PERGAMON

Evaluation and Program Planning 22 (1999) 55–64

EVALUATION and PROGRAM PLANNING

Using an evaluability assessment to select methods for evaluating state technology development programs: the case of the Georgia Research Alliance

Jan Youtie[a,*], Barry Bozeman[b], Philip Shapira[b]

[a] *Economic Development Institute, Georgia Institute of Technology, Atlanta, GA 30332-0640, U.S.A.*
[b] *School of Public Policy, Georgia Institute of Technology, Atlanta, GA 30332-0345, U.S.A.*

Accepted 1 August 1998

Abstract

Although increasing attention is being paid to all fields of government program performance, to what extent can the effectiveness of investments in programs involving technology-based economic development be addressed? Questions such as this raise the issue of the 'evaluability' of technology-based economic development programs—the degree to which the particular characteristics of the program affect the ability to provide effective evaluation. This article discusses how an evaluability assessment was conducted of the Georgia Research Alliance (GRA). The article presents the steps involved in conducting an evaluability assessment, including the development of an understanding of the structure and operations of the program, the perspectives of key stakeholders and participants as to potential program impacts and how these might be measured, and the evaluation of technology-based economic development programs in other states. Different methods through which GRA could be evaluated are analyzed and compared. © 1999 Elsevier Science Ltd. All rights reserved.

Keywords: Evaluation; State; Technology; Programs; Georgia

1. Introduction

U.S. states have been increasing their investments in technology development programs in recent years. From 1992 to 1995, state investments in university/non-profit centers, joint industry-university research partnerships, direct financing grants, incubators, and near-term assistance programs using science and technology for economic development grew by more than 32%, reaching $405 million in 1995 (Coburn & Berglund, 1995; Berglund, 1998). These state investments are augmented, in most cases, by multiple other funders including the federal government, industry, venture capital, consortia, and private sources.

The 1990s have also been a period in which more attention has been paid to government program performance. Thirty-five states have some type of performance-based budgeting initiative, either through legislation, executive order, or budget agency initiative. The field of technology development has not been immune from this growing desire for performance measurement. A recent survey of such programs found that 95% of states employ methods for collecting performance data or conducting program evaluations. But, despite the prevalence of some type of performance measurement or evaluation efforts among state technology development programs, few states have well-conceived evaluation plans. For example, activity reporting, client survey data, and informal client contact are the most commonly used evaluation methods (Melkers & Cozzens, 1996). More systematic evaluation approaches are less common. Only in part is this due to lack of funding or interest; there are also complex issues about how best to apply evaluation methodologies to assess the often diffuse and indirect effects of technology promotion policies.

This article reports on an evaluability assessment conducted of one of Georgia's major technology development programs—the Georgia Research Alliance (GRA). The objective was to examine and identify research approaches and strategies for evaluating the impacts of the Georgia Research Alliance and its associated program investments. This involved interviewing program managers, university administrators, research faculty, private sector partners, and state sponsors to

* Corresponding author. Tel.: 1-404-894-6111; fax: 1-404-894-0069; e-mail: jan.youtie@edi.gatech.edu

PII: S0149–7189(98)00041–X

develop an understanding of the structure and operations of GRA and the perspectives of key stakeholders and participants as to potential program impacts and how these might be measured. Information was also gathered about the evaluation of technology-based economic development programs in other states. Drawing on this research, we then analyzed and compared different methods through which GRA could be evaluated.

2. Methodological approaches to evaluating state technology development programs

Almost every methodological approach employed in the social and behavioral sciences has, at some point, been adapted to the purpose of evaluating technology policies and programs. Indeed, a body of literature has emerged that appraises the experience of using particular methods in the field of technology policy (see, e.g., Meyer-Krahmer, 1988; Evered & Harnett, 1989; Capron 1992; Bozeman & Melkers, 1993; Georghiou, 1995; Capron & van Pottelsberghe de la Potterie, 1997; Piric & Reeve, 1997). It is apparent that each evaluation method has specific strengths and weaknesses, both in general and for any specific application. In this section, we review the principal methods that have been used to evaluate applied technology development programs, with a concise assessment of each method's advantages and disadvantages. Subsequently, these methods will be viewed in terms of combinations, on the presumption that the weaknesses of one method can often be offset by using another in combination.

