousand Oaks, CA: Sage.

Rousseau, R. (2004). Impact factors and databases as instruments for research evaluation. Paper presented at the 4th International Conference on University Evaluation and Research Evaluation, October, Wuhan, China.

Russell, J. M., & Rousseau, R. (2002). Bibliometrics and Institutional Evaluation. In UNESCO (Ed.), *Encyclopedia of Life Support Systems: Part 19.3, science and technology policy* (pp. 1–20). Paris, France: UNESCO.

Saegusa, A. (1999a). Japanese Labs Balk at Bid to Boost External Evaluation. *Nature*, *397*, 378.

Saegusa, A. (1999b). …And to use External Scrutiny to Increase Competitiveness as Universities. *Nature*, *399*, 511.

Salmon, W. C. (1998). Rational Prediction. In M. Curd & J. A. Cover (Eds.), *Philosophy of Science: The central issues* (pp. 433–459). New York, NY: Norton.

Scherer, F. M. (1967). Research and Development Resource Allocation under Rivalry. *The Quarterly Journal of Economics*, *81*(3), 359–394.

Scriven, M. (1967). The Methodology of Evaluation. In R. Tyler, R. Gagne & M. Scriven (Eds.), *Perspectives of Curriculum Evaluation* (pp. 39–83). Chicago, IL: Rand-McNally.

Scriven, M. (2006). The Evaluation of Research Merit Versus the Evaluation of Funding of Research. *Journal of MultiDisciplinary Evaluation*, *5*, 120–123.

Swinbanks, D., Nathan, R., & Triendl, R. (1997). Western Research Assessment Meets Asian Cultures. *Nature*, *389*, 113–117.

van Raan, A. J. F. (2005). Fatal Attraction: Conceptual and methodological issues problems in the ranking of universities by bibliometric methods. *Scientometrics*, *62*(1), 133–143.

von Tunzelmann, N., & Mbula, E. K. (2003). *Changes in Research Assessment Practices in Other Countries Since 1999: Final report*. Brighton, England: University of Sussex, SPRU Science and Technology Policy Research.

Weinberg, A. M. (1963). Criteria for Scientific Choice. *Minerva*, *1*, 159–171.

Weinberg, A. M. (1989). Criteria For Evaluation, a Generation Later. In D. Evered & S. Harnett (Eds.), *The Evaluation of Scientific Research* (pp. 3–15). New York, NY: John Wiley & Sons.