ELSEVIER

Research Policy 29 (2000) 657–678

research policy

www.elsevier.nl/locate/econbase

Evaluating technology programs: tools and methods

Luke Georghiou [a], David Roessner [b,c,*]

[a] *University of Manchester, Oxford Rd., Manchester M13 9PL, UK*
[b] *Department of Public Policy, Georgia Institute of Technology, Atlanta, GA 30332-0345, USA*
[c] *Science and Technology Policy Program, SRI International, 1611 N. Kent St., Arlington, VA 22209, USA*

Abstract

This article reviews the analytical tools, methods, and designs being used to evaluate public programs intended to stimulate technological advance. The review is organized around broad policy categories rather than particular types of policy intervention, because methods used are rarely chosen independent of context. The categories addressed include publicly-supported research conducted in universities and public sector research organizations; evaluations of linkages, especially those programs seeking to promote academic-industrial and public-private partnerships; and the evaluation of diffusion and industrial extension programs. The internal evaluation procedures of science such as peer review and bibliometrics are not covered, nor are methods used to projects and individuals ex ante. Among the conclusion is the observation that evaluation work has had less of an impact in the literature that it deserves, in part because much of the most detailed and valuable work is not easily obtainable. A second conclusion is that program evaluations and performance reviews, which have distinctive objectives, measures, and tools, are becoming entangled, with the lines between the two becoming blurred. Finally, new approaches to measuring the payoffs from research that focus on linkages between knowledge producers and users, and on the characteristics of research networks, appear promising as the limitations of the production function and related methods have become apparent.

Keywords: Technology programs; Tools; Methods

1. Introduction

Demand for evaluation has been fueled by the desire to understand the effects of technology policies and programs, to learn from the past and, more instrumentally, to justify the continuation of those policies to a sometimes skeptical audience. The relation between the evolution of evaluation methods and approaches and that of technology policies, we shall argue, is complex. The development of evaluation approaches has responded to varied stimuli. In the mid-1980s, the first OECD review in this field noted the convergence of the internal evaluation traditions of science with a growing demand for evaluation of public policy in general (Gibbons and Georghiou, 1987). The latter trend would later be characterized as a feature of ''new public management'' and reach its apotheosis in the current requirement for programmatic and institutional performance indicators. A third influence was the growing tendency to associate science with competitive performance and the search for more effective ways to achieve that linkage. The growth of activity was catalogued in other reviews at that time (Roessner, 1989; Meyer-Krahmer and Montigny, 1990) that ex-

* Corresponding author. Department of Public Policy, Georgia Institute of Technology, Atlanta, GA 30332-0345, USA

0048-7333/00/$ - see front matter
PII: S0048-7333(99)00094-3

tended the analysis to innovation programs from the US and European perspectives, respectively. Since that time, as policies have increasingly focused upon fostering linkages of various types within innovation systems, evaluation methods have been developed which aim to characterise and measure those linkages.

In this paper, we have chosen to focus on technology policy rather than the broader scope of innovation policy. The evaluation literature is now so extensive that it has been necessary to be selective even within this constraint. Evaluation methods tend to cluster around particular types of policy intervention. Hence, we have found it more useful to organize the review by broad policy category than by an attempt to group methods independently of their context. Following this approach, three related focal points for evaluation are used to structure what follows. These appear to follow a sequential model of innovation from basic research, through academic–industrial linkages to industrial collaborative R&D and finally diffusion and extension. However, it will emerge from what follows that similar questions (for example those concerning economic returns) may be posed at any stage of this sequence. Furthermore, evaluations have been instrumental in exposing the many feedback loops and unexpected consequences that have modified the way in which the innovation process is understood.

The first of our focal points concerns evaluation of publicly supported research carried out in universities and public-sector research organizations. One reason this is included is the growing significance attached to the economic and social rationales for public support of research. Hence, the relevant section addresses means by which the economic and social value of science is being assessed. The internal evaluation procedures of science, principally peer review with some contribution from its offshoots in scientometrics, are not covered here. Modified peer review, or merit review as it is sometimes called, extends the terms of reference of a peer review panel to cover the broader socio-economic issues that are discussed below. Often, the outcomes of such panel reviews are highly influential because of the status of their members. However, from a methodological perspective, interest in such approaches is limited to the variations in structure and composition of consultation and to the ways in which information inputs and reporting are organized. In general, the organizational and political dimensions of evaluation, though important, are not covered in this review.

In the second part of this section, the scope is broadened to include evaluations that focus upon linkages, including those of programs seeking to promote academic–industrial and public–private partnerships. The centrality of these interfaces to technology policy has stimulated a large body of research, much of which has an evaluative character. The selection made here aims to give a flavor of the range of evaluation methods used rather than to review the topic as a whole. These interfaces are also present in the next focus, collaborative R&D programs, though industrial collaboration is usually their main aim. This policy instrument has had a special relationship with the development of evaluation, with the two growing together during the 1980s. It continues to attract a large amount of evaluation activity.

Finally, experience in the evaluation of diffusion and extension programs is discussed. In methodological terms, this involves a transition from the rather lumpy and unpredictable distribution and attribution of benefits that characterize research, to a domain where large numbers of client firms are in receipt of less visible assistance. Evaluation here is thus characterized by treatment of large data sets, though we will argue that these conceal a wide range of services and clients.

Two other points need to be made concerning the scope of this review. The first is that evaluation is a social process, which means that methods cannot be equated with techniques for collection of data or their subsequent analysis. The choice of what is significant to measure, how and when to measure it, and how to interpret the results are dependent upon the underlying model of innovation that the evaluator is using, implicitly or explicitly. Much of the data collected by evaluators are themselves conditioned by the positioning of the evaluation and those who execute it. In consequence, it is usually necessary to understand the setting of the evaluation and the discourse in which its results are located before the choice of approach can be fully appreciated.

The second qualification about our scope addresses the level of evaluation covered. The central unit of analysis is the program evaluation. In many

ways, this is the easiest territory for evaluators in that a program almost by definition has boundaries in space and time and certainly should have objectives. However, when moving across national and cultural borders this criterion for inclusion may be over-restrictive. Evaluation in France, in particular, is mainly at the level of institutions, even if many of the same issues are addressed (Larédo and Mustar, 1995).

Aspects that remain excluded from this review are ex ante evaluation, evaluation of projects and of individuals (though each of these may well be linked to the evaluations we discuss). The last consideration in terms of scope is that of the most aggregated evaluations, which address whole sectors, policies, and national or international systems. Many of these tend to intrude into the general realm of technology policy studies, but some are more firmly rooted in the tradition of evaluation. We will argue in the concluding section of this paper that understanding a program may well require more work on the systemic context than is presently common practice.

2. Evaluation of the socio-economic impacts of research in universities and public laboratories

2.1. Measuring the returns

Within the past 15 years, there have been several comprehensive assessments of methods for measuring the returns, benefits, and/or impacts of public support for research. The Office of Technology Assessment's 1986 Technical Memorandum was stimulated by Congressional interest in improving research funding decisions through the use of quantitative techniques associated with the concept of investment (U.S. Congress, Office of Technology Assessment, 1986). The OTA review covered economic and output-based, quantitative measures used to evaluate R&D funding. Economic methods included macroeconomic production functions, investment analysis and consumer and producer surplus techniques. Output methods included bibliometric, patent count, converging partial indicators, and science indicators approaches.

Hertzfeld (1992) provided a probing critique of methods employed to measure the returns to space research and development, but did so within a larger framework of general approaches to measuring the economic returns to investments in research. He classified these approaches into three distinct types (Hertzfeld, 1992, p. 153):

1. The adaptation of macroeconomic production function models to estimate the effects of technological change or technical knowledge that can be attributed to R&D spending on GNP and other aggregate measures of economic impact.
2. Microeconomic models that evaluate the returns to the economy of specific technologies by estimating the consumer and producer surpluses these technologies create.
3. Measuring the direct outputs of research activity through reported or survey data on patents, licenses, contractor reports of inventions, value of royalties or sales, and so forth.

Most recently, the Critical Technologies Institute of the RAND Corporation (now renamed the Science and Technology Policy Institute) published a report prepared for the White House Office of Science and Technology Policy that reviewed methods available for evaluating fundamental science (Cozzens et al., 1994). The RAND study classified evaluation methods into three types:

1. Retrospective, historical tracing of the knowledge inputs that resulted in specific innovations.
2. Measuring research outputs in aggregate from particular sets of activities (e.g., programs, projects, fields, institutions) using bibliometrics, citation counts, patent counts, compilations of accomplishments, and so forth.
3. Economic theory/econometric methods employing, as measures of performance, productivity growth, increase in national income, or improvements in social welfare as measured by changes in consumer and/or producer surpluses (Cozzens et al., 1994, pp. 41–43).

What is striking about these critical assessments is that, despite their nearly 10-year span, there is virtually complete agreement among all three concerning the strengths and weaknesses of the various approaches to determining the benefits or payoffs from investments in research generally, and fundamental science in particular. Moreover, improvements over the period in the several techniques described have been modest and incremental. There have been numerous examples of the application of these tech-

niques to measure the impacts of research that span the past 30 years. There is considerable agreement on which studies represent the best of each genre, and on the limitations of each approach.

