Chapter 3

Noneconomic Quantitative Measures: The “Output” of Science

chapter 3

Noneconomic Quantitative Measures: The "0utput" of Science

Having identified severe drawbacks to the use of econometric models to evaluate Federal R&D, OTA looked elsewhere for objective quantitative measures. The only quantitative approach to the evaluation of research output is bibliometrics, which analyzes scholarly publications for indications of quantity and quality. The underlying assumption of this approach is that knowledge is the essential product of research and publications are the most readily identified manifestations of that knowledge. With a gradually evolving methodology, bibliometricians have attempted to measure objectively the quantity and quality of research results. They have achieved some success in comparing projects within a discipline, and less in comparing disciplines. Bibliometric analysis does not address the most important policy question: how to compare the value of Federal research with other Federal programs.

A considerable amount of quantitative information about the U.S. science and engineering enterprise is published regularly by the National Science Foundation (NSF), the National Institutes of Health (NIH), and the National Research Council (NRC) in their reports on funding, personnel, degree attainment and graduate education. Every 2 years NSF publishes a 300-page compilation of this information, Science *Indicators.* Science indicators could be used to provide a rough measure of the health of the research enterprise in the United States if some agreement could be achieved in the science policy community about which of the thousands of numbers published by NSF are most relevant to that task. The use of science indicators to measure the quality of the research process in the United States is discussed later in this chapter.

BIB BIOMETRICS

The quantitative analysis of scientific publications is in its second generation. The first generation, spurred by Eugene Garfield's founding of the *Science Citation Index* and Derek de Solla Price's efforts, [1] explored the feasibility of understanding science through its literature alone. Price boldly named this approach the "science of science" and published demonstrations of its heuristic, if not immediate policy, value, [2]

The second generation, now a decade old, sought to develop and exploit publication and citation data as a tool for informing decisionmakers, especially in Federal agencies and universities.[3] This current generation has many of the features 01 an

For a first person retrospective, see Eugene Garfield, ***Essays of an Information Scientist, vol.*** 1, **1962-73; vol. 2, 1974-76 (Philadelphia,** PA: ISI Press, 1977). For examples, see Eugene Garfield, et al., The ***Use of Citation*** **Data** ***in Writing the History of Science*** (Philadelphia, PA: Institute for Scientific Information, 1964); Derek de Solla Price, "Networks of Scientific Papers," Science, vol. 149, July 30, 1965, pp. **510-515; Derek de** Solla Price, "Is Technology Historically independent of Science? A Study in Statistical Historiography, " ***Technology and Culture, vol. 6,*** fall 1965, pp. **553-568.**

'Derek de Solla Price, *Little* ***Science, Big*** *Science* (New York Columbia University Press, 1963).

[3]For reviews, see Yehuda Elkana, et al. (eds.), *Toward A Metric of* Science: ***The Advent of Science Indicators (New*** York. John Wiley & Sons, 1978); Francis Narin, "Objectivity Versus Relevance in Studies of Scientific Advance, " *Scientometrics,* ***vol. 1,*** September 1978, pp. 35-41. The use of projected citation data in a controversial promotion and tenure case is described in N.L. Geller, et al., "Lifetime-Citation Rates to Compare Scientists Work, " Social ***Science Research,*** vol. ***7, 1978,*** pp. 345-305

institutionalized scientific specialty: multidisciplinary journals and practitioners, a clientele (both consumers and patrons), and numerous claims to the efficacy of "bibliometrics" as a policy tool.[4] The quantitative analysis of scientific publications has arguably established its place in the evaluation of research outcomes and as an input both to the allocation of resources for research and to the expectation that the growth of scientific knowledge can be measured, interpreted, and indeed, manipulated.

This chapter focuses on the second generation of noneconomic quantitative measures of scientific research results and evaluates its usefulness to policymakers. The chapter assesses the most promising approaches and methods that have been employed and suggests how quantitative data and models could be refined to augment decisionmaking processes in science.

The First Generation of Bibliometrics (1961-74)

The pioneers of bibliometrics searched for ways to understand science independent of the scientists themselves. First-person accounts, questionnaires, and historical narratives all require some form of cooperation or consent of the scientists involved. This dependence on self-interest sources could bias the results. Bibliometric pioneers of the early 1960s saw a need first to reconstruct, then to monitor and predict, the structure and products of science. Eugene Garfield and Derek de Solla Price talked about "invisible colleges" and the tracing of "intellectual influence" as a mirror held up to science, imperfect but public, using the formal communication system of science. Science literature could be studied—without recourse to the authors—to open new vistas, both practical and analytical, once it was cataloged, indexed and made retrievable.

With the creation of the Science *Citation Index* (SCI), the scientific literature became a data source for the quantitative analysis of science. It generated both the concepts and measurement techniques that formed the bedrock of bibliometrics.[5] These include the principal units of analysis: publications (papers, articles, journals), citations (bibliographic references), and their producers (individual authors and collaborators in teams). When subjected to the primary methods of analysis—counting, linking, and mapping—these units yield measures of higher order concepts: coherent social groups, theory groups, networks, clusters, problem domains, specialties, subfields, and fields.

Computers aided the increasingly sophisticated manipulation of documents in the growing SCI database. Journal publications could be counted by author, but also aggregated into schools of

[4] Cofounded in 1978 by Garfield and Price, *Scientometrics* became the flagship journal of bibliometrics. Its contributors seem to come primarily from information science, psychology, and sociology. Other spurs to the institutionalization and visibility of bibliometrics has been, since 1972, the National Science Board's biennial *Science* Indicators series and the ongoing work of the Institute for Scientific Information (especially Henry Small) and Francis Narms Computer Horizons, Inc. (discussed below). For historical perspectives on the development of this specialty, see Daryl E. Chubin, "Beyond Invisible Colleges: Inspirations and Aspirations of Post-1972 Social Studies of Science," *Scientometrics, vol. 6, 1985,* pp. 221-254; Daryl E. Chubin and S. Restive, "The 'Mooting' of Science Studies: Strong Programs and Science Policy," in K.D. Knorr-Cetina and M. Mulkay (eds.), Science Observed (London and Beverly Hills, CA: Sage, 1983), pp. *58-83. Also see* the special issue of *Scientometrics, vol. 6, 1985,* dedicated to the memory of Derek Price.

[5] These are touted, debated, and assailed in Daryl E. Chubin, "The Conceptualization of Scientific Specialties," *The Sociological Quarterly, vol. 17,* autumn 1976, pp. *448-476;* Daryl E. Chubin, "Constructing and Reconstructing Scientific Reality: A Meta-Analysis," *International Society for the Sociology of Knowledge Newsletter, vol. 7,* May 1981, pp. 22-28; Susan E. Cozzens, "Taking the Measure of Science: A Review of Citation Theories," *ISSK Newsletter, vol. 7,* May 1981, pp. 16-21; D. Edge, "Quantitative Measures of Communication in Science: A Critical Review," *History of Science vol. 17, 1979,* pp. *102-134;* and in various chapters in Elkana, et al., op. cit.

thought or whole institutions.[b] Consistent and influential contributors to the literature could be identified by co-citations (the number of times two papers are cited in the same article) and separated from occasional authors. The resultant co-citation clusters could be depicted as a "map of science" for a given year showing the strength of links within clusters and the relations, if any, among them.[7]

By the mid-1970s, bibliometricians were constructing structural and graphical maps of the domains and levels of research activity in science. Further, they were comparing these pictures to other accounts, built on biographic and demographic information, informal communication, and other informant-centered data, to depict how research communities—their research foci, intellectual leaders, and specialized journals-change over time. They thus offered a more comprehensive perspective on the growth of knowledge, at least in terms of its outputs, than was ever previously available." Analysts differed in their interpretation and application of the data, and the life of the community responsible for the outputs tended to remain unobserved. Nevertheless, bibliometric analysis began to offer the promise of the independent baseline implied in Price's phrase "the science of science."[9]

The Second Generation (1975=85)

The legacy of the first generation was the promise of its scholarly literature. The second generation has attempted to deliver on the promise that bibliometric analysis could be predictive and reliable for decisionmaking. That promise has yet to be fulfilled, for reasons that will be discussed below. However, there is growing evidence that the quantitative assessment of science warrants the attention it is now receiving from policymakers both in the United States and Europe.