2.1. Case studies

Recently, there have been considerable advances in the use of case studies to determine the impact of specific state-sponsored R&D and technology development programs (Brown, Berry & Goel, 1991; Kingsley, Bozeman & Coker, 1996). These studies suggest that when applied with care, case studies can give an indication not only of the extent of program success or failure but the reasons for success or failure. A case study can also serve to document successes to stakeholders and funding agents. The great strength of case studies is they provide a sense of context and a richness of detail that exceeds virtually every other approach to analysis (Yin, 1989). The chief weaknesses of case studies are: (1) they are not typically cumulative; (2) their results are not easily communicated in a summary fashion for decision-makers; and (3) they require a good deal of interpretation. Finally, although case study methods are readily understandable, they require a good deal of expertise to perform, and can be quite expensive and call for the use of consultants. Case studies are an excellent complement to objective and quantitative techniques because the tradeoffs entailed in case studies are usually the mirror image of such quantitative techniques as cost-benefit analysis.

2.2. Peer review/external review

Peer review has long been recognized as a legitimate approach to evaluating scientific results (Kruytbosch, 1989; Chubin & Hackett, 1990; Bozeman, 1993). Uses of peer review for evaluating technology development programs include evaluating results of scientific papers to determine the scientific significance of the research and the evaluation of research programs and proposals. A somewhat different kind of review panel—an external review panel—has often been used to assess the management or economic development performance of research and technology programs. Panel reviews of this kind are used by agencies such as the National Science Foundation and the National Institute of Standards and Technology to assess the efficacy of major sponsored programs and centers. Panel selection can be problematic: individuals selected to serve on such panels must have no connection or beneficial interest with the program, yet at the same time must become familiar enough with the program in a short time to provide useful insights and recommendations. Peer reviews have great strength in the application of scientific research evaluation. They are also useful in assessing strategic, management, and operational performance. However, purely scientific reviews are weak mechanisms to identify economic development impacts, as most review participants have no claims on ascertaining these impacts.

2.3. Content analysis

Many research and technology development programs have an important but often not explicitly stated objective to change perceptions (e.g., state image, business climate). Content analysis is a method that can be used to answer questions about perceptions and image. There are several well-developed computer models for content analysis, but simple analysis of mass media references can be useful as well. Analysis involves examining number, direction, and favorability and changes over time. Although direct links cannot necessarily be determined, examination of the content of the references can provide findings beyond descriptive information.

2.4. Surveys

Survey research is widely used in evaluation studies. In assessing technology-based economic development programs, end-user surveys are especially pervasive and often prove quite useful (Shapira & Youtie, 1998). In the most complete evaluation designs, control groups of non-assisted companies are established (Cook & Campbell, 1979). Surveys are often used to try to develop measures of actual program outcomes (e.g., 'how many jobs retained?'). However, particularly for longer-term technology-based projects, companies often find it hard to accurately provide answers to such questions. Difficulties in implementing surveys include the high cost of conducting surveys, the potential problem of low response rates depending on the data collection technique (phone, mail, in-person), and how to deal with non-respondent bias.

2.5. Bibliometric analysis and other measures of research output

Since about the mid-1970s, there has been an explosion in the development of various techniques, including bibliometrics, for analyzing measures of research output such as citations of research publications, patents, and licenses. Recently, new software programs have been developed that are able to search and analyze citation indices and patent databases (Porter & Detampel, 1995). A number of studies have shown that citation analysis can be extremely useful for determining not only the value of research but also in charting its course (see Irvine & Martin, 1984). Citation techniques range from the simple—such as simply counting citations to a researcher's work—to sophisticated studies examining citation networks and citation communities, and providing weights and indices pertaining to the 'quality' of citations, citation 'decay curves' and so forth (Rip, 1988). Unpublished works and the time-lag between research production and publication notwithstanding, citation analysis is a good starting point for assessing the scientific impacts of research. However, from a state program perspective, it is hard to link citation measures of research impact to economic development outcomes. Similar issues arise with the analyses of patents and licenses. The links between patents or licenses and state economic development remain difficult to assess, with such measures being much affected by particular institutional intellectual property policies and the underlying character and maturity of specific research fields.