Historical trace studies such as Hindsight (Sherwin and Isenson, 1967), TRACES (Illinois Institute of Technology Research Institute, 1968), and ''TRACES II'' (Battelle, 1973) offer the possibility of detailed information about the relative contribution of basic vs. applied research, institutional contexts, and sources of research support. But the method requires subjective judgments about the significance of particular research events, and suffers from lack of generalizability because of non-random innovation selection processes. It also traces single products of research backwards in time to identify the full range of contributions, rather than tracing forward from a single research activity (e.g., a project) to identify multiple impacts or consequences. Further, it is an extremely costly method; given the complexity of knowledge flow processes and variations across fields of science and technology, the number of cases necessary for generalization would be prohibitively expensive. Finally, historical trace studies fail to account for the indirect effects of research, including dead ends (from which substantial learning takes place), spillovers, and synergistic effects.

Other, more aggregate, approaches seek to overcome the limitations of historical trace studies. Measures of basic research outputs have been refined to the point where reasonably valid and reliable data on the quantity and quality of research outputs can be obtained. It has been argued that while any single output measure is insufficient, the relative effectiveness and efficiency of research programs, organizational units, and institutions can be measured using combinations of measures (e.g., Martin and Irvine, 1983). The problem for such indicators generally is one of linking outputs to impacts, especially impacts that are valued in markets or the political arena. For some policy purposes, such as resource allocation decisions among several institutions conducting basic research, these kinds of studies can be of considerable value. But they are inadequate for questions concerning the impact of basic research on larger societal goals, especially when measures of value are sought. In addition, except in the most general sense, these approaches offer little guidance to research program managers seeking to maximize the benefits that result from their programs of research. As the RAND/CTI study concluded,

> Most direct indicators of R&D outcomes such as citation and other bibliometric indicators are only indirectly (and loosely) related to those aspects of economic and other goals of public improvement which are of greatest interest (Cozzens et al., 1994, p. 44).

As noted above, economic assessments of research fall into two basic categories: production function analyses and studies seeking social rates of return. Production function studies assume that a formal relationship exists between R&D expenditures and productivity. While there is ample evidence that such a relationship exists, there are numerous problems with this approach. Prominent among the studies that employ this approach are those by Denison (1979), Link (1982), Levy and Terleckyj (1984), Jaffe (1989) and Adams (1990). Regardless of the assumptions underlying these studies, all conclude that the relationship between private R&D expenditures and various measures of economic growth and productivity are positive and substantial. But there are fundamental problems with this approach that are exacerbated when applied to public R&D expenditures. First, the ''technical information'' term in the production function is only approximated by R&D expenditure data, and in any event its effect on the larger production system is not well understood. Second, the approach does not account adequately for externalities that result from R&D activity. Third, if the focus is on publicly supported R&D, additional problems arise because the intent of most public R&D is not to stimulate economic growth, but to achieve (public) agency missions. Any contribution to economic growth is thus due to indirect knowledge transfers. Fourth, the contributions of basic and applied research are difficult to distinguish, yet these are important for many policy purposes. Finally, this approach is designed to address economic benefits that result from incremental changes in production efficiencies, but cannot adequately account for the effect of new products or radical changes in production efficiency (Cozzens et al., 1994, p. 48).

Aggregate social rate of return studies attempt to estimate the social benefits that accrue from changes in technology and relate the value of these benefits to the cost of the investments in research that produced the changes of interest. These studies employ, by analogy, what amount to the internal rate of return calculations often used by private firms. Social benefits are measured as the sum of consumer and producer surpluses. Prominent examples of this approach are studies by Mansfield (1980; 1991), Mansfield et al. (1977), Tewksbury et al. (1980) and Link (1981). At a more disaggregate level, the consumer surplus approach has been used to estimate the returns from public investments in technology development programs as well. We discuss this, and the use of an alternative technique, briefly in a subsequent section of this article.

As in the case of production function studies, aggregate consumer surplus studies tend to show that the aggregate rate of return to research is positive, and that the social rate of return exceeds, on average, the private rate of return. Also, as in the case of production function studies, this approach has significant drawbacks. First, the causal mechanism that links R&D investment to social or private returns is imputed but not direct. Second, because calculations of consumers' and producers' surplus depend on the existence of supply and demand curves, they are inappropriate for goods and services that are radical departures from their predecessors (which is just the case for the most interesting products of basic research). Third, because they rely on data from individual cases, like historical trace studies they are difficult to generalize from and expensive to conduct. Fourth, they rely on an adaptation of an investment model intended for short-term investment calculations, not the long-term benefits that might accrue from basic research. Thus, the discounting requirements of the models may severely underestimate the contribution of basic research. Finally, as with other approaches, spillovers are difficult to identify and account for (Cozzens et al., 1994, p. 55).

In summary, the dilemma the limitations of these approaches pose for the research evaluator lies in the question of attribution, this being that to realise economic effects of research a range of complementary factors from within and beyond the innovation process are brought to bear, obscuring the relationship under study.

Hertzfeld concludes his assessment of the various techniques with a recommendation:

> ... the most promising type of economic study for measuring returns to space R&D is the documentation of actual cases based on direct survey information. Case studies provide relatively clear examples of the returns, both theoretical and actual, to space R&D investments. A well-structured series of case studies coupled with good theoretical modeling aimed at integrating the data and incorporating those data with more general economic measures of benefits may well be the way to establish reliable and 'believable' economic measures of the returns to space R&D (p. 168).

Hall (1995), in a review of the private and social returns to R&D supports the employment of a broader range of approaches, cautions that case study evidence may be hampered by focus on 'winners', and needs to be supplemented by a computation of returns.

3. Evaluating linkages

Since the RAND/CTI assessment was published, a number of papers have appeared that are critical of the entire conceptual framework economists have been using to analyze scientific activity. This ''new economics of science,'' perhaps best formulated by Dasgupta and David (1994), is critical of the ''old economics of science'' based on production function-based approaches to assessing the payoffs from science (and its linkages to technology), and introduces broader perspectives that may lead to new approaches for evaluating basic science. By incorporating the insights gained from sociological studies of science, economists' perspectives are expanding to incorporate the characteristics of research institutions and the professional networks that bind knowledge producers and users. One of Dasgupta and

David's propositions emerging from their work reads as follows:

> The organization of research under the distinct rules and reward systems governing university scientists, on the one hand, and industry scientists and engineers, on the other, historically has permitted the evolution of symbiotic relationships between those engaged in advancing science and those engaged in advancing technology. In the modern era, each community of knowledge seekers, and society at large, has benefited enormously thereby (Dasgupta and David, 1994, p. 518).

Although much of the ''new economics'' is devoted to the implications of this broader thinking to resource allocation efficiencies, its inclusion of the processes of knowledge transfer and use and of the networks by which such transfers take place offers opportunities for new ways of thinking about how to value science beyond attempts to monetize the value of the knowledge produced and/or to infer such value from the applications that enter economic markets.

Bozeman et al., as part of an ongoing evaluation of the US Department of Energy's Office of Basic Energy Sciences, have recognized this opportunity and are currently developing new theory about knowledge and innovation that, in our view, shows promise for methodological advances in evaluating basic research (Bozeman and Rogers, 1999; Bozeman et al., 1998; Rogers and Bozeman, 1998). Bozeman and Rogers propose a ''pre-economic'' approach to evaluating scientific knowledge, one that does not purport to supplant economic evaluations but to complement them. Their approach is based on the range and repetition of uses of knowledge both by scientific colleagues and by technologists. One of the several advantages of this approach is that, unlike most evaluations that use the program or project as the unit of analysis, it incorporates the role of research collectives and networks engaged in knowledge production and use. In their view,

> innovation and knowledge flows cannot be assessed independently of the collective arrangement of skilled people, their laboratories and instruments, their institutions, and their social networks of communication and collaboration. Innovation gives rise to diverse scientific and technological outcomes including, among others, scientific knowledge, new technological devices, craft and technique, measures and instruments, and commercial products. But it is the collective arrangements and their attendant dynamics, not the accumulated innovation products, that should more properly be considered the main asset when assessing the value of research (Bozeman and Rogers, 1999).

Thus, for Bozeman and Rogers, the appropriate evaluation unit of analysis should be the set of individuals connected by their uses of a particular body of information for a particular type of application — the creation of scientific or technical knowledge. They term this set of individuals the ''knowledge value collective.'' In this scheme, science can be valued by its capacity to generate uses for scientific and technical information. That capacity is captured by the dynamics of the knowledge value collectives associated with use, as measured by their growth or decline, their productivity or barrenness, and their ability to develop (or deflect) human capital.

In many ways, this work builds upon the perspective developed by the Centre for Sociology of Innovation (CSI) in France, who have focused their evaluation activities upon the emergence of networks. One approach, applied mainly at the institutional level, uses the concept of techno-economic networks to characterise in a dynamic manner the nature, types and durability of alliances between public laboratories and companies, with particular emphasis on the ''intermediaries'', meaning documents, embodied knowledge, artifacts, economic transactions and informal exchanges (Callon et al., 1997). The emphasis in the CSI approach moves away from economic appraisal, focusing instead upon how networks are assembled to realize innovations in collective goods. This brief treatment cannot do justice to the developing work on collectives, but suffice it to say that it appears to offer considerable promise for capturing many of the payoffs from basic research, such as the development of human

capital, that thus far have been systematically underestimated or ignored.