The analysts discussed below have used bibliometrics to anticipate the source of "greatest contributions" and identify promising research projects. They produce policy-relevant documents and recognize intervention decisions as a desirable consequence of their work. Several governments have funded their efforts. A look at the leading

[b]Seminal work here is D. Crane, ***Invisible Colleges: Diffusion of Knowledge in Scientific Communities*** (Chicago and London: University of Chicago Press, 1972); B.C. Griffith and 14. C. Mullins, "Coherent Social Groups in Scientific Change," ***Science vol. 177***, Sept. 15, 1972, pp. ***959-964;*** N.C. Mullins, 'The Development of a Scientific Specialty: The Phage Group and the Origins of Molecular Biology," ***Minerva, vol. 10, 1972,*** pp. 52-82; N.C. Mullins, ***Theory and Theory Groups in Contemporary American Sociology (New*** York: Harper Row, ***1973);*** **and Garfield, op. cit., "Corporate Index" that** lists publications by institution of author.

[7]The methodological groundwork for co-citation analysis is presented in B.C. Griffith, et al., "The Structure of Scientific Literatures II: Toward a Macro- and Microstructure for Science," ***Science Studies,*** vol. 4, 1974, pp. 339-365; H.G. Small, "Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents," *Journal of the* ***American Society for Information Science,*** vol. 24, 1973, pp. 265-269; H.G. Small, "Multiple Citation Patterns in Scientific Literature: The Circle and Hill Models," ***Information Storage and Retrieval, vol.*** 10, 1974, pp. 393-402; H.G. Small and B.C. Griffith, "The Structure of Scientific Literatures I: Identifying and Graphing Specialties," Science ***Studies, vol.*** 4, 1974, pp. 1740.

[8]Noteworthy illustrations are discussed in Daryl E. Chubin, 'The Conceptualization of Scientific Specialties," op. cit., and G.N. Gilbert, "Measuring the Growth of Science: A Review of Indicators of Scientific Growth," *Scientometrics,* vol. 1, September 1978, pp. 9-34.

[9]For example, in 1969, Price's "Measuring the Size of Science," *Proceedings of the Israel Academy of Sciences and Humanities*, vol. 4, 1969, pp. 98-111, tied national publication activity to percent of GNP allotted to R&D. By 1975, F. Narin and M. Carpenter ("National Publication and Citation Comparisons," *JASIS*, vol, 26, pp. 80-93) were computing shares, on a nation-by-nation basis, of the world literature, and characterizing interrelations among journals (Francis Narin, et al., "Interrelationship of Scientific Journals," *JASIS* vol. ***23, 1972,*** pp. ***323-331), as*** well as the content of the literature in broad fields (Francis Narin, et al., "Structure of the Biomedical Literature," *JASIS*, vol. ***27, 1976,*** pp 25-45). These analyses employed algorithms for tallying, weighing, and linking keywords in article titles to citations aggregated to journals and authors nation - of-affiliation at the time of publication. Some would call this methodology "crude"; others would herald its sophistication for discerning patterns in an otherwise massive and perplexing literature The latter is preciseiy the mentality guiding the *Science Indicators* volumes and foreshadowed in two other pioneering papers of the first generation: Eugene Garfield, "Citation Indexing tor Studying Science," *Nature*, vol. ***227, 1970,*** pp. ***659-671;*** Derek de Solla Price, "Citation Measures of Hard Science, Soft Science, Technology and Non-science," ***Communication Among Scientists and*** *Engineers,* ***C.*** Nelson and D. Pollock (eds.) (Lexington, MA: D.C. Heath, 1970), pp. 3-22.

practitioners will reveal how they approach the question:

> How can we **characterize** the effects of decisions about funding programs as they reverberate into the various levels of the scientific community: up from "fields" into disciplines and down from "fields" into research areas or teams?[10]

Francis Narin of Computer Horizons, Inc., is the veteran performer, linking the two generations. His computerized approach is based on the components of the Science *Citation Index* and used in conjunction with other data, such as the National Library of Medicine's MEDLINE and the NIH in-house grant profile system, Information for Management Planning, Analysis, and Coordination (IMPAC). Although Narin's work tends toward the macroscopic, its manipulations have grown more sophisticated in their capability of addressing micro-level questions. Narin's methodology answers quantitatively the following kinds of questions:

- Are articles published in basic journals referenced in clinical and practitioner journals [these types derive from Narin's own classification of article content in journals]?
- Is there a relationship between priority scores on research applications and number of articles produced and citations received?
- Are grants to medical schools more productive than grants to academic departments?
- Are young researchers more productive than older researchers?
- Is the return on investment mechanism [investigator-initiated proposals] more productive than other support mechanisms?
- How often do National Institute on Drug Abuse, National Institute on Alcohol Abuse and Alcoholism, and other NIH-supported researchers cite work supported by the National Institute of Mental Health?

A common criticism of Narin's work is that it is too descriptive and relies on ad hoc explanation for the observed patterns and trends. Some feel it is excessively dependent on a literature baseline and does not reflect an understanding of the sciences it appraises. In a current project sponsored by the National Cancer Institute (NCI), "An Assessment of the Factors Affecting Critical Cancer Research Findings, " Narin consciously tries to remedy the problem by working closely from the outset with a panel of cancer researchers. He is tracing key events through participant consensus, the historical record, and various bibliometric indicators. Discrepancies are apparently negotiated as the project unfolds, though the exact negotiation procedure is not specified .11

Another departure for Narin stems from his acquisition and computerization of U.S. Patent Office case files that will permit mapping of literature citations in patents at the national, industry, and inventor levels. An infant literature has crystallized around the notion of "technology indicators" with patents signifying the conversion of knowledge into an innovation with commercial and social value—another tangible return on investment. [12]

Irvine and Martin's (Science Policy Research Unit, University of Sussex, UK) evaluation program in "converging partial indicators" has gained attention for three important reasons:

1. They claim to assess the basic research performance of large technologydependent facilities, such as the European Organizations for Nuclear Research (CERN) accelerator and the Isaac Newton Telescope.
2. They have made cost-effectiveness the central performance criterion in their input-output scheme.
3. Their "triangulation" methodology is an impressive codification of many separate proce-

---

'Susan E. Cozzens, "Editor's Introduction, " m "Funding and Knowledge Growth, " Theme Section, *Social* Studies *of Science*, vol. 16, February 1986, forthcoming (quote from mimeo version, p. 9).

"The objective of this project is to estimate knowledge returns from the U.S. war on cancer. What has been the extent and character of NCI funding in the cancer literature: are highly cited papers and **authors supported by NCI grants and contracts?** More on this genre of study is presented below and in other chapters of this technical memorandum, but see Francis Narin and R.T. Shapiro, "The Extramural Role of the NIH as a Research Support Agency, " ***Federation Proceeding,* vol. *36,*** October 1977, pp. 2470-2475.