2.6. Cost-benefit Analysis (CBA) and Return-On-Investment (ROI) techniques

These methods seek to provide a framework within which a range of sometimes diverse benefits and costs can be arrayed and aggregated to provide estimates of net benefits and paybacks. (For examples of studies that have applied cost-benefit or return on investment techniques to technology programs, see Feller & Anderson, 1994; Nexus Associates, 1996; Roessner et al., 1996; Thompson, 1998.) The attractiveness of these techniques is that they provide a 'number'. Policy-makers are accustomed, especially in the field of economic development, to findings such as 'five dollars of economic benefit resulted from each dollar invested in the program'. The chief problem (assuming that reliable cost and benefit data can be collected—which is not universally true) is that these methods are highly sensitive to the particular assumptions of the model being used. Simple changes in the model can lead to drastically different results. Some factors that differ from one analysis to the next include: the measure of the opportunity costs for investment (discount rates), the degree to which and ways in which overhead investments and equipment costs are internalized; and, particularly, calculation of multiplier effects (which can take on the characteristics of 'numerical fiction'). None of this means that cost benefit analysis is invalid, only that it is of little use without in-depth understanding of the assumptions upon which the analysis is based.

2.7. Benchmarking

Currently a fashionable method of assessing impacts of technology-based economic development programs, benchmarking can be quite useful if its two core requirements are met: (1) identifying the appropriate benchmark programs and (2) developing valid benchmark measures. Neither of these requirements is easily met. In one sense, all programs are unique. The program elements, geographic setting, and implementation structures are never identical. Thus, the question becomes 'how similar is similar enough?' In addition, benchmarking analysis does generate numbers, but it is important to underscore the strong interpretive element involved in the method. Benchmarking is closer to case studies in its methodological character than it is to more quantitative approaches, although the common use of measures can easily disguise this fact. Benchmarking is most useful when making management and operational comparisons (see Wilkins, 1998).

2.8. Input–output analysis

Input–output models are often used to predict the additional household income, business revenues, state and local tax revenues, and employment associated with program-induced expenditures and business investment decisions. Input–output models generally have large data requirements if built from scratch, but once a model has been built, they can provide results quickly and outputs are easy to understand and communicate. Input–output models cannot address issues such as attribution, strategic advantage, business cycles, and changes in technology. Further, input–output models are often based on dated information, and they do not include an analysis of the impacts of the additional economic activity on the costs of providing state or local services, which increases the risk that interpretations will overstate impacts (Riall, 1991).

2.9. Systems and flow analysis

Although there is no generally agreed upon definition, flow analysis or logic models can be viewed as models requiring the specification of means and ends and showing paths from particular activities to particular outcomes. A wide variety of techniques are available for such analysis (see Behn & Vaupel, 1982, for an overview). Unlike other methods which require considerable professional expertise, charting the flow of activities is actually an activity better performed by program staff because, in the first place, they learn by doing and, in the second place, they have first-hand knowledge of the program elements and intended consequences. While this is a useful management analysis tool and is helpful for formative evaluation, it is not much help in measuring impacts. Despite its limits, it is an approach that is useful, inexpensive, and has immediate management applications.

2.10. Performance indicators

This term refers, quite simply, to the development of a set of critical indicators that can be used for program monitoring. This information allows counts to be made of aspects of the program, which are often requested by state sponsors and can be critical to understanding what the program is doing and what services it is providing (Shapira & Youtie, 1998.) Such indicators should be revealing of progress if not causality. Performance indicators can be program-specific, drawing activity and outcome indicators from the program's own management information and reporting system, or comparative, using government sources or rankings listed in business publications to obtain research, technology, and economic development indicators (Southern Technology Council, 1997).

2.11. Diffusion and network studies

There is a long history of diffusion studies in technology policy and, particularly recently, powerful techniques of network analysis are now available as a result of developments in various social sciences outposts (see, e.g., Shrum & Mullins, 1988; Tijssen & Korevaar, 1997). But an elementary approach that charts networks of users and providers can be useful to model a program's reach. This can be accomplished very easily by simply interviewing individuals as to persons encountered and sources of information. Elementary network analysis and charting of diffusion requires very little professional expertise and can be performed relatively inexpensively.

No single evaluation method is strong in all aspects. A method strong in one aspect (e.g., precision) is usually less strong in another (e.g., cost). The 'science' of policy evaluation is in the proficient application of method, but the 'art' is in the choice of method and in creative approaches to shoring up weaknesses while exploiting the strengths of methods. It is with this trade-off perspective that the evaluability assessment is made, in the programmatic context of GRA.