Citations to scientific papers made in patents have been used as an indicator of growing linkage between academic science and industry (Narin and Norma, 1985; Narin et al., 1997). While supporting this indicator, Pavitt (1998, p. 109) has argued that patenting by universities gives ''a restricted and distorted picture of the range of the contributions that university research makes to practical applications.'' The explanation of this view lies in the complex nature of the relationship between universities and the corporate sector, involving flows of ideas and people.

Roessner and Coward (1999, forthcoming) are conducting an analytical review and synthesis of the research literature on cooperative research relationships between US academic institutions and industry, and US federal laboratories and industry. The scope of the review and synthesis was restricted to:

- empirical studies based on original or archival data only rather than theoretical, normative, or policy analyses;
- studies of research cooperation among US institutions;
- published or at least readily available reports and studies (i.e., conference papers, dissertations, theses, contractor reports, and unpublished papers generally were excluded);
- published or printed since 1986.

As of this writing, 33 studies have been reviewed of the approximately 50 to be examined. While only a handful of these studies were intended as formal program evaluations, most were intended to yield policy-relevant results. Surveys, case studies, and personal interviews accounted for most of the methods employed in these studies (22); two conducted quantitative analyses of archival data, two employed modeling, and the remainder were literature reviews or essays or used a combination of methods. As a general observation, the preponderant use of surveys, case studies, and interviews is an indication of both the recent emergence of cooperative R&D relationships as a subject of study and the complex, dynamic characteristics of these relationships.

The following three brief case studies of evaluations of programs to stimulate academic–industry linkages demonstrate both general lessons for methodological practice and the importance of the specific and individual context of each evaluation. All were subject to a mix of approaches (the normal case for evaluations) and demonstrate the strengths and limitations of each method employed in meeting the needs of different clients and stakeholders.

3.1. Evaluation of the Center for Advanced Technology Development (CATD)

The Center for Advanced Technology Development (CATD) was established in 1987 at Iowa State University (ISU) with funds from the US Department of Commerce. CATD seeks to bridge two gaps in the innovation process:

- between the research results of the university and the commercial market; and
- between a company's problems and the expertise resident at the university.

To address the first gap, CATD funds research intended to demonstrate proof-of-concept and develop advanced prototypes. To address the second, CATD can match funds provided by industry to address industrial problems.

An extensive evaluation of CATD was conducted in 1995–1996 (Roessner et al., 1996). The evaluation is significant for several reasons: there was sufficient financial support to conduct a relatively thorough study (a situation uncharacteristic of most single, university-based technology transfer organizations), there were multiple clients for the evaluation, and the evaluation design combined multiple strategies and types of data. There were three primary clients for the evaluation: the National Institute of Standards and Technology (NIST) of the US Department of Commerce, which paid for the evaluation, the Iowa State Legislature, and CATD staff. NIST was interested in knowing whether it should support long- or short-term projects. For the state, CATD needed to validate its successes and show evidence of success, as developed by an independent evaluator, to its constituencies. Finally, CATD management wanted to know which of its activities generated the greatest payoff, how its activities were perceived by its industrial clients, and what changes were needed to increase the payoffs from its activities.

A project with multiple clients, each asking different questions, called for a design employing multiple evaluation strategies. In addition, multiple strategies (and multiple measures of outcomes and impacts) would yield greater confidence in the results if the several strategies reached similar conclusions. The research team employed three research strategies:

- a survey of CATD client firms and a formal but simplified benefit/cost analysis, which would address payoff from investment questions;
- a series of detailed case studies, which would address the client perception and management questions;
- a ''benchmarking'' analysis involving similar programs, which would also address the payoff questions using a different method.

Of these, the benchmarking exercise was the most novel approach. Benchmarking programs such as CATD presents formidable challenges to analysts and evaluators, most of which have to do with the lack of data on programs that are truly comparable. For these reasons, the Georgia Tech team used three different approaches to benchmarking CATD outputs and impacts:

- data from the FY 1993 survey of members of the Association of University Technology Managers;
- four selected state cooperative technology programs; and
- results of a survey of companies conducting cooperative research with federal laboratories (Bozeman et al., 1995).

Methodologically, the case studies offered the best fit between client questions and the evaluation results. The least successful effort was benchmarking, primarily because it proved so difficult to identify comparable programs that had been evaluated at a comparable level of detail. The benefit/cost framework was required to address the questions posed by decisionmakers concerned about the future of the program itself, but the results offered no more than a modest justification for the public expenditures and little for purposes of research management. The case studies yielded three models of university–industry collaboration that could be associated with measures of impact. Managers could then decide, based on their preferred outcomes (e.g., jobs vs. increased private investment vs. startup companies), what project selection criteria to emphasize. The fact that quite different evaluation methods produced similar overall (positive) results strengthened the acceptability of each type of result.

3.2. Impact on industry of participation in the engineering research centers program

The National Science Foundation's Engineering Research Center (ERC) program, initiated in 1985, was intended to stimulate US industrial competitiveness by encouraging a particular brand of industry–university research collaboration (the university-based industrial research consortium), emphasizing interdisciplinary research, and fostering a team-based, systems approach to engineering education.

Evaluating such a program presents formidable challenges, especially if, as was the case here, some clients of the evaluation (e.g., Congress) wished to obtain information on benefits to companies in terms of dollar payoff of their participation in ERCs. The overall objectives of the evaluation were to examine patterns of interaction between ERCs and participating companies, and to identify results of that interaction in terms of types of impact and value of impacts (benefits to firms). These deceptively simple objectives posed several challenges because, among other reasons, the consequences of industry–university interaction take multiple forms over time, flow via multiple paths within the firm, and are likely to have intermediate effects not valued in monetary terms — but set in motion (or prevent) activities that do have such value.

In an effort to deal with these challenges in the initial research design, SRI International, the evaluator, first conducted a series of case studies within a small number of firms to provide details on intra-firm activities related to ERC participation (Feller and Roessner, 1995; Ailes et al., 1997). Next, drawing upon the case studies, a focus group consisting of representatives from companies that participate in ERCs was held. Third, the research team designed a survey instrument based on the results of the case studies, focus group, and model, and surveyed the more than 500 firms that were participating. Following the survey, SRI conducted telephone interviews with about 20 survey respondents who reported par-

ticularly detailed or varied benefits from their participation.

Results of the case studies of ERC-company pairs and the focus group required that ''typical'' approaches to survey design (e.g., how respondents are identified; the unit of analysis) be rethought, and that prevailing ideas about how firms value their participation in ERCs (as well as other investments in external R&D) are incomplete and simplistic. To be specific, the preliminary investigations suggested that:

- The ''known'' direct consequences (observable) to a firm of its participation in an ERC are extremely limited, often restricted to a handful of people in the firm.
- The business unit that pays the fee to participate (and is identified as the ''member'') may not be the business unit that receives the (observable) benefits from participation.
- Companies regard benefits received from participation in an ERC to be a function of their efforts to make use of ERC results and contacts, rather than of the amount of their membership fee.
- Firms rarely attempt to estimate the precise dollar benefits gained from their participation in an ERC.

These results had a number of very important implications for any effort to employ surveys to measure impacts or benefits from industry participation in collaborative R&D relationships with universities or with other R&D suppliers. First, the business unit, not the firm or even the division, is the active response category. Thus, the survey instrument must be directed to the unit that is the primary direct beneficiary of the company's participation in the ERC. Second, surveys must distinguish carefully between at least two key roles within the unit: the ''champion'' and the decision maker who approves the budget for ERC membership. Data must be obtained from both. The SRI team addressed this problem by identifying these roles as (1) the ''champion,'' the person who interacts most intensively with the ERC, and (2) the ''approver,'' the person who reviews or approves the budget for the unit's participation in the ERC. Third, the survey must establish the nature of each respondent's involvement with the ERC. Valid analysis of any impact or benefit data must control for the nature of each respondent's interaction. Finally, it is important to separate ''results'' or ''impacts'' from ''benefits,'' because different respondents within the same unit may agree on results or impacts but impute quite different levels of benefit to them.

Companies surveyed derived a multiplicity of benefits from their participation in ERCs, but few made a systematic effort to monetize these benefits or develop formal measures of payoff or cost-effectiveness in order to justify the cost of membership. Instead, they evaluated their participation in ERCs as a dynamic process rather than expecting a fixed set of benefits that can be assessed by present or retrospective rating of outcomes. Occasionally, specific outcomes were realized and an economic value estimated, but this was rare. Telephone interviews revealed that firms concluded that efforts to measure benefits in dollars would cost more than the company's investment in the ERC. Only a small proportion of companies developed a new product or process or commercialized a new product or process obtained as a result of ERC interaction, but those that did valued these benefits particularly highly. Thus, it proved to be extremely important to distinguish between whether a benefit was experienced and the value placed on that benefit.