[12]M.P. Carpenter, et al., "Citation Rates to Technologically Important Patents, " ***World Patent Information, vol. 3, 1981,*** pp. 161-163; and various case study reports on patent activity emanating from Battelle's Pacific Northwest Laboratories, for example, R.S. Campbell and L.O. Levine, ***Technology indicators Based on Citation Data: Three Case Studies,*** Phase II Report prepared for the National Science Foundation Grant, PRA **78-20321 and Contract 2311103578, May 1984**.

dures and measures that have been advocated both by policymakers and analysts. '3

The synopsis presented below is based primarily on a review of Irvine and Martin's articles, four critiques, and a reply.[14] The Irvine and Martin rationale for developing indicators of past research performance is to provide "a means to keep the peer-review system 'honest. ' " Irvine and Martin caution us further "to distinguish between conventional peer-review (involving a small number of referees or 'experts' on a panel) and our extensive peer-evaluations drawing in very large numbers of researchers across different countries and based on structured confidential interviews and attitude surveys. "[15] For the two investigators, conventional grants or journal peer review is but a single indicator; when combined with bibliometric data on research performance and external assessments of the likely future performance of new facilities, a series of multiple indicators is formed. If the indicators converge, Irvine and Martin regard the evaluation results as relatively reliable. [16]

As proxies, partial indicators must stand for a lot that goes unmeasured—by choice or otherwise. Sometimes the interpretive burden is overwhelming (see table 6). No matter how systematic, quantitative, and convergent their findings appear, Irvine and Martin's use of triangulation is problematic, as they admit (see table 7 for a summary):

> The fact that the indicators converge in a given case does not "prove" that the results are 100 percent certain—the indicators may all be "wrong" together. However, if a research facility like the Lick 3-meter telescope produces a comparatively large publication output at fairly low cost, if those papers are relatively highly cited, . . . and if large numbers of astronomers rate it highly in the course of structured interviews, we would place more credibility on the resulting conclusion that this was a successful facility than if the same finding were arrived at by a panel of three or four "experts" without access to the systematic information that we have collected. [17]

If the output measures do not converge, the results become quite problematic. There is no straightforward means of resolution except intuition and judgment.

Tables 7, 8, and 9 present typical samples of the information one can obtain from Irvine and Martin's analyses. Table 7 shows for four different but comparable optical telescopes the average number of papers published per year over the decade 1969-78, the cost per paper, the number of citations to work done on that telescope over the 4-year period 1974-78, the average number of citations per paper, and the number of papers cited 12 or more times. This table was part of a paper that demonstrated that the Isaac Newton Telescope (INT) in Great Britain was more costly and less productive than several comparable facilities. (This was largely due to a political decision to locate the INT at a poor observing site on British soil. Subsequently, it was moved to a more favorable site at La Palma.) The table compares the various facilities in terms of output (papers per year), cost-effectiveness (cost per paper), influence (citations), and significance of scientific work (citations per paper and number of papers cited more than 12 times).

Table 8 presents similar output data for world experimental high energy physics facilities from 1977 to 1980. The table shows, for example, that although the largest number of papers were produced at the CERN proton synchrotrons in 1978, this facility did not have the greatest influence in terms of the number of citations to work done there, nor was it producing the most significant

[13] J. Irvine and B.R. Martin, Foresight *in Science: Picking the Winners* (London: Frances Pinter, 1984). See especially B.R. Martin and J. Irvine, "Assessing Basic Research: Some Partial Indicators of Scientific Progress m Radio Astronomy, " ***Research Policy,*** vol. 12, 1983, **pp. 61-90.**

"The five components are: J. Knge and D. Pestre, "A Critique of Irvine and Martin's Methodology for Evaluating Big Science, " *Social* ***Studies of*** Science, vol. 15, 1985, pp. **425-539;** H.F. Moed and A.F.J. van ***Ram, "Critical*** Remarks on Irvine and Martin's Methodology for Evaluating Scientific Performance, " ***Social Studies of Science, vol.*** 15, 1985, pp. **539-547; R. Bud,** 'The Case of the Disappearing **Caveat: A Critique** of Irvine and Martin's Methodology, " ***Social Studies of Science, vol.*** **15, 1985,** pp. 548-553; H.M. Collins, 'The Possibilities of Science Policy, " ***Social Studies of Science,*** VOI. 15, 1985, pp. **554-558; and** B.R. Martin and J. Irvine, "Evaluating the Evaluators: A Reply to Our Critics, " ***Social Studies of Science, vol.*** **15, 1985, pp. 558-575.** For brevity, quotes from the critics will be noted in the text by (page number) only, those from Martin and Irvine as (IM, page number).

'Ibid., p. 566.

"Ibid., p. 527.

'"Ibid., p. **568**

Table 6.—Main Problems With the Various Partial Indicators of Scientific Progress and Details of How Their Effects May Be Minimized

Partial indicator based on	Problem	How effects may be minimized
A. Publication counts	1. Each publication does not make an equal contribution to scientific knowledge	Use citations to indicate average impact of a group's publications, and to identify very highly cited papers
	2. Variation of publication rates with specialty and institutional context	Choose matched groups producing similar types of papers within a single specialty
B. Citation analysis	1. Technical limitations with ***Science Citation Index:***	
	a. first-author only listed	Not a problem for research *groups*
	b. variations in names	
	c. authors with identical names	Check manually
	d. clerical errors	
	e. incomplete coverage of journals	Not a serious problem for "Big Science"
	2. Variation of citation rate during lifetime of a paper—unrecognized advances on the one hand, and integration of basic ideas on the other	Not a problem if citations are regarded as an Indicator of impact, rather than quality or importance
	3. Critical citations	
	4. "Halo effect" **citations**	
	5. Variation of citation rate with types of paper and specialty	Choose matched groups producing similar types of papers within a single specialty
	6. Self-citation and "in-house" citation (SC and IHC)	Check empirically and adjust results if the incidence of SC or IHC varies between groups
C. Peer	1. Perceived implication of results for own center and competitors may affect evaluation	1. Use a complete sample, or a large representative sample (25°/0 or more)
	2. Individuals evaluate scientific contributions in relation to their own (very different) cognitive and social **locations.**	2. Use verbal rather than written survey so can press evaluator if a divergence between expressed opinions and actual views is suspected
	3. "Conformist" assessments (e.g., "halo effect") accentuated by lack of knowledge of contributions of different centers	3. Assure evaluators of confidentiality 4. Check for systematic variations between different groups of evaluators

SOURCE: B.R. Martin and J. Irvine, "Assessing Basic Research: Some Partial indicators of Scientific Progress in Radio Astronomy, " Research *Policy*, vol 12, 1983

Table 7.—Output indicators for Optical Telescopes—A Summary

	Lick 3-meter	KPNO[a] Z. I-meter	CTIO[b] 1 ,5-meter	INT[c] 2,5-meter
Average number of papers pa., 1969-78	42	43	35	7
Cost per paper m 1978	*f* 13k	*f* 7 k	*f* 6 k	*f* 63 k
Citations to work of past 4 years m 1978 . . .	920	710	580	140
Average citations per paper in 1978, ., .,	4.2	3.3	3.3	3.6
Number of papers cited 12 or more times in a year, 1969-78, .	41	31	21	4

[a] Kitt Peak National Observatory (U.S.).
[b] Cerro Tololo InterAmerican observatory (Chile).
[c] Isaac Newton Telescope (Great Britain).

SOURCE: B.R. Martin and J. Irvine, "Evaluating the Evaluators: A Reply to Our Critics, " *Social Studies of Science, vol.* 15, 1985, p. 569.

scientific work in terms of average citations per paper or number of highly cited papers. One can also see the decreasing importance of the CERN proton synchrotrons as newer machines such as the CERN super proton synchrotrons and the Gernman Electron Synchrotrons Laboratories (DESY) accelerator at Hamburg come on-line and begin to produce important results.