3. The GRA case: methods in practice

The GRA is a collaborative initiative among six research universities in Georgia to use research infrastructure invested in targeted industry areas to generate economic development results (see also the GRA's world-wide web site at *http://www.gra.org*). Research infrastructure investments in advanced telecommunications, environmental technologies, and human genetics are administered by three centers. GRA has several key programmatic elements. Eminent scholars in each of the three research areas are recruited to the university system based in part on a GRA supplementary endowment to be used for facilities, equipment, and other non-salary expenses. Three operating centers administer GRA funds, dispersing them to researchers at the six GRA universities, as well as engaging in auxiliary educational and policy initiatives and programs in their particular target research area. A recent addition—the Technology Development Investment program—funds the university side of industry-university collaborative research projects with significant commercial potential. GRA management acts as a 'holding company' for the program, developing strategy and finding financial resources. Between fiscal years 1992 to 1997, the state invested approximately $126 million in eminent scholar endowments and equipment and facilities (as well as leveraging additional federal and private funds). All GRA investments are channeled through participating universities and research centers.

GRA did not have a formal evaluation program in place at the time this study was conducted, although on alternate years the program conducts project-by-project reporting and monitoring of macro-level indicators such as research and development expenditures. As an effort to further its performance-based budgeting initiatives, the Georgia Governor's Office of Planning and Budget (OPB) desired that evaluative information about GRA's impacts be developed. They asked the authors to recommend methods and approaches for evaluating GRA, which would then be implemented by GRA or outside evaluators.

3.1. Evaluability assessment

The authors determined that OPB's request could best be addressed by conducting an evaluability assessment. An evaluability assessment examines the degree to which the particular characteristics of the program affect the ability to provide effective evaluation (see also Wholey, 1994). An evaluability assessment asks, to what extent is evaluating the program's effects actually possible? What should be measured? And, what actually can be measured, given the constraints of the resources that can reasonably be allocated to evaluation? The following elements and factors that affect GRA's evaluability were identified.

3.1.1. Differences in stakeholder perspectives

GRA has several distinct groups of stakeholders: program management; university administrators; research faculty; private sector partners; and state sponsors. Each of these groups has somewhat different perspectives as to what the key measures of GRA are, or should be. For example, the state, in general, aims to promote economic development, and therefore emphasizes job-related measures. Universities hope also to improve their recognition (measured through discoveries, patents, and publications), and desire facilities to support students. Private companies usually seek market success, sales, and profitability (not job creation, per se). Thus, although many state technology development programs focus on job-related measures as indicators of impacts, other stakeholders pursue different measures of success.

3.1.2. Time-lags

Most of GRA's investments in technological capability can be expected to have significant results only over medium-to-long time horizons (of perhaps 7–15 years). While costs and some initial benefits are evident now, in many cases it is still rather early to judge the full range of benefits, spillover effects, and even the full magnitude of costs.

3.1.3. Indirect links between program intervention and desired program outcomes

GRA 'intervenes' by making additional investments in facilities and equipment available for researchers, including eminent scholars, at research universities. The program's desired outcomes, with variations according to different stakeholders, are focused on attracting increased research funding, state economic development and improvements in the perception of Georgia as a location for technology-oriented companies. However, particularly for the last two objectives, the link with facility and equipment funding is indirect. First, making the link between program investments and outcomes requires information about specific GRA program investments available through the GRA program office. Information is also available on matching direct investments, including those by private organizations, and on additional grants attracted to GRA facilities. Less information is readily available on sources and streams of resources flowing to GRA centers and projects, especially the flow of matching and complementary resources. Second, making the link requires the occurrence of a series of additional downstream steps before changes occur. For instance, for economic development effects to materialize, technologies have to be transferred and commercialized, private companies established in Georgia, production started, sales generated, and new jobs created. While GRA staff pay attention to these links, program funds cannot be expended here and the program itself has, at best, rather indirect and limited influence on the key downstream steps. The 'attribution' to GRA investments of changes in downstream outcomes thus presents difficulties to evaluators.