3.3. Teaching company scheme

An example of a European study in this domain saw evaluations addressing what is widely perceived as a successful initiative, the UK's Teaching Company Scheme (TCS) (Senker and Senker, 1997). The TCS is a long-running initiative that provides access to technology for firms and facilitates academic–industrial technology transfer by employing graduates to work in a partnering company on a project of relevance to the business. However, the study found little evidence of a positive association between participation in the Scheme and the performance of academic departments. In fact, there was a higher positive association for departments involved in other forms of industrial linkage. From the perspective of evaluation methodology, this example illustrates the difficulty of isolating the effects of a single stimulus from a wide range of variables and indeed of finding suitable proxies for collective academic performance when linkages are still primarily at the level of individuals. In another evaluation study, the TCS

scored highest on a cost-effectiveness index devised by the UK National Audit Office to compare a range of programs which supported innovation. This index was based upon nineteen performance indicators which were intended to reflect a combination of the efficiency of delivering assistance and the measurable effect on innovation (National Audit Office, 1995). While the attempt to compare across widely differing schemes is laudable, the report serves mainly to illustrate the pitfalls involved in trying to capture complex effects with quantitative indicators, and the importance of implicit or explicit weighting of criteria. TCS did well because it involved low administrative overheads by comparison with schemes offering grants to firms. However, granting schemes have different objectives that may not be achievable through low overhead policy instruments.

4. Evaluation of collaborative R&D

For Europe, the emergence of publicly supported collaborative R&D between firms acting in research consortia (and usually with academic partners) was a distinctive feature of the 1980s, despite long antecedents in industrial research associations. The US was initially inhibited in this area by anti-trust legislation (Guy and Arnold, 1986), but private initiatives such as the Microelectronics and Computer Technology were followed by the Advanced Technology Program (ATP) in 1990. The latter was much more similar to its European counterparts in structure and aims, though somewhat smaller. It also differed in offering single company support targeted at small firms in addition to support for joint ventures.

From the perspective of this review, an important feature of such programs is the role they have played in the stimulation of research evaluation method and practice (Georghiou, 1998). The novelty of this policy instrument and the expectation that new types of effects would be significant provided a stimulus to program managers to seek a deeper understanding of their actions. Added to this has been an ongoing desire to demonstrate to policy-makers and other stakeholders that, on the one hand the programs are contributing to the competitiveness of firms, but that nonetheless, those firms have been induced to undertake R&D that would not have taken place in the absence of an intervention. For the ATP, a further hurdle has been the need to demonstrate to political critics that the program is satisfying the theoretical criteria for intervention by operating in the margin where private returns alone do not justify the R&D investment without the additional consideration of social returns. In Europe, questions of rationale have generally been dealt with at the ex ante stage. In most cases, it has been sufficient to demonstrate that economic and social benefits have been achieved. As in the previous section, methodological issues are discussed in terms of a series of case studies, selected both for the prominence of the programs involved and for the evaluation issues they raised.

4.1. The Alvey Programme

From the earliest days of evaluation of collaborative programs, it was clear that a broad range of effects needed to be considered. The evaluation of the UK's Alvey Programme for Advanced Information Technology (Guy and Georghiou, 1992) set the style for most subsequent UK national evaluations of collaborative R&D, as well as forming an input to broader European practice. That particular exercise was a real-time evaluation, tracking the program from shortly after its inception, with periodic topic-oriented reports fed back to the management, many of which addressed the complex process issues arising from the program, for example the problems with intellectual property conditions (Cameron and Georghiou, 1995). The final report, delivered 2 years after the initiative ended, found that Alvey had succeeded in meeting its technological objectives as well as its main structural objective of fostering a collaborative culture, particularly in the academic–industrial dimension. However, it had manifestly not succeeded in its commercial objective of revitalising the UK IT industry. The problem, the evaluators found, was in the objective itself. What was, despite its size, still an R&D program had been sold as a substitute for an integrated industrial policy with goals to match. For reasons that went well beyond technological capability, the market position of UK IT firms declined substantially during that period. This was perhaps the first example of a long-running evaluation problem in this sphere: such programs have multiple and sometimes conflicting goals that

are often not articulated in a format amenable to investigation through an evaluation. Hence, it was necessary to reconstruct the objectives in an evaluable form.

Since that time, the UK has developed the so-called ROAME (or ROAMEF) system, whereby prior-approval for a program requires a statement detailing the rationale, objectives (verifiable if possible), and plans for appraisal of projects, monitoring and evaluation (plus feedback arrangements) (Hills and Dale, 1995). This clearly simplifies the situation for evaluators but also emphasizes the need for evaluators to engage with policy rationales, a topic returned to below in the discussion on the evaluation of the ATP.

4.2. The EU Framework Programmes

For most Europeans, the development of evaluation has gone hand-in-hand with the history of pan-European collaborative initiatives, notably the European Commission's Framework Programmes and the inter-governmental EUREKA Initiative. The Framework Programmes have been evaluated in many different ways at the behest of a variety of stakeholders (Olds, 1992; Georghiou, 1995; Guzzetti, 1995; Peterson and Sharp, 1998). The legally based evaluation process has always been based upon convening a panel of independent experts for each sub-program and asking them to report upon its scientific, managerial and socio-economic aspects. Reports focused heavily on the first two criteria, confining themselves to generalized remarks on the third. Since 1995, this approach has been elaborated, following pressure from Member States, and now consists of continuous monitoring, reporting annually, and 5-year assessments, carried out midway through program implementation but including within the frame of analysis the preceding program (Fayl et al., 1998).

However, from a methodological perspective, the most interesting developments have taken place at the margins of this system or outside it altogether. The first significant development came from a study carried out in France that aimed to look at the impact of all EU R&D funding upon the national research system (Larédo et al., 1990). This was the prototype of a family of ''national impact'' studies which, using a survey and interviews, took a cross-cutting view of effects and revealed a number of unexpected results, for example the significance of European projects in providing a basis for doctoral training. By working at the level of the organisation rather than the project, some idea of the aggregate and interactive effects of participation could be gained.

Also from France came the best known European attempt to calculate the economic effects of programs upon participating organizations. Known as the ''BETA method'' after the laboratory at Strasbourg University where it originated, this approach is based upon in-depth interviews during which managers in the organizations are asked to identify ''added value'' from sales and cost-savings, and to attribute a proportion of this to their participation in the project in question. The most innovative aspect of the model is a formula for calculating indirect benefits that arise from transfer of technology from the project to other activities of the participant, from business obtained through new contacts or reputation enhancement gained via the project, from organizational or management improvements, or from the impact on human capital or knowledge in the organization. The principal difficulty surrounding this approach is that of attributing particular economic effects to a single project, when in many cases, the innovation may have drawn upon multiple sources from inside and outside the company concerned. The method produces ratios attractive to research sponsors because in a well publicized example (Bach et al., 1995) they show direct effects some 14 times greater than the initial public investment. The figure cannot, of course, be taken as a rate of return as it does not include in the calculation any other investments necessary to realize the returns.

Future directions for evaluation of the Framework Programme are discussed in a recent report from the European Technology Assessment Network (Airaghi et al., 1999). A new emphasis upon broader social objectives in the current Fifth Framework Programme implies the need to deal with a broader range of stakeholders and to enter the difficult territory of measuring the effects of R&D on employment, health, quality of life and the environment, all areas mainly driven by other factors. A second challenge is the need to deal with the sometimes elusive concept of ''European value-added'', in other words, to assess whether the R&D objectives were most

efficiently pursued at the European Community level, as opposed to national, regional, or indeed global levels.

4.3. EUREKA

Evaluation of the EUREKA Initiative has evolved largely independently from that of the Framework Programme. EUREKA is normally described as a ''bottom-up'' initiative in which firms may seek entry for any projects that meet the broad criterion of developing advanced technology with a market orientation. There is no central budget and only a very small central secretariat coordinating decisions taken by representatives of the Member States who choose whether or not, and by how much, to fund their own nationals' participation in collaborative projects. For the first decade of EUREKA's existence, evaluation took place optionally at the behest (and expense) of the Member State holding the Chair for that year.

After an administratively oriented panel exercise in 1991, the largest evaluation of EUREKA, and indeed of any European initiative to date, took place during 1992/1993. This involved teams from 14 countries working together, with a survey of all participants and interviews with about three participants from each of a selection of projects. Findings are in Ormala et al. (1993), and a description of the process appears in Dale and Barker (1994). This evaluation stood at a crossroads in terms of European approaches as it involved an explicit combination of tools developed during the Alvey and French impact studies, with further inputs from the Nordic and German evaluation traditions (described respectively in Ormala, 1989 and Kuhlmann, 1995). Not surprisingly, it has provided a template for many exercises that followed, within and outside EUREKA. In particular, the questions developed for the survey, on topics such as the additionality of collaborative research and on the relation of R&D to firm strategy, have been extensively reproduced and adapted.

Since 1996, EUREKA has adopted a new approach which, though conventional in the tools it uses (adaptations of the survey instruments from the 1993 evaluation), has been innovative in its structure (Sand and Nedeva, 1998; Georghiou, 1999). Known as the Continuous and Systematic Evaluation, this involves the automatic despatch of a questionnaire on outputs and effects upon completion of the project (this functions as a Final Report). Shorter follow-up questionnaires are despatched after one, three and (in the future) 5 years to participants indicating commercial effects. In addition a sample of some 20% of participants who completed 3 years before is interviewed. An annual report is prepared by an independent expert panel on the basis of these findings. While simple in concept, this is a rare example of an evaluation that systematically follows-up effects during the market phase. A further innovation by EUREKA in late 1999 was the convening of key participants in all past and present evaluations to review their collective findings and the degree to which these had been adopted. One key recommendation (the need for post-project support for small firms) had recurred in each evaluation. Its lack of adoption was seen as a measure both of its importance and of the lack of an adequate system to follow-up and learn from evaluation findings.