Table 9 presents the high energy physicists' own evaluations of the relative contributions of the different facilities described in table 8, based on a mail survey of 182 researchers in 11 countries. These evaluations are based on the relative outputs of the different accelerators over their entire

Table 8.—Experimental High-Energy Physics, 1977-80

	Percent of papers published in past 2 years		Percent of citations to work of past 4 years		Average citations per paper		Highly cited papers: number cited in times			
	1978	1980	1978	1980	1978	1980	n >15	n >30	n >50	n >100
CERN proton synchroton	22.0°/0	11.5 %	14.50/0	12.50/o	2.2	2.2	13	2	1	0
Brookhaven/AGS	5.50/0	5.5%	5.0 "/0	3.0%	2.7	1.6	0	0	0	0
Serpukhov	12.0°/0	14.00/0	4.0%	5.0 0/0	1.2	1.2	0	0	0	0
CERN ISR[a]	4.5%	5.5 %	7.0%	7.5%	5.4	4.4	11	2	0	0
Fermilab	16.5°/0	1 9.0%	32.0%	21.5%	7.3	3.6	40	10	5	1
CERN super proton synchroton	2.50/o	8.50/o	4.0%	8.50/o	12.7	5.0	19	7	3	0
SLAC [b]	9.5%	6.00/0	15.0%	11.5%	5,7	4.4	26	6	1	1
DESY	4.0 "/0	6.50/o	5.5%	15.50/0	5.7	8.8	36	16	4	0
Rest of world	23.5°/0	24.00/o	13.0 "/0	15.00/0	2.0	1.9	19	5	0	0
World total	1,115	930	8,190	5,090	3.5	3.0	164	48	14	2
	100"%0	100%	100%	100%						

[a]Intersecting storage rings.
[b]StanfordLinear Accelerator Center.

SOURCE: J Irvine and B R Martin, "Quantitative Science Policy Research, " testimony to the House Committee on Science and Technology Oct 30 1985

Table 9.—Assessments (on a 10-point scale[a]) of Main Proton Accelerators in Terms of "Discoveries" and "Providing More Precise Measurements"

	Self -rankings	Peer-rankings	Overall rankings (sample size= 169)
Discoveries:			
Brookhaven/AGS	9.5(± 0. 1)	9.0(± 0.1)	9.2(± 0. 1)
CERN PS	7 1(± 0.2)	6.7(± 0.2)	6.9(± 0. 1)
CERN ISR	6.8(± 0. 3)	5.9(± 3.2)	6.1(±0.2)
CERN SPS	5.9(±0.3)	5.6(±0.2)	5.7(±0. 1)
Fermilab	7 4(± 0.3)	7.1 (±0. 1)	7.2(±0.1)
Serpukhov	3.8(±0.5)	2.6(±0.1)	2.7(±0. 1)
More precise measurements:			
Brookhaven/AGS	7 1(±0.2)	7,2(±0.2)	7 2(*o. 1)
CERN PS	8.5(±0. 1)	8.5(±0.l)	8.5(±0.. 1)
CERN ISR	7 3(*0.3)	6.9(±0..2)	7 0(±0. 1)
CERN SPS	8 2(±0.2)	8.2(*0.2)	8.2(±0. 1)
Fermilab	6.3(*0.2)	6.0(±0. 2)	6. 1(±0..1)
Serpukhov	4.3(±0.5)	3.5(*0.2)	3.6(±0..2)

[a]10-top. The assessements are based on the relative outputs from the accelerators over their entire operational careers up to the time of the !nterviews with hlgh-energy physicists in late 1981 to early 1982.

SOURCE J Irvine and B.RMartin, "Quantitative Science Policy Research, ' testimony to the House Committee on Science and Technology, Oct 30 1985

operational careers and therefore do not necessarily match the output indicators from table 8, which are for a 3-year period. Comparable indicators exist for the entire 22-year period, 1960-82, in Irvine and Martin's papers.

Overall, the contribution of Irvine and Martin's work to research evaluation can be summarized as follows:

- They have collected, synthesized, and published a colossal amount of information—all original data—about the scientific performance of big and expensive scientific institutes.
- They have shown that when peers are assessing their own fields they can be reliable judges of scientific performance.
- Where choices have to be made in a field, among several similar research units competing for resources, Irvine and Martin provide policymakers with sound information for assisting a rational decision.

On the negative side, it is not known how the Irvine and Martin approach would fare in non-Big Science areas. Would the methodology transfer, as Irvine and Martin assert, to different cultural and research contexts? Even if converging indicators can validate contribution to scientific progress or the impact of a research team on its peer community, judgments of applicability and quality of these findings do not automatically follow. These are properties of interpretation, not analysis. Irvine and Martin tend to confuse the two.

The Irvine and Martin methodology is based solely on bibliometric and peer ratings among facilities in the same science, not knowledge-producing facilities in different sciences. However, the strategic choices between fields are the tough

ones in a zero-sum world. Like peer review, "converging partial indicators" are useless for strategic choices.

Finally, knowledge is produced by scientific communities, not individual institutions. Therefore, comparing facilities may bean empty exercise. The implication of Irvine and Martin's recommendations is that reducing or eliminating funding to the least cost-effective facility has no adverse effect on progress, and in fact diverts scarce resources to more productive facilities elsewhere. Such a strategy, however, undermines the knowledge-producing community and runs counter to the view of science as a cultural activity that intertwines local teams with distant peers through literature, informal communication, and the training of new generations of practitioners. These are not ignored by Irvine and Martin, but they are minimized.

The Center for Research Planning (Coward, Franklin, and Simon) has developed the method of "bibliometric modeling" based on Institute for Scientific Information (ISI) data and co-citation clusters to monitor the research front of a given specialty. [18] Each model consists of an intellectual base and the current work of the specialty. When brought together in a computer, these two sets of papers contain the building blocks of a model: "title words (keywords) of its current papers" and "the demographics of the specialty" (performing organizations and countries). The "age of its intellectual base" is an indicator of the specialty's "development potential. "

Working closely with the Economic and Social Research Council of the United Kingdom, the Center for Research Planning (CRP) built specialty-specific models and met in workshops with key participating research teams and technical experts from the respective research councils responsible for the funding. This hands-on approach allowed data to be passed to the scientists for their own independent analysis. In other words, CRP works with representatives of the knowledge-producing communities being evaluated and the policy users themselves to increase credibility and relevance of their studies. As their 1984 final report to the Advisory Board for the Research Councils states:

> The models are not intended to function as computer-based decision algorithms in the science resource allocation process, but should be viewed as a potential decision-support system .19

Though the CRP approach is, like Narin's, data-intensive and unobtrusive at its source, it is more interactive with the relevant actors. [20] It is unclear how this iterative and interactive process affects interpretations. While the modeling notion is made explicit by CRP, their methodology is not as well codified as Irvine and Martin's. Perhaps recognizing this, CRP is championing interactive computing with their models to moderate the suspicions of researchers and policymakers alike. Such interactions will allow users to ask specific questions of a model on-line and receive immediate answers. Though this innovation will have obvious appeal, the jury is still out on its efficacy.

Other Important Teams

In Holland, the research team of H.F. Moed, W.J.M. Burger, J.G. Frankfort, and A.F.J. Van Raam has carried out extensive comparative studies of the research productivity of different departments at their home University of Leiden. They have used publication and citation counts to track trends in the quantity and impact of research published by individuals and teams in the Faculties of Medicine and Mathematics and Natural Sciences over a 10-year period (1970-80).[21]

"H. R. Coward, et al., "ABRC Science Policy Study: Co-Citation Bibliometric Models" (abridged), presented to the Advisory Board for Research Councils, Department of Education and Science, United Kingdom, July 1984, pp. 1-3, 65.