3.1.4. External factors

GRA's investments are targeted towards three main industrial fields: advanced telecommunications; environmental technologies; and biotechnology. The development of each of these industrial areas in Georgia is affected by a series of contingencies and broader forces, of which GRA is but one. For example, regulatory changes, shifts in federal R&D priorities, other state and local economic development policies, business cycles, the availability of local downstream investment capital and entrepreneurial management skill, technological developments elsewhere, and the growth of market demand are among the factors which can greatly affect whether GRA's investments yield economic development returns to the state. Sorting out the effects of program investment in the context of broader industry, economic, and technological change is a complex and uncertain task.

3.1.5. Difficulty of developing counter-factual evidence and controls or benchmarks

Ideally, program evaluation design should incorporate elements that can consider counter-factual evidence and arguments, i.e. what would have happened without the investment of program funds? For some technology-based programs, it is possible to establish control groups which can provide a basis for understanding what happens to those not served by the program, attempting to hold other key variables constant (Jarmin, 1998). In other situations, alternative explanations for outcomes can be probed through logic-driven questioning and evidence collection designs (Cosmos, 1996; Youtie, 1997). It might be possible to examine GRA in these ways, but not without complications. For example, how can participants and non-participants be reasonably distinguished and their performance then controlled? Is the unit of participation the six GRA research universities? In this case, there are no other comparable institutions in Georgia. If eminent scholars are chosen as the participants, what is the control group? Other faculties in GRA universities? But, is this a fair control, as eminent scholars are preselected to be among the very best in their field? Moreover, by design, eminent scholars and GRA research facilities aim to promote collaborative research teams within specific technological fields. How can the roles and effects of others in the team be separately determined?

It is also evident that while some program inputs and outcomes can be measured, other effects are hard or even impossible to gauge. Similarly, there are likely differences in the degree of confidence that can be expressed as to the causal associations between program investments and particular outcomes. Thus, specific additional research awards can be measured (in dollars) and possibly attributed to GRA investments. But it is likely to be much harder to separate GRA's impact (from other factors) on any overall increases in research funding in Georgia (although we may again know by how many dollars research funding has increased). Equally, while it may be possible to track the transfer of specific new technologies developed in GRA centers out to particular companies, the companies themselves often find it very difficult to precisely estimate benefits and costs, particularly if the technology is still at an early stage of commercialization (Roessner et al., 1996). Finally, such elements as effects on overall economic development or improvements in the 'image' or 'research perception' of the state are harder both to measure precisely and to causally associate with the program. Strategies and approaches to overcoming some of the challenges affecting the evaluability of GRA are discussed below.

3.2. Recommended approaches

In conducting our evaluability assessment, we considered the eleven major methodologies discussed earlier and their applicability to the context, needs and constraints facing the GRA. The criteria we adopted to distinguish between the different options included the level of technical training and expertise needed to implement the method; the method's validity (power to ascertain causal relations) and reliability (test/re-test correspondence); the resources required; and timeliness. We also took into account other factors such as the method's orientation ('summative' vs 'formative'), likely perceptions of the credibility of the method, the 'usability' of the results of a method, and ease of communicating the results of a method. Informed by the background case interviews and our experience and knowledge of how these methods had been used in comparable studies, we used our best judgment to assess the applicability of each method (see Table 1).

Based on this assessment, we recommended two different evaluation regimes for assessing GRA and its activities. We referred to these as 'routine evaluation' and 'comprehensive evaluation'. Routine evaluation implies the investment of modest resources and does not require expertise of external consultants. Comprehensive evaluation is more thoroughgoing but requires greater resources and the use of credible external evaluation consultants. We recommended that each be pursued, but at different intervals. Routine evaluation should be performed annually or bi-annually; comprehensive evaluation should be performed on a longer cycle, of three to four years.

3.2.1. 'Routine' evaluation

Typically, valid evaluation requires considerable technical expertise and commitment of substantial resources. But often it is possible to engage in useful evaluation activity even when evaluation is performed on a modest budget and by persons who are not highly trained evaluators. In the GRA context, two types of useful evaluation activities can be performed at very little expense and without the need for great evaluation expertise. We recommended that GRA be evaluated regularly (on an annual or bi-annual basis) on the basis of (1) performance indicators; and (2) flow analysis. The use of any evaluation or management decision tool exerts some costs. But the amount of direct outlays required to adopt these approaches to routine evaluation should prove minimal.

Performance indicators would provide base-line data for management and evaluation, to help managers and other stakeholders track and monitor progress. These same performance indicators can be useful as part of more comprehensive evaluations employing rigorous analysis of data. The set of performance indicators adopted should have widespread support among stakeholders and should be amenable to routine collection. Ideally, agreement should be sought on a manageable number of priority performance indicators.