4.4. SEMATECH

SEMATECH, the US government industry research consortium founded in 1987, has been the subject of several academic studies, each seeking different measures of its effectiveness and cost-effectiveness. SEMATECH originally consisted of a consortium consisting of 14 US firms in the semiconductor industry (together accounting for 80% of US semiconductor component manufacturing) and the Defense Advanced Research Projects Agency (DARPA) of the US Department of Defense. Private investment in the consortium has typically totalled US$100 million annually, matched by government funds. As of 1998, SEMATECH is totally privately supported, and the number of member firms has declined.

The broadest assessment of SEMATECH was conducted by Grindley, Mowery, and Silverman, who published their results in 1994. The authors drew upon the available literature, SEMATECH archives, and interviews with SEMATECH managers and members of the technical advisory board. They discuss evaluation criteria, observing that these are particularly problematic because of several features

of research consortia generally and SEMATECH in particular:

- agreement among the several stakeholders on goals was lacking;
- similarly, agreement on appropriate evaluation criteria was lacking;
- the relevant time horizon for achievement of SEMATECH's broadest goals extends beyond the time of its existence (5 years at the time of this evaluation);
- it is extremely difficult to establish a counterfactual against which to assess the impact of what is basically a unique arrangement;
- SEMATECH has been highly flexible, changing its goals as conditions changed (Grindley et al., 1994, p. 736).

They conclude that flexibility in goals was crucial for SEMATECH's survival; that allowing for flexibility, its goals have been achieved; and that the political goals of achieving ''world leadership for the US semiconductor industry'' and ''industry competitiveness'' constituted unrealistic criteria against which the program could be judged.

Link et al. (1996) evaluated SEMATECH with considerably more restrictive criteria in mind. They sought to measure the returns to member companies from their investments in SEMATECH, restricting the returns to those accruing from research alone, excluding benefits from improvements in research management, research integration, and spillovers (Link et al., 1996, p. 739). They drew a representative sample of eleven projects and then interviewed the people in member companies who were most knowledgeable about each project. Many companies either estimated a range of benefits or said that they experienced significant benefits but could not estimate them accurately. Respondents also reported that intangible benefits related to research management, integration, and spillovers were more important than benefits related to research. The authors proceeded to use these benefit estimates and project cost data obtained from SEMATECH accounting records to calculate an internal rate of return for member companies. The ''best'' estimate of IRR was 63%, a figure that included both public and private funds invested in each project, burdened with an appropriate overhead figure, and with projects weighted by size. This study demonstrates that it is possible to obtain benefit estimates in quantified form from members of research consortia but, consistent with similar results from a variety of other studies of research collaboration and technology transfer, these estimates fail to capture the bulk of benefits association with consortium membership (Roessner and Bean, 1994; Feller and Roessner, 1995; Roessner et al., 1998).

The third evaluation, carried out by Irwin and Klenow and published in 1996, used similarly narrow evaluation criteria: basically several measures of member firm performance as compared with non-member firms. Although the major strength of this evaluation design is, in principle, the use of an implicit control group (via regression analyses using dummy variables to represent member/non-member status), in reality the fact that SEMATECH members are the dominant firms in the industry weakens the value of this usually desirable feature. (See the discussion of the Irwin and Klenow evaluation in the chapter by Klette, Moen, and Griliches in this issue.)

4.5. ATP

The US Department of Commerce's ATP is drawing substantial critical attention from policy-makers, and partly for this reason is also the supporter — and subject — of numerous formal evaluation studies. These studies are worth describing briefly here, less because any of them individually represents a methodological advance, but because of the sheer size and variety of the evaluation effort. In the words of the Director of ATP's Economic Assessment Office, which supports and conducts many of these evaluations, ''It [ATP] is probably the most highly scrutinized program relative to its budget size of any government program to date'' (Ruegg, 1998a,b, p. 3). The range of research topics being supported by ATP as of mid-1998 suggests a rich future source of (largely economic) analyses and evaluations of virtually every aspect of the program (Ruegg, 1998b, p. 8).

Papers prepared for a recent ATP-sponsored symposium on evaluation reflect the range of studies being conducted and a number of issues related to evaluation methods. The paper by Jaffe (1998) reviews the extensive past efforts to measure the social

rates of return from R&D and compare them with private returns. In the case of ATP, of course, measuring the average rates of net social return to a portfolio of projects may serve to justify (or undermine) the rationale for the program, but in the absence of predictive models directed at the individual project level, project selection decisions remain uninformed about possible spillover effects from individual projects. Identifying technological and market factors that tend to be systematically related to large positive differences between the social and private returns at the project level continues to be a formidable task facing economists.

One of the ATP program's objectives is to accelerate the development and commercialization of new technologies. Laidlaw (1998) surveyed 28 projects funded in 1991 to obtain estimates of the amount that ATP funding reduced development cycle time, and estimates of the economics of reducing cycle time by 1 year. Estimates of cost reductions or savings resulting from reduced development cycle time ranged from one million dollars to ''billions,'' suggesting that such estimates may be highly unreliable. In another project reported in the symposium, Link (1998) sought estimates of the reduction in research costs attributable to ATP funding, realized by seven companies participating in the Printed Wiring Board Research Joint Venture. Companies estimated research cost savings due to workyears saved and testing and machine time saved, and production cost savings due to productivity improvements. Total estimates are reported 2 years into the project and at the project's end. Link provides no discussion of the reliability of these estimates, however.

Researchers at CONSAD Research attempted to estimate the preliminary impacts on the national economy of an ATP-sponsored project whose objective was to control dimensional variation in automobile manufacturing processes. At the plant level, experts were asked to estimate the direct impacts of the newly developed technologies on the production processes across different assembly plants; other experts estimated the rate of adoption of the technologies by automobile manufacturers and the magnitude of the impact of the resulting increase in product quality on the sales of automobiles. A macroeconomic inter-industry model was then employed to estimate the impacts on the US economy of the increased demand due to quality improvements. The results showed cumulative effects over 5 years of US$8.7 billion in increased total industrial output and an increase of more than 160,000 jobs (CONSAD Research, 1998).

The Republican-led US Congress has been highly critical of ATP, and consequently has initiated a number of inquiries concerning its cost-effectiveness and political justification. Prominent among the formal evaluation efforts launched by the Congress is a large-scale, survey-based evaluation of ATP conducted by the General Accounting Office (GAO). An unusual aspect of the evaluation was the use of a comparison group, quasi-experimental design. In addition to 89 successful applicants for ATP awards, 39 ''near winners'' who had been scored as having very high scientific and technical merit, had satisfied the program's requirements, and had strong technical and business merit but had not received ATP funding were surveyed. The 128 winners and near winners were surveyed by telephone using a 76-item, closed-ended instrument. Data were obtained on applicants' efforts to obtain funding before applying to ATP, and if they intended to pursue their projects whether or not they received ATP funding. Near winners were asked about the fate of their projects after failing to obtain ATP support. The results were used by both critics and supporters of ATP to support their views (U.S. General Accounting Office, 1996).

If nothing else, the GAO evaluation demonstrates the difficulties associated with evaluating public programs intended to support pre-competitive technology development in private firms. While relevant and, apparently in the case of the GAO evaluation, reliable data on key questions about pre- and post-award project histories can be obtained from award winning companies and from a comparison group of companies, drawing conclusions about whether the program is addressing a market failure or not is a political rather than an analytical exercise.

The GAO experience has been echoed in similar European work on the issue of additionality — what difference did the intervention make. Traditional treatments have focused on input additionality — that is whether the incremental spending by an assisted firm was greater than or equal to the amount of subsidy. Papaconstantinou and Polt (1997, p. 11),

summarizing an OECD conference, concluded that several participants saw:

> a focus on additionality (the changes in behaviour and performance that would not have occurred without the programme) as a criterion for success is simply a reflection of the difficulty of accurately measuring spillovers or externalities and thus the net social benefits of programmes.

They present a view more consistent with the network models described earlier in this article: that the concept of 'behavioral additionality' (induced and persistent changes in the innovative behavior of firms) provides a sounder measure of the effects of intervention (Buisseret et al., 1995) in keeping with policy rationales founded in the notion of systemic rather than market failure.

4.6. Economic estimates of the net benefits of public support of technology development

Although our focus in this article is on methods for evaluating technology programs, in contrast to either more aggregate levels of analysis (e.g., policies) or less aggregate levels (e.g., individual projects or technologies), we would be remiss if we omitted reference to methods that economists have developed to measure the net social benefits of public support for technology development. The consumer/producer surplus method, pioneered by Griliches (1958) and Mansfield et al. (1977), uses estimates of producer and consumer surplus to measure the social and private rates of return to investment in particular technologies. Link and Scott (1998) have used what they term a ''counterfactual'' model to determine the relative efficiency of public vs. private investment in technologies with public good characteristics. In principle, these two approaches could be applied to the individual projects or technologies that comprise the portfolio of technologies in a publicly supported program of technology R&D, and the results aggregated to evaluate the entire program. The consumer/producer surplus approach would yield an estimate of whether, and by how much, the average social rate of return over all projects or technologies in the portfolio exceeds the average private rate of return. The counterfactual approach would yield an estimate of whether the average benefit/cost ratio summed over all projects is greater than one and, if so, by how much.