"Ibid.

[20] L. Simon, et al., "A Bibliometric Evaluation of the U. S.-Italy Cooperative Scientific Research Program, " ***Evaluation of U.S.-Italy Bilateral Science Program*** **(Washington, DC: National Science Foundation, February 1985);** J.J. Franklin and H.R. Coward, "Planning International Cooperation in Resource-Intensive Science: Some Applications of Bibliometric Model Data, " papers presented at the National Science Foundation symposium entitled "International Cooperation in Big Science, " February 1985.

[21] H.F. Moed, et al., ***On the Measurement of Research Performance: The Use of Bibliometric*** Indicators (Leiden, the Netherlands: University of Leiden, Research Policy Unit, Diensten OWZ/PISA, **1983);** H.F. Moed, et al., "A Comparative Study of Bibliometric Past Performance Analysis and Peer Judgment, " *Scientometrics*, ***vol. 8, Nos.*** **3-4,** 1985, **pp.** 149-159; and H.F. Moed, et al., The Application of Bibliometric Indicators: Important Field- and Time-Dependent Factors to be Considered, " vol. 8, Nos. 3-4, 1985, pp. 177-203.

In France, the team of Michel Calon, Jean-Pierre Courtial, William Turner, and Ghislaine Chartron at the School of Mines in Paris has used the technique of "co-word" analysis to identify principal problem areas being worked on by the laboratones of a major French research institute and to situate that research in its international context. Co-word analysis monitors the number of times that keywords, identified by researchers as describing a research problem, occur in pairs in the research literature. A map of the pairings of these "co-words" can give one a sense of the structure of a research field .22

[22] A. Rip and M. Courtial, "Co-word Maps of Biotechnologies: An Example of Cognitive Scientometrics," *Scientometrics*, vol. 6, 1984, pp. 381-400. M. Callon, et al., "The Transition Model and Its Exploration Through Co-Word Analysis: Using Graphs for Negotiating Research Policies," Centre de Sociologie, Ecole des Mines de Paris and Centre Nationale de la Recherche Scientifique, mimeo.

THE USE OF BIBLIOMETRICS TO EVALUATE RESEARCH AT THE NATIONAL INSTITUTES OF HEALTH

Of all the Federal agencies supporting R&D, NIH conducts the most extensive ex post evaluation of its research through bibliometric studies and other activities carried out at the individual institutes. In 1970, the Public Health Service Act was amended to set aside for evaluation activities up to 1 percent of the funds appropriated to any program authorized by the Act for evaluation. Each of the 11 institutes of NIH receives a separate appropriation from Congress, so that each can evaluate its own programs. The Program Evaluation Branch in the Office of the Director studies cross-cutting issues, develops new approaches to evaluation, and supports the development of data resources for this purpose .23 The budget for evaluation studies at NIH was $5.8 million in 1985 ($2.8 million from set-aside funds and $3 million from the regular budget). A review of NIH's use of bibliometrics illustrate the range of useful information that can be produced.

NIH Databases

NIH maintains several extensive databases that are used for evaluation. The IMPAC database contains detailed information about all active reviews and awards of NIH grants, including the names of all principal investigators, the applicant's institution, the type of grant, the review group, priority score awarded through peer review, the funding institute, and the amount of support. Two longitudinal databases developed from IMPAC track the training or funding history for any investigator who has applied for NIH support. A separate financial database holds year-end data on all appropriations and obligations since 1950 for all NIH institutes and mechanisms of support .[24]

NIH maintains substantive research classification systems: Computer Retrieval of Information on Scientific Projects (CRISP) assigns subdisciplinary classification terms to each grant and contract, Medical Subject Heading (MeSH) provides subject description and classification information for every publication indexed in the Medical Literature Analysis and Retrieval System (MEDLARS) and MEDLINE. MeSH identifies source of research support, subject, author, title, journal, data, and descriptors of the research for every indexed research article published since 1981. These databases are used for literature searches and for evaluation of NIH activities in a given research area.

NIH uses a special database, MEDLINE, for bibliometric analysis. This database contains records of all articles, notes, and reviews that have appeared since 1970 in a selected group of biomedical journals, along with the sources of financial support acknowledged by the authors of each article. There are over 300,000 papers in the database, as well as a record of nearly 2.5 million citations to those papers. Originally, the 240 journals of the database covered about 80 percent of the publications resulting from NIH-supported research. The size of the journal base was expanded in 1981 to include the entire MEDLARS system of the National Library of Medicine, nearly 1,000 journals, accounting for 95 percent of NIH-supported research. This extensive database has been

Helen Hoter Gee, "Resources for Research Policy Development at the National Institutes of Health," typescript, presented before the Health Policy Research Working Group, Harvard University, Mar. 20, 1985.

[24] Ibid.

the subject of the bulk of U.S. bibliometric studies, most of which seek to measure the long-term scientific payoffs from NIH-supported research.

Bibliometric Studies at NIH

Grace M. Carter of the Rand Graduate Institute conducted the first NIH commissioned bibliometric study in 1974.[25] Carter explored the use of citations as a measure of the research output of 747 research project grants and 51 program project grants awarded on a competitive basis in fiscal year 1967. The three output measures for research grants were the priority score received on renewal applications, the production of at least one frequently cited article, and the average citation rate for publications cited at least twice. Using statistical multivariate analyses, Carter tested a series of hypotheses about the three research output measures. Her examination of study section judgments of renewal applications revealed that on average grants proposed for renewal produced more useful research results than other grants. She also found that a grant that produced a highly cited publication was more likely to be renewed than one that did not. Thus, peer group evaluation and citation analysis produced comparable results.

In addition, Carter found a high correlation between priority scores on the first grant application and the number of subsequent publications and citations. She also found that research proposals perceived by study sections to have a high probability of being "exceptionally useful" received higher priority scores and more years of funding than those not so perceived. Carter concluded, rather cautiously, that the concept of "scientific merit" contains enough objective content that different groups of people meeting several years apart will agree that one set of grants is more scientifically meritorious than another set of grants.[26]

Francis Narin furthered the work of Carter by using bibliometric techniques to obtain quantitative indicators of research performance that were in general accord with the intuitive expectations of the research community .27 He was able to establish a degree of concordance between the structure of biomedical research literature and the structure of biomedical knowledge, which enabled him to use bibliometric databases and analyses to demonstrate a number of interesting points:

- Utilizing correlational techniques, he was able to establish correspondence between bibliometrically measured research productivity indicators and quantitative, nonbibliometric measures, including institutional funding and institutional ranking based on formal peer assessment.
- International biomedical publication rates are highly correlated with the GNP and national affluence (GNP per capita). [28]
- Changes in U.S. research funding can be associated with changes in the number and content of research publication 3 to 5 years later.
- Basic biomedical information is both published and cited by scientists supported by many bureaus, institutes, and divisions at NIH, forming a pool of fundamental research knowledge. In contrast, clinical information is produced and used by a narrower set of largely clinical researchers. Basic research is more highly cited than clinical research.
- Differences exist in the kinds of research publications produced by scientists in medical schools of different sizes and levels of national prestige. The number of publications

[25] Grace M. Carter, ***Peer Review, Citations, and Biomedical Research Policy: NIH Grants to Medical School Faculty,*** prepared for the Health Resources Administration and the Office of the Assistant Secretary for Planning and Evaluation of the Department of Health, Education, and Welfare, R-1583-HEW (Washington, DC, HEW, December 1974).

[26] Ibid., p. v.