The counterfactual approach has been applied to several technologies supported by NIST (e.g., Link, 1996), but not, as far as we know, to overall programs such as ATP and SEMATECH that support multiple technologies. Both the Griliches/Mansfield and the Link/Scott approaches rest on industry-reported estimates of private investments in technology development. In addition, the consumer/producer surplus approach requires knowledge of the supply–demand parameters for process technologies, and must make fairly heroic assumptions to obtain equivalent estimates for new products. The counterfactual approach requires estimates by industry of what their costs would have been in the absence of government support (thus the label, counterfactual). Both approaches provide useful insights into the issue of net returns from public investments in support of new technologies, but since both rely, in part, on expressed preferences rather than revealed preference they embody a subjective element and must be used and interpreted with that subjectivity in mind.

5. Evaluation of diffusion and extension programs [1]

For several decades, industrialized nations have initiated a variety of programs intended to promote the diffusion of industrial technology within their borders. The primary purpose of most of these programs has been to enhance the competitiveness of target firms, which in turn is expected to increase the competitiveness — and, hence, the economic growth — of the nation or region in which the target firms operate. Some of these programs single out new, state-of-the-art technologies for diffusion; others emphasize best practice technologies and techniques; still others focus as much or more on business

[1] Although purists may argue that the terms signify differences among program goals or practices, for our purposes, the programs whose evaluation we discuss in this section can be labelled interchangeably as industrial modernization, industrial extension, or manufacturing extension.

practices, training, and marketing as on technology itself. In most instances, target firms are small- and medium-sized manufacturing enterprises (SMEs), typically defined as firms with fewer than 500 employees. Program staff seek to deliver a range of services whose number and mix vary widely across programs in the general terrain of training, advice to firms and promotion of linkages.

Each of the 50 states in the United States now has at least one industrial modernization/extension program. States have been the initiators of these types of programs, with significant industrial extension programs in several states dating back to the 1960s. Not until very recently did the US federal government begin to support a nationwide program of such programs under the Manufacturing Extension Program (MEP) within the Commerce Department's NIST. The MEP has an active evaluation program, and most larger evaluative efforts to date and the attempt to create a community of evaluation researchers focused on MEP-supported extension centers have been supported or encouraged by this group. Although state-supported programs have existed for decades, and the national effort is now ten years old, the state of the art in evaluating industrial modernization programs (at least in the US) is just beginning to evolve from a fairly primitive state. In a 1996 special issue of this journal devoted to the evaluation of industrial modernization programs, Shapira and Roessner (1996) observed that at the time there was considerable experimentation, innovation, and mutual learning, but ''to date there have been comparatively few systematic evaluation efforts in the industrial modernization field'' (p. 182). They cited the newness of the field, the small amount of resources devoted to evaluation, and lack of agreement on designs, measures, and who should manage the evaluations.

Feller et al. (1996) offer a harsh assessment of the reasons for the relatively primitive state of the evaluation art as of 1996. They discount the usual reasons for such situations, namely inherent conceptual problems or difficulties in obtaining data, claiming that in other program areas such as manpower training, medicine, and education similar issues have not forestalled the emergence of a substantial evaluation research literature, tradition, and practitioner community (p. 313). With few exceptions, they say, referring to the state of US evaluation, ''the ability to repeat the litany of formidable barriers has served as a convenient rationale for not conducting evaluations or as an umbrella justification for limited expectations or egregious shortcomings.'' In the remainder of this section, we focus on these few exceptions and on the progress that has been made in the last 4 years.

In their 1996 review of methods used to evaluate industrial modernization programs, Shapira et al. (1996) concluded that ''most of the evaluation methods used by industrial modernization programs are either indirect (program monitoring or external reviews) or implemented with one data point after service completion'' (p. 202). Cost and the lack of demand for more sophisticated evaluation strategies largely account for the situation; as Shapira and Youtie (1998) observed recently, there appears to be no direct correlation between the usefulness of an evaluation method with that method's degree of sophistication or use of controls. The state of the evaluation art is thus defined by a small number of studies that employ relatively rigorous designs: the identification and use of comparison groups and the use of statistical controls to achieve internal validity of results.

Identification of appropriate comparison groups [2] is a daunting task because, among other reasons, firms receiving services from industrial modernatization programs (clients) are self-selected, automatically distinguishing them in unknown ways from non-client firms. At least two prominent evaluations have used different approaches to identifying a comparison group of non-client firms. The Michigan Manufacturing Technology Center (MMTC), one of the first NIST manufacturing centers, has evaluated the impact of 5 years of service delivery to SMEs by comparing a set of performance measures collected from client firms to performance measures on the same variables from firms participating in the Performance Benchmarking Service (PBS), a separate ser-

[2] In this article, we use the term ''comparison'' groups rather than ''control'' groups to distinguish the use of quasi-experimental designs employing the logic of comparing firms that received services from similar firms that did not receive services, from true experimental designs in which treatment and control groups are selected at random from a population of potential target firms.

vice operating under the auspices of the MMTC. [3] PBS currently has about 3000 firms in its data set. Shapira and Youtie's evaluation of the Georgia Manufacturing Extension Alliance (GMEA) uses a different method of identifying a comparison group (Shapira and Youtie, 1998). The GMEA evaluation sent a ''control'' survey to all manufacturers in the state of Georgia with more than ten employees. The 1000 responses received were weighted to reflect the actual distribution of manufacturers by industry and employment size. Then GMEA client performance, measured by value added per employee, was compared with the weighted sample of all Georgia manufacturers.

Jarmin (1999), in an interesting effort to measure the performance of MEP client firms against non-client manufacturers, used the US Census Bureau's Longitudinal Research Database (LRD) to create a nonclient comparison group. Jarmin's evaluation used data from eight MEP centers in two states. He linked client data obtained from these eight centers to the LRD by using the Standard Statistical Establishment List (SSEL), which uses the same identifiers as the LRD. Then he linked clients to the SSEL by creating matching variables such as firm name and zip code from the several fields that are common between the two data sets. Jarmin controlled for selection bias, a serious problem in all such comparison group analyses, by estimating a Heckman-style two-stage model (Jarmin, 1999, p. 103).

The GMEA evaluation employed a very wide range of evaluation techniques to gather and analyze data, thus exemplifying perhaps the most comprehensive state-of-the-art in industrial modernization evaluation. In particular, they developed:

- a customer profile assembled by program personnel at the point of initial contact with a customer;
- activity reports that track field agent activities and customer interactions;
- client valuation surveys administered by mail to each customer when all major GMEA services had been completed;
- progress tracking via benchmarking and non-customer controls via the state survey of all manufacturers and a 1-year follow-up of GMEA customers; and
- a series of case studies to provide in-depth information about the effects of GMEA services on firm operations and profitability.

In addition to the obvious analyses of these data, Shapira and Youtie combined data on the private and public returns to GMEA-related investments and divided this figure by private plus public investment to obtain a benefit/cost ratio. Private returns included increases in sales, savings in labor, materials, energy, or other costs, reductions in the amount of inventory, and avoidance of capital spending. Private investment included estimates of the value of customer staff time commitment, increased capital spending, and fees. Public investment included federal, state and local program expenditures. Public returns were measured by federal, state, and local taxes paid by companies and their employees, estimated from sales increases or job creation/retention. The analysis accounted for zero sum effects (i.e., sales increases may be at the expense of other firms in the state) by adjusting sales to about 30% of the reported sales in the benefit/cost model.

The results of these evaluations raise two kinds of issues. First, there are considerable methodological issues. Most are related to the complexity of the programs themselves and to the inability of evaluators to identify a ''true'' control group of non-client firms. Second, there are political issues related to differences between what the evaluation community and the practitioner/policy community regard as the most useful and credible attributes of evaluations. With regard to the first issue, it has proven highly problematic to sort out the effects of widely varying lengths and types of services delivered to client firms, which are accompanied by widely varying levels of resource commitment by the clients themselves. Aggregating across any significant number of clients masks enormous variations in these two key variables. Further complicating this aspect of evaluations is the difficulty that firms have estimating the dollar value of industrial modernization program services for their operations.

The clients for evaluations of industrial modernization programs are as widely varied as the programs themselves. Federal officials, state-level politicians, and practitioners all seek and value dif-

[3] This description of the MMTC evaluation is based on Dziczek et al. (1998).

ferent kinds of information, and have different criteria of ''value.'' As Shapira and Youtie (1998) point out, US federal sponsors have little interest in measures of customer satisfaction, but instead seek measures of economic impacts. Information about the relative impact of different program services is of vital interest to program managers and field staff, but has little value for those charged with justifying the program's existence. Testimonials — basically superficial case studies — prove useful for program justification, but more formal cases, conducted using rigorous social science methods, have no greater impact for justification purposes. They can be highly valued, however, by program managers and field staff because they reveal the paths by which services are translated into changes in client operations and, in turn, into changes in client productivity.

European experiences of evaluating diffusion and extension programs are most extensive in Germany, as a consequence of that country's technology policy traditions. However, an assessment equally as harsh as those in the US has been made by Kuhlmann and Meyer-Krahmer (1995) about the state-of-the-art:

> The practice of evaluation of technology policy in Germany, when critically surveyed, does not nurture any naïve illusions about satisfying impact control. So far, evaluations have scarcely done more than provide evidence that technology policy interventions correlate with trends in technological and industrial development.

They conclude that progress is dependent, among other things, upon more objective data (less reliant on perceptions of participants), integration of evaluation with prospective methods for technology assessment in order to assess longer term unintended impacts of public intervention in technology. In response, a network of European evaluators has been attempting to unify evaluation with both TA and foresight approaches under the umbrella of a European Strategic Intelligence System (Kuhlmann et al., 1999).