[27] Francis Narin, ***Concordance Between Subjective and Bibliometric Indicators of the Nature and Quality of Performed Biomedical Research,*** a Program Evaluation Report for the Office of Program, Planning and Evaluation, National Institutes 01 Health (Washington, DC: NIH, April 1983); Francis Narin, ***Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity,*** monograph prepared for the National Science Foundation, Accession #PB252339/AS (Springfield, VA: National Technical Information Service, March 1976).

[28] J. Davidson Frame and Francis Narin, "The International Distribution of Biomedical Publications," ***Federation Proceedings,*** vol. ***36, No. 6,*** May **1977, pp.** 1790-1795. Frame and **Natin** investigated the U.S. role in international biomedical publication based on counts of articles, notes, and reviews in 97s biomedical journals, Thev found that the United States authored 42 percent of these papers which were far more heavily cited than papers from other countries. Only 4 percent of publications were found to originate from underdeveloped regions.

produced per dollar of research funding is the same for the large and small institutions, indicating neither economies or diseconomies of scale. However, scientists from the larger medical schools publish their papers in more prestigious journals, and in a much wider set of subfields than smaller schools. Smaller schools can attain a critical mass of research activity only if they concentrate their research effort in a select area. In addition, faculty perceptions of the ranking of medical schools are very much in accord with bibliometric measures of the ranking of the same schools.

- The research supported and performed by the different institutes is appropriately concentrated in the clinical areas corresponding to their missions.
- Publications resulting from research conducted at NIH are more highly cited than publications in the same research areas supported through other sources.

Narin's work was the basis of the most widely accepted application of bibliometric techniques—the *Science Indicators* series of the National Science Foundation[29]—and provided the fuel for more extensive use of bibliometrics for evaluation and planning at NIH. Evaluations include the effectiveness of various research support mechanisms and training programs, the publication performance of the different institutes, the responsiveness of the research programs to their congressional mandate, and the comparative productivity of NIH-sponsored research and similar international programs.

Bibliometric analysts tested the ability of citation maps and co-citation clusters to detect transitions between basic research and clinical research in the biomedical sciences. [30]Citation maps of research in Lesch-Nyhan syndrome, Tay-Sachs disease, and the effects of the drug methotrexate displayed the anticipated transitions, though the extent of the transition varied from case to case. Co-citation cluster analysis identified the basic or clinical orientation of the different research areas but did not find the sought-for transition points.

Most recently, NIH has undertaken bibliometric studies to determine the effectiveness of different research support mechanisms. For example, a study of the research centers' programs using subsequent grant applications and publications as the criteria, found that a grant to a center is a more effective mechanism for supporting clinical research than it is for supporting basic science .31 The Dental Institute used bibliometric analysis to determine that their centers' program has been effective in recruiting new scientists to research relevant to the institute's mission. Similarly, the National Cancer Institute and Francis Narin are conducting a study to identify the contributors to the most important research findings of the last 15 years and to determine where the research was conducted and the mechanisms of support. The research will test some assumptions of biomedical research grants policy, for example, that the individual investigator grant is superior to the contract and that extramural science is better than intramural .[32]

Bibliometrics have also been used for evaluating biomedical manpower training programs. Peter Coggeshall, a staff member of the National Academy of Sciences (NAS), used the NIH database, NSF grant data, and the NAS Survey of Doctoral Recipients to compare investigators who had received NIH predoctoral training support with two other groups—a group that had been trained in departments that had received training grants and a group that had received no NIH support in any form. The study concluded that individuals who received NIH predoctoral support produced superior subsequent career records in terms of publications and citations, were more likely to be working on NIH-sponsored activities, and were more successful in obtaining grants .33

Although Narin's early work showed that the distribution of papers supported by each institute

[29]*Science Indicators 1972; 1974; 1976; 1978; 1980; 1982; 1984*, **reports to the National** Science Board, National Science Foundation (Washington, DC: U.S. Government Printing Office, 1973, 1975, 1977, 1979, 1981, 1983, 1985)

[30]U. S. Department of Health and Human Services, *Applications of Bibliometric Methods* ***to the Analysis and Tracing*** *of Scientific* ***Discoveries,*** HHS-NIH Evaluation Report NTIS #PB80-210586 (Springfield, VA: National Technology Information Service, 1981).

[31]Gee, op cit, p. 9

[32]Lou Carrese, Associate Director for Program Planning and Analysis, National Cancer Institute, personal communication 1985

[33]Institute of Medicine, *The Career Achievements of NIH Predoctoral* ***Trainees and Fellows*** (Washington, DC National Academy Press, 1984).

in the basic and clinical medical disciplines follow very closely the institute's mission, several institutes continue to pursue the use of bibliometrics to validate accountability. For example, Narin has recently used bibliographic methods to evaluate trends in pulmonary and hypertension research. He found that National Heart, Lung, and Blood Institute actions since the passage of the National Heart, Blood Vessel, Lung, and Blood Act of 1972 have led to quantifiable progress in the research areas listed in the mandate .[34] Narin's work with the National Cancer Institute will be applied to the same purpose. In addition, the National Institute of Mental Health is conducting a 10-year analysis of the publication record of its grantees for purposes of accountability.

In almost all cases, bibliometric studies evaluated program performance and conformity to agency or institute mission. In some cases, they helped to identify areas for future research funding. The National Institute of Mental Health has begun to use bibliometrics and cluster groups to conduct a form of "portfolio analysis. " Looking at their program portfolios and the clustering of research publications by field, they are identifying leading edges of research that might require more support from their institute. 35 Narin has shown empirically that bibliometric data may qualify as an important adjunct measure to more subjective measures applied through peer review.

The Utility of Bibliometrics for Research Decisionmaking

Bibliometric techniques provide rough indicators of the quantity, impact, and significance of the output of a group of scientists' research. They are not generally considered valid for measuring the productivity of individual scientists due to differences in publishing styles and journal requirements, and the questionable validity of small statistical samples. However, Cole and others have shown that publication counts correlate positively with other measures of individual scientists' research quality such as peer review, Nobel Prizes, and prestige of academic appointment .36

Publication counts give a rough measure of the quantity of work produced by a research team or facility. Citation counts are an indicator of the influence that work has had on the larger scientific community. And the number of citations per article or the number of highly cited articles provide a rough measure of the significance of the work, since important papers tend to be cited most often. These indicators can help a funding agency compare the quantity, quality, and visibility of research done by various individuals or institutions. They can help identify the strong research groups and the relative cost-effectiveness of research sponsored at different centers.

However, they have two important limitations with respect to research decisionmaking. First, they are entirely retrospective. They have no inherent future predictive capability, unless one believes that past performance is an indicator of likely future achievement—not an unreasonable assumption. Therefore, they are more applicable to research program evaluation than to research planning. Second, they are not applicable to *strategic* decisions about resource allocation between fields. Most bibliometricians contend that the techniques can only be validly applied within individual disciplines. Publication and citation practices vary too widely between fields to allow for interdisciplinary comparisons. This, unfortunately, makes bibliometric techniques of *limited value for the most important decisions facing agency heads and congressional decisionmakers—allocating resources among fields.*

It should be noted that some analysts dissent from this view. Derek de Solla Price, in an unpublished article for the National Research Council, argues that the relative strength of different

'Public Health Service, ***Bibliographic Methods for the Evaluation of Trends in Pulmonary and Hypertension Research,*** NTIS #PB82-159724, 1982.

"Lawrence J. Rhoades, Science Policy Planning and Evaluation Branch, Office of Policy Analysis and Coordination, National institute of Mental Health, personal communication, **1985.**

'G. A. Cole, ***The Evaluation of Basic Research in industrial Laboratories*** (Cambridge, MA: Abt Associates 1985) p 43

fields in the United States can be assessed by comparing the ratios of the numbers of citations of U.S. articles in foreign journals to the numbers of citations of foreign articles in U.S. journals by field, normalized to account for differences in national research "output" by field. Price's scheme is quite complicated and involves measures of quality, quantity, and "internationality, " but it is a first attempt to compare fields using dimensionless indicators that have been normalized to remove the effects of different publishing and citing practices between fields. 37

Co-citation analysis enables one to monitor how specialties or subfields evolve over time. Cogitation analyses display the relationships among highly cited papers by *showing how many times such papers are cited together in single articles.* Based on co-citations, two-dimensional diagrams or maps of specialties can be created which illustrate the clustering of the most important works in that specialty, based on the number of citations By examining changes in the clusters one can track the evolution of the specialty over time. For ex - ample, figures 2A, B, C, and D illustrate the evolution of the collagen specialty cluster in biochemistry between 1970 and 1973. As can be seen by comparing figures 2D and A, the cluster map has become much larger by 1973, and most significantly, an entirely new set of research papers has replaced the cluster of most important works identified in 1970. This change coincided with the discovery of a new substance, pre-collagen, in 1971, which totally reoriented the research front in the specialty. Thus co-citation maps can, in principle, help one to identify important changes in research specialties over time.

'-Derek de Solla Price, "Science Indicators ot Quantity and Quality for Fine Tuning of United States Investment in Research in Major Fields of Science and Technology, " typescript draft, paper prepared at the request of the Commission on Human Resources, National Research Council, April 1980, typescript draft.

Even without maps, co-citation cluster analysis can help one identify the level of research activity in different specialties. Table 10 takes specialty

Table 10.—Changes in Sample of Continuing Clusters, 1970-73

Specialty	Direction of change	1970-71 (%)	1971-72 (%)	1972-73 (Ye)
Nuclear levels	c	58	**45**	25
	d	21	**55**	17
	n	21	0	58
Adenosine trlphosphatase	c	67	25	67
	d	0	50	22
	n	**33**	25	11
Australia antigen	c	**55**	54	**57**
	d	**4**	26	30
	n	**41**	20	13
Proton-proton elastic scattering	c	50	7	44
	d	50	21	25
	n	0	72	31
Ultrastructure of secretory cells	c	50	43	60
	d	12	57	0
	n	38	0	40
Nuclear magnetic resonance	c	37	55	23
	d	13	9	54
	n	50	36	23
Polysaccharides	c	46	44	36
	d	46	34	7
	n	8	22	57
Crystallizatlon of polymers	c	100	100	100
	d	0	0	0
	n	0	0	0
Affinity chromatography	c	60	67	72
	d	20	0	14
	n	20	33	14
Leukocytes: chonic granulomatous disease	c	40	63	33
	d	13	5	53
	n	47	32	14
Collagen	c	80	40	27
	d	20	0	40
	n	0	60	33
Erythrocyte membranes	c	9	15	58
	d	64	5	42
	n	27	60	3
Delayed hypersensitivity	c	77	46	50
	d	15	27	29
	n	8	27	21

c - continuing, d = **dropping, n - new documents**

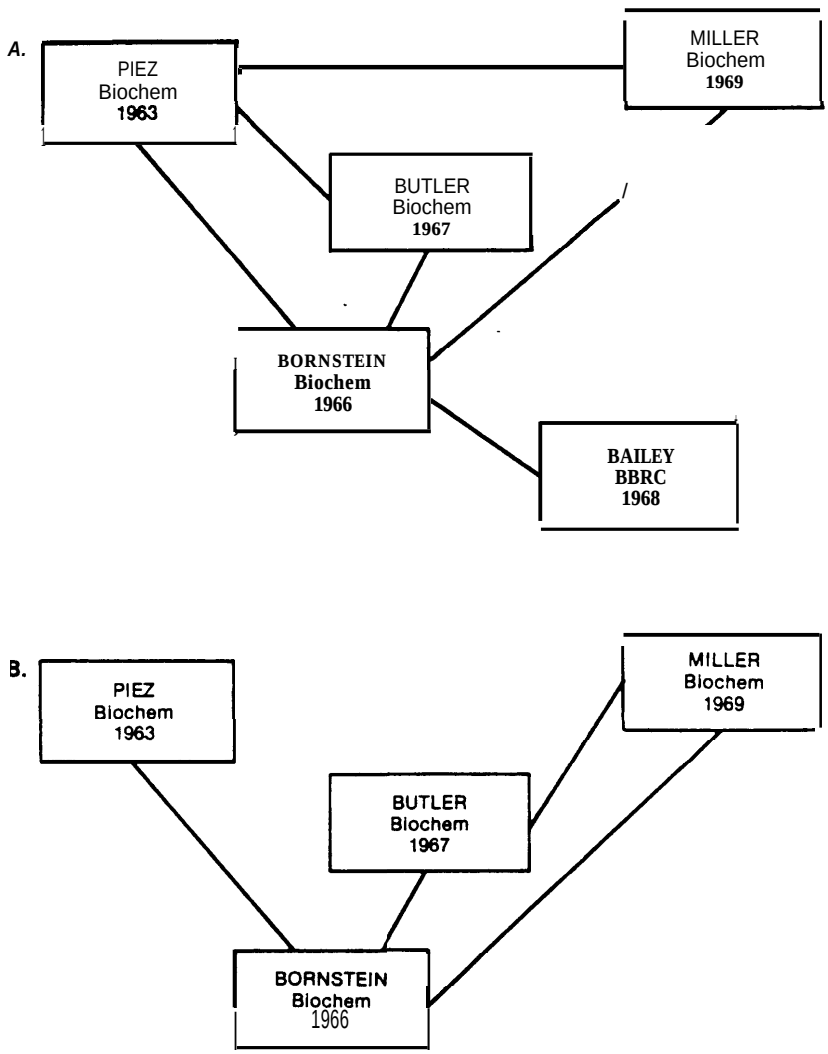
SOURCE Yehuda Elkana, et al., (eds.) Toward a *Metric of* Science *The Advent of Science Indicators* **(New York: John Wiley & Sons.** 1978), pp 199-201

clusters in biochemistry and shows, for each, the percentage of key papers that are identical to the previous year's, the percentage that have dropped out from the previous year, and the percentage that are new. As can be seen, the specialties vary appreciably, from crystallization of polymers, for which the same papers defined the subfield cluster over all 3 years, to erythrocyte membranes, in which 64 percent of the important papers dropped out in 1970, and 80 percent of the papers were totally new in 1971. If research decisionmakers are eager to fund specialties where new ideas are emerging rapidly, data on the evolution of cluster specialties over time could help to identify fields of rapid change.[38]

It should be stressed, however, that most bibliometncians view publication, citation and cocitation analyses as complements to, not substitutes for, informed peer evaluation. All three analyses are, of course, ultimately indirect measures of the scientific community's peer evaluation of researcher's productivity.

[38] "Eugene Garfield, et al., "Citation Data as Science Indicators, ***Toward a Metric of Science: The Advent of Science indicators,*** Yehuda Elkana, et al., (eds.) (New York: John Wiley & Sons. 1978), pp. 196-201.