6. Conclusions

In Section 1, we observed that evaluation methods and the methods used for more general technology policy studies are related — in other words, that broad social and political trends influence the practice of technology program evaluation. At certain times, evaluation studies have been at the leading edge of technology policy studies, for example in eliciting an understanding of collaborative R&D and of networking more generally. In other respects, they have tended to lag behind for methodological and political reasons. The demand for evidence of direct economic effects has left many evaluation studies coupled to a more linear view of the world than is common in the mainstream of technology policy studies. In general, though, evaluation may be seen as complementary to work carried out in other branches of the field, for example, more aggregated studies of national performance or more detailed work at the level of the organization. Evaluation work has probably had less of an impact in the literature than it deserves, in part because much of the most detailed and valuable work is not easily obtainable. There is a disturbing tendency for evaluation data that could form a valuable reference point for future studies to be lost in the grey literature, especially for those in countries other than the one in which the study was performed. Evaluators should make greater efforts to ensure that their main findings are also published in more conventional media. This problem also has inhibited the effectiveness of evaluation itself. Despite their limitations, most studies would benefit from a greater use of benchmarking and comparison group approaches. The use of archival data and multivariate methods should continue to be explored as one means of constructing analytical controls.

Looking at the broader role of evaluation in policy, some interesting changes are on the horizon. [4] In the US, passage of the Government Performance and Results Act (GPRA) in 1993 has galvanized the attention of the R&D community. GPRA calls for annual ''performance reviews'' at relatively high levels of agency activity, generally well above the

[4] For a comparison of research evaluation, practices in two different political settings, the US and Canada, and of the differences between GPRA and Canada's earlier, formal requirement that program evaluation be an integral part of the budget process, see Roessner and Melkers (1997).

program level. It also contains an ''escape clause'' that enables agencies supporting fundamental research to utilize qualitative measures of performance rather than the quantitative ones favored in the legislation. Program evaluation activity has increased in response to GPRA, under the apparent assumption that these evaluations will help support agency GPRA requirements. Appropriately or not, program evaluations and performance reviews are becoming entangled, and it appears to us that, as pressure to develop performance measures and justify program budgets increases as a result of GPRA, the lines between the two will become blurred. A similar tendency has also emerged in Europe, with monitoring and performance indicator approaches intersecting with evaluations, and in some cases, competing. This raises some important issues, not the least of which echoes our earlier observation that valid program evaluations must, increasingly, account for the systemic context in which they are performed. Technology programs do not exist in a vacuum, either politically or theoretically. The short-term yet aggregate perspective of GPRA's performance reviews conflicts with the longer-term, program-level, yet context-dependent perspective of technology program evaluations. The fundamental requirement for the design of a performance indicator regime is a clear understanding of context, goals and the relationships which link goals to effects. Whether this important distinction will be recognized and dealt with by government officials and the evaluation community remains to be seen.

Limitations of the production function approach to measuring the payoffs from research are now apparent, but promising new directions are just emerging. Return on Investment and Internal Rate of Return measures, despite their quantitative appeal, have limited value for program justification and even less for R&D program management decisions. Methods are needed that capture more fully the noneconomic benefits from research — or at least the benefits not easily translated into monetized form by those who receive them directly (e.g., private firms) or who seek to develop valid metrics (e.g., professional evaluators). We look forward to the outcome of efforts that focus on the characteristics of networks and linkages between knowledge producers and knowledge users as possible units of analysis, and measures of value, for the products of research. Methods that emphasize the dimensions of human capital and institutional development offer fertile ground for development.

Despite their widespread use as techniques for obtaining estimates of the benefits of research, and especially of linkages between research institutions, surveys of the intended beneficiaries are problematic. The studies we reviewed in this article document some of the difficulties: within the ''user'' organization, those benefiting most from the interaction between external sources of knowledge and technology may not be the same as those who interact directly with such sources and know the most about the interaction; the greatest benefits are longer term and qualitative rather than short-term and quantitative, and are thus difficult to estimate in monetized form; efforts to obtain and validate such estimates conflict with firms' proprietary concerns and with their priorities for allocating staff time (i.e., responding to surveys); the inherently risky and long-term nature of the innovation process itself often precludes reliable estimates of the payoffs from research. Where monetized estimates are obtained it must be recognised that these embody the (unknown) model of innovation and consequent attribution of benefits that the respondent is applying. The tendency to broaden the range of acknowledged stakeholders in R&D programs that has emerged recently in Europe provides another problem for survey approaches as many of these ''social'' groups are too distant from the research to provide detailed and structured responses. We are not advocating that surveys be shelved as appropriate evaluation tools, but we do urge that the subtleties now appearing in reported research be anticipated and incorporated in future evaluation designs. We also encourage, as mentioned above, the creative construction of comparison groups, as was done in the GAO evaluation of the ATP program, and the use of case studies as the preferred method for understanding what actually transpires in the complex process of technological innovation.

We made the point that progress in technology program evaluation methods has been incremental over the past 15 years. Evaluation, as the applied end of technology policy studies, will continue to grapple with real-world problems such as lack of clarity in objectives and the need to meet multiple and often

conflicting stakeholder needs. On the positive side evaluation is an adaptable beast and through a creative combination of methods has often managed to provide illumination and rational analysis in the most difficult of circumstances. It also continues to have much to offer to technology policy studies as a whole. There is a trade-off in which evaluators have to work within more restrictive terms of reference than other researchers, but in return gain access to a depth of data that would otherwise not be possible. They also provide a conduit by means of which the broader concepts of technology policy studies are carried into the policy arena, often at the highest levels. Our review suggests to us that there are a number of promising directions that evaluation methods might take that could lead to significant advances. We hope we are right, and that agencies supporting such work recognize the potential and act upon it.

References

Ailes, C.P., Roessner, D., Feller, I., 1997. The Impact on Industry of Interaction with Engineering Research Centers. SRI International, Arlington, VA, January 1997. Final Report prepared for the National Science Foundation, Engineering Education and Centers Division.

Airaghi, A., Busch, N.E., Georghiou, L., Kuhlmann, S., Ledoux, M.J., van Raan, A.F.J., Viana Baptista, J., 1999. Options and Limits for Assessing the Socio-Economic Impact of European RTD Programmes. ETAN, Commission of the European Communities, January 1999.

Bach, L., Conde-Molist, N., Ledoux, M.J., Matt, M., Schaeffer, V., 1995. Evaluation of the economic effects of BRITE-EURAM programmes on the European industry. Scientometrics 34 (3), 325–349.

Battelle Columbus Laboratories, 1973. Interactions of Science and Technology in the Innovative Process: Some Case Studies. Battelle Columbus Laboratories, Columbus.

Bozeman, B., Rogers, J., 1999. ''Use-and-Transformation'': An Elementary Theory of Knowledge and Innovation. School of Public Policy, Georgia Institute of Technology, Atlanta, GA (draft paper).

Bozeman, B., Papadakis, M., Coker, K., 1995. Industry Perspectives on Commercial Interactions with Federal Laboratories. Georgia Institute of Technology, School of Public Policy, Atlanta, GA, January 1995.

Bozeman, B., Rogers, J., Roessner, D., Klein, H., Park, J., 1998. The R&D Value Mapping Project: Final Report. Report to the Department of Energy, Office of Basic Energy Sciences. Georgia Institute of Technology, Atlanta, GA.

Buisseret, T., Cameron, H., Georghiou, L., 1995. What difference does it make? Additionality in the public support of R&D in large firms. International Journal of Technology Management 10 (4–6), 587–600.

Callon, M., Larédo, P., Mustar, P., 1997. Technico-economic networks and the analysis of structural effects. In: Callon, M., Larédo, P., Mustar, P. (Eds.), The Strategic Management of Research and Technology — Evaluation of Programmes. Economica International, Paris.

Cameron, H., Georghiou, L., 1995. Managerial performance — the process evaluation. In: Callon, M., Larédo, P., Mustar, P. (Eds.), The Strategic Management of Research and Technology — Evaluation of Programmes. Economica International, Paris.

CONSAD Research, 1998. Estimating economic impacts of new dimensional control technology applied to automobile body manufacturing. Journal of Technology Transfer 23 (2), 53–60.

Cozzens, S., Popper, S., Bonomo, J., Koizumi, K., Flanagan, A., 1994. Methods for Evaluating Fundamental Science. RAND/CTI DRU-875/2-CTI, Washington, DC.

Dale, A., Barker, K., 1994. The evaluation of EUREKA: a pan-European collaborative evaluation of a pan-European collaborative technology programme. Research Evaluation 4 (2), 66–74.

Dasgupta, P., David, P.A., 1994. Toward a new economics of science. Research Policy 23, 487–521.

Denison, E.F., 1979. Accounting for Slower Economic Growth: The United States in the 1970s. Brookings, Washington, DC.

Dziczek, K., Luria, D., Wiarda, E., 1998. Assessing the impact of a manufacturing extension center. Journal of Technology Transfer 23 (1), 29–35.

Fayl, G., Dumont, Y., Durieux, L., Karatzas, I., O'Sullivan, L., 1998. Evaluation of research and technological development programmes: a tool for policy design. Research Evaluation 7 (2).