Figure 2.–Davelopment of a Speciality Cluster, 1970-73

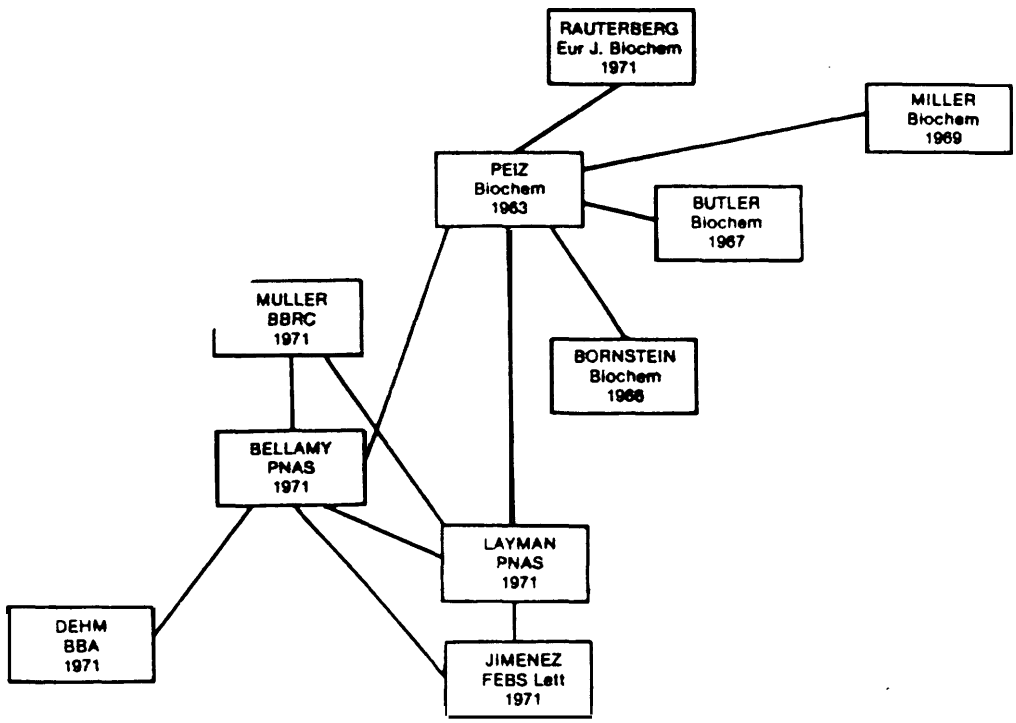


The figure shows the evolution of the collagen cluster over the 4-year period 1970-73. Boxes contain the names of first authors of the highly cited papers and years of publication. Lines connect papers co-cited at least 11 times in the corresponding source year.

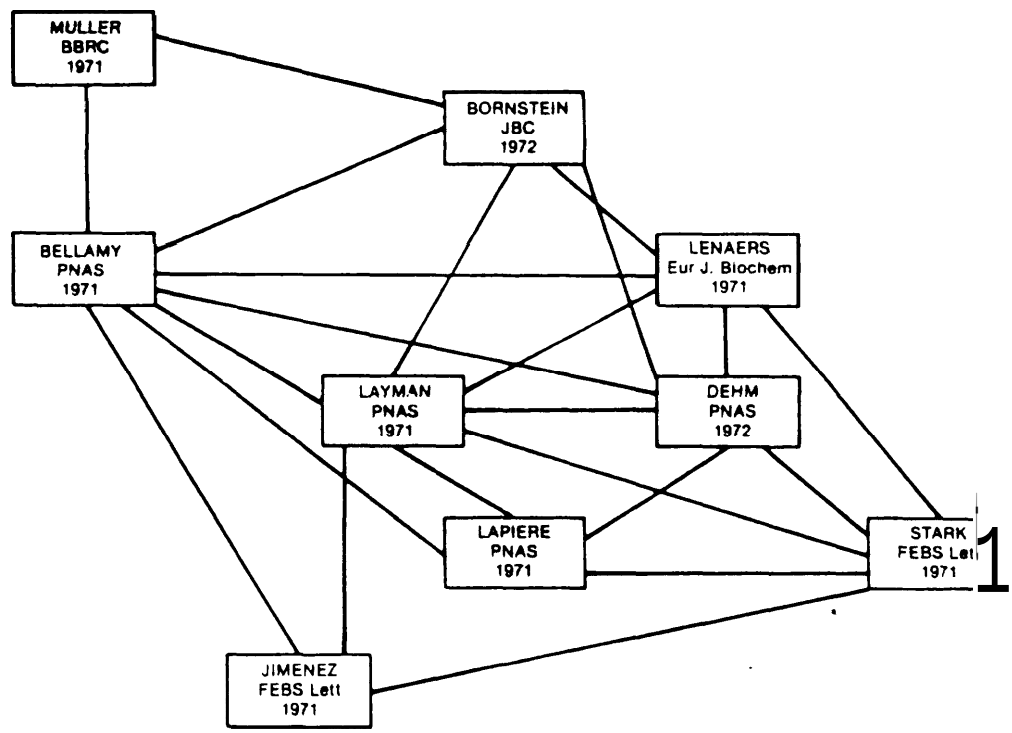
NOTE A IS collagen, 1970, and B IS collagen, 1971

Figure 2.—Deveiopment of a Specialty Cluster, 1970-73 (continued)

C.



D.



NOTE: C. is collagen, 1972; and D. is collagen, 1973.

SOURCE: Eugene Garfield, et. al., "Citation Data as Science Indicators," *Toward a Metric of Science: The Advent of Science Indicators*, Yehuda Elkana, et al., (eds.) (New York: John Wiley & Sons, 1978), pp. 199-201.

SCIENCE INDICATORS

One method of assessing the health of the research enterprise is to directly question the scientists and administrators involved in it. This is done in a variety of ways in this country. Individual scientists are asked to testify at congressional hearings. Federal agencies create scientific advisory panels to help guide research.

The National Research Council and its constituent bodies—the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine—carry out numerous reviews of research programs and research fields for the executive and legislative branches. The most comprehensive of these are the Research *Briefings* and Five *Year Outlooks* prepared by the Committee on Science, Engineering, and Public Policy. Since all NRC reports are prepared by committees of scientists, they represent, to some degree, the informed, consensus-based peer judgments of the scientific community on the state of the research enterprise. As a check on the validity of such reports, the government could support detailed surveys of the scientific community.

However, scientists' judgments on the state of their own field of research can rarely be totally disinterested and often reflect the researcher's characteristic desire to investigate more problems in greater depth than available funding will allow. These inherent biases can be compensated, to some degree, by the use of a variety of science indicators that are measured by NSF, NIH, and NRC and published on a regular basis. These indicators include the amount of funds devoted by the Nation to research and development by source, sector, nature of the work, performer, and scientific fields; statistics on the distribution of scientific and engineering personnel, graduate students, and degree recipients by field, sector, and institution; and the support for graduate education and training. The NSF Science *Indicators* tables include funding levels by agency and even program; the specific institution receiving the funds, employing the scientists, and training the graduate students; and funding for specific Standard Industrial Classification codes within industry. Most lacking in the policy community is a consensus on which indicators are most relevant and how the different indicators might be used in combination to measure the health of the research enterprise. A workshop or report on the use of Science *Indicators* to measure the health of the research enterprise might be a useful first step in that direction.

One must remember, however, that all measures or "indicators" of research are inevitably flawed. Any number describing research is an abstract symbol that depicts, imperfectly, only one aspect of it. Choosing one measure over another implies that the measurement use has made some assumption about what is important. The chosen measure has meaning only through interpretation.

Even if an acceptable measure of an aspect of research can be devised, interpretation remains problematic:

> . . . the inputs [to science]—of dollars, of working scientist, males, females, and Hispanics, graduate students, post-docs, and professors—are well known and further broken down into industry, government, education, or lost to view. We also have counts of outputs—of papers, citations of papers, and Nobel Prizes arranged according to national origin. But how do we know what the numbers "ought" to be? . . . such indicators help very little in determining the health of science in any absolute sense or, more practically, in relation to what it might be if organized and financed at some theoretical optimal level .[39]

These difficulties illustrate the limitations of science indicators for research evaluation.

[39] R.S. Morison, "Needs, Leads, and Indicators, " Science, *Technology, and Human Values,* vol. 7, winter 1982, pp. 6-7.