Feller, I., Roessner, D., 1995. What does industry expect from university partnerships? Issues in Science and Technology, Fall 1995, 80–84.

Feller, I., Glasmeier, A., Mark, M., 1996. Issues and perspectives on evaluating manufacturing modernization programs. Research Policy 25 (2), 309–319.

Georghiou, L., 1995. Assessing the framework programmes — a meta-evaluation. Evaluation 1 (2), 171–188.

Georghiou, L., 1998. Issues in the evaluation of innovation and technology policy. Evaluation 4 (1), 37–51.

Georghiou, L., 1999. Socio-economic effects of collaborative R&D — European experiences. Journal of Technology Transfer 24, 69–79.

Gibbons, M., Georghiou, L., 1987. Evaluation of Research — A Selection of Current Practices. OECD, Paris.

Griliches, Z., 1958. Research costs and social returns: hybrid corn and related innovations. Journal of Political Economy.

Grindley, P., Mowery, D.C., Silverman, B., 1994. SEMATECH and collaborative research: lessons in the design of high-technology consortia. Journal of Policy Analysis and Management 13 (4), 723–758.

Guy, K., Arnold, E., 1986. Parallel Convergence. Frances Pinter, London, pp. 32–67.

Guzzetti, L., 1995. A brief history of European Union research policy, Commission of the European Communities, October 1995, pp. 76–83.

Hall, B.H., 1995. The private and social returns to research and development: what have we learned? Paper presented at AEI-Brookings Conference on The Contributions of Research to the Economy and Society, Washington, DC, October 3 1994, Revised 1995.

Hertzfeld, H., 1992. Measuring the returns to space research and development. In: Hertzfeld, H., Greenberg, J. (Eds.), Space Economics. American Institute of Astronautics, Washington.

Hills, P.V., Dale, A., 1995. Research Evaluation 5 (1), 35–44.

Illinois Institute of Technology Research Institute, 1968. Technology in Retrospect and Critical Events in Science. National Science Foundation, Washington.

Irwin, D., Klenow, P., High-tech R&D subsidies — estimating the effects of SEMATECH. Journal of International Economics, 40, 323–44.

Jaffe, A., 1989. Real effects of academic research. American Economic Review 79 (5), 957–970.

Jaffe, A.B., 1998. The importance of 'spillovers' in the policy mission of the advanced technology program. Journal of Technology Transfer 23 (2), 11–19.

Jarmin, R.S., 1999. Evaluating the impact of manufacturing extension on productivity growth. Journal of Policy Analysis and Management 18 (1), 99–119.

Kuhlmann, S., 1995. German government departments' experience of RT and D programme evaluation and methodology. Scientometrics 34 (3), 461–471.

Kuhlmann, S., Meyer-Krahmer, F., 1995. Practice of technology policy evaluation in Germany: introduction and overview. In: Becher, G., Kuhlmann, S. (Eds.), Evaluation of Technology Policy Programmes in Germany. Kluwer, Dordrecht.

Kuhlmann et al., 1999. Final Report of the Advanced Science and Technology Policy Planning Network (ASTPP) A Thematic Network of the European Targeted Socio-Economic Research Programme (TSER): Report to Commission of the European Communities Contract No. SOE1-CT96-1013.

Laidlaw, F.J., 1998. ATP's impact on accelerating development and commercialization of advanced technology. Journal of Technology Transfer 23 (2), 33–41.

Larédo, P., Mustar, P., 1995. France, the guarantor model and the institutionalisation of evaluation. Research Evaluation 5 (1), 1995.

Larédo, P., Callon, M., et al., 1990. L'impact de programmes communitaires sur le tissu scientifique et technique français. La Documentation Française, Paris.

Levy, D.M., Terleckyj, N., 1984. Effects of government R&D on private R&D investment and productivity. Bell Journal of Economics 14, 551–561.

Link, A., 1981. Basic research and productivity increase in manufacturing: additional evidence. American Economic Review 71 (5), 1111–1112.

Link, A., 1982. The impact of federal research and development spending on productivity. IEEE Transactions on Engineering Management EM/29 (4), 166–169.

Link, A.N., 1996. Evaluating Public Sector Research and Development. Praeger, New York.

Link, A.N., 1998. Case study of R&D efficiency in an ATP joint venture. Journal of Technology Transfer 23 (2), 43–51.

Link, A.N., Scott, J.T., 1998. Public Accountability: Evaluating Technology-Based Institutions. Kluwer, Boston.

Link, A.N., Teece, D.J., Finan, W.F., 1996. Estimating the benefits from collaboration: the case of SEMATECH. Review of Industrial Organization 11 (5), 737–751.

Mansfield, E., 1980. Basic research and productivity increase in manufacturing. American Economic Review 70.

Mansfield, E., 1991. Academic research and industrial innovation. Research Policy 20, 1–12.

Mansfield, E., Rapoport, J., Romeo, A., Wagner, S., Beardsley, G., 1977. Social and private rates of return from industrial innovations. Quarterly Journal of Economics 91 (2), 221–240.

Martin, B., Irvine, J., 1983. Assessing basic research: some partial indicators of scientific progress in radio astronomy. Research Policy, North-Holland.

Meyer-Krahmer, F., Montigny, P., 1990. Evaluations of innovation programs in selected European countries. Research Policy 18 (6), 313–332.

Narin, F., Hamilton, K.S., Olivastro, D., 1997. The increasing linkage between US technology and public science. Research Policy 26 (3), 317–330.

National Audit Office, 1995. The Department of Trade and Industry's Support for Innovation, HMSO HC715, London.

Olds, B.J., 1992. Technological Eur-phoria? An Analysis of European Community Science and Technology Programme Evaluation Reports. Beliedsstudies Technologie Economie, Rotterdam.

Ormala, E., 1989. Nordic experiences of the evaluation of technological research and development. Research Policy 18 (6), 343–359.

Ormala, E., et al., 1993. Evaluation of EUREKA Industrial and Economic Effects. EUREKA Secretariat, Brussels.

Papaconstantinou, G., Polt, W., 1997. Policy evaluation in innovation and technology: an overview. In: OECD Proceedings, Policy Evaluation in Innovation and Technology — Towards best practices, OECD, Paris.

Pavitt, K., 1998. Do patents reflect the useful research output of universities? Research Evaluation 7 (2), 105–111.

Peterson, J., Sharp, M., 1998. Technology Policy in the European Union. Macmillan, Basingstoke, p. 210.

Roessner, D., 1989. Evaluating government innovation programmes: lessons from the US experience. Research Policy 18 (6).

Roessner, D., Bean, A.S., 1994. Payoffs from industry interaction with federal laboratories. Journal of Technology Transfer 20 (4).

Roessner, D., Coward, H.R., 1999. An Analytical Synthesis of Empirical Research on U.S. University–Industry and Federal Laboratory–Industry Research Cooperation. SRI International, Washington, DC. Final Report to the National Science Foundation, (forthcoming).

Roessner, D., Melkers, J., 1997. Evaluation of national research

and technology policy programs in the United States and Canada. Journal of Evaluation and Program Planning 20 (1).

Roessner, D., Ailes, C., Feller, I., Parker, L., 1998. Impact on industry of participation in NSF's engineering research centers. Research-Technology Management 41 (5), 40–44.

Rogers, J., Bozeman, B., 1998. Creation of Knowledge Value: A Taxonomy of Knowledge Value Alliances. Paper prepared for presentation at Workshop on Research Evaluation, Ecoles des Mines, Paris, France, June 30–July 2.

Ruegg, R.T., 1998a. Symposium overview. Journal of Technology Transfer 23 (2), 3–4.

Ruegg, R.T., 1998b. The Advanced Technology Program, its evaluation plan and progress in implementation. Journal of Technology Transfer 23 (2), 5–9.

Sand, F., Nedeva, M., 1998. The EUREKA Continuous and Systematic Evaluation procedure: an assessment of the socio-economic impact of the international support given by the EUREKA Initiative to industrial R&D co-operation. Proceedings of the APEC Symposium on the Evaluation of S&T Programmes among APEC Member Economies, 2–4 December, Wellington, New Zealand, National Center for Science and Technology Evaluation, Ministry of Science and Technology of China.

Senker, J., Senker, P., 1997. Implications of industrial relationships for universities: a case study of the UK teaching company scheme. Science and Public Policy 24 (3).

Shapira, P., Roessner, J.D., 1996. Evaluating industrial modernization: introduction to the theme issue. Research Policy 25 (2), 181–183.

Shapira, P., Youtie, J., 1998. Evaluating industrial modernizatin: methods, results and insights from the georgia manufacturing extension alliance. Journal of Technology Transfer 23 (1), 17–27.

Shapira, P., Youtie, J., Roessner, J.D., 1996. Current practices in the evaluation of US industrial modernization programs. Research Policy 25 (2), 185–214.

Sherwin, C.W., Isenson, R.S., 1967. Project hindsight: defense department study of the utility of research. Science 156, 1571–1577.

Tewksbury, J.G., Crandall, M.S., Crane, W.E., 1980. Measuring the societal benefits of innovation. Science 209, 658–662.

US Congress, General Accounting Office. Measuring Performance: The Advanced Technology Program and Private-Sector Funding. GAO/RCED-96-47, January 1996.

U.S. Congress, Office of Technology Assessment, 1986. Research Funding as an Investment: Can We Measure the Returns? Office of Technology Assessment, Washington, DC.