As a management tool, we also suggested that GRA principal researchers and managers adopt some form of flow analysis or logic model. This analysis would be used to chart 'how the program is supposed to work' to understand how expectations and reality diverge. The analysis would chart specific targets and expected paths between the production of output and the dissemination to targeted users, even where there is an intention that research or program outputs diffuse to a user community. It is often assumed that good, technically viable work will, as a matter of course, be used. Evidence shows that this is sometimes not the case. Consistent tracking through flow analysis would assist in project monitoring and management and in identifying the conditions and future steps that will affect subsequent technology diffusion.

3.2.2. 'Comprehensive' evaluation

As a second part of the evaluation regime, we recommended that GRA be subject to comprehensive evaluation. While each of the recommended methods presented in Table 1 was worthy of consideration, we recommended that a good balance would be provided by using (in addition to the methods employed in the routine evaluation): (1) a survey-based cost benefit analysis; (2) case studies; (3) content analysis; and (4) external and peer review. By using these approaches in combination

Table 1
Assessment of methods available for the Georgia Research Alliance (Applicability: H = High, M = Medium, L = Low)

Method	Technical needs[i]	Validity and reliability[ii]	Resources[iii]	Time need[iv]	Overall rating (including other factors)[v]
Case study	H	H/L	H	H	M
Peer review	L/M	M/M	M	L	H
Content analysis	H	H/H	M	M	M
Bibliometrics/other research output measures	H	M/H	M	M	L
CBA/ROI	H	M/M	M	M	M
User survey	H	H/H	M	M	H
Benchmark	M	M/M	M	M	L
Network analysis	H	H/H	M/H	H	L
Input–output	H	L/H	M/H	M/L	L
Systems/flow analysis	L	M/M	L	L	H
Indicator systems	L	H/H	L	L	H

[i] Technical needs refers to the degree of technical training and expertise required to perform the method. Generally, 'High' means that the method should be performed by experts, 'Low' means it can be performed by program personnel, and 'Medium' means it can be performed chiefly by program personnel with some consultation from experts.
[ii] Validity refers to the power of the method to ascertain the casual relation in hypotheses about program effects; reliability refers to test–retest correspondence.
[iii] Generally, the approaches are not equipment-intensive and, thus, resources chiefly refers to number of person hours required of evaluators (not including others' time for providing data, e.g., questionnaire respondents).
[iv] Time refers to amount needed for the evaluation study, not for the program to have its required and measurable effects.
[v] The rating is based on subjective assessment of the suitability of the technique for the purpose of evaluating GRA. Criteria include, but are not limited to, the factors explicated in the table. Other factors included in the assessment: 'summative' (chiefly for final program effects) vs 'formative' (useful for program improvement in an ongoing evaluation) orientation; likely availability of expertise; likely perceptions of the credibility of the method; the 'usability' of the results of a method; and ease of communicating the results of a method.

for a comprehensive evaluation, one could ensure an in-depth portrait of program activities with attention to the details that contribute to success (through case studies); provide for objective monetary-based impacts (by using cost benefit analysis); give insight into the crucial issue of changes in perceptions generated by the program (by using content analysis); and examine the overall strategic direction, research, operation, and management of the program (through external and peer review panels).

To properly conduct a cost-benefit analysis, it would be necessary to conduct interviews and survey customer firms (usually by mail) to define, identify, and attribute appropriate treatments for the full set of costs and benefits. A significant problem with most survey research efforts, especially mailed surveys, is response bias. Often there are systematic and relevant differences between companies responding to a questionnaire and those who choose not to respond. In addition, some customers—particularly smaller companies—can find it difficult to estimate quantitative outcomes. Several techniques have been developed for documenting and adjusting for response bias. If used with caution, we suggested that cost-benefit analysis could be useful for a wide range of GRA programs and, particularly, can provide a convenient index of the economic impacts of GRA programs. One inexorable problem is that benefits and costs for GRA occur in different streams: investments are being made now for benefits that may take many years to fully materialize. If models do not properly address lag effects and time horizons, the program's effects can easily be under- or over-valued. A danger of cost benefit analysis, especially for the unsophisticated user, is that it appears extremely precise and 'scientific', but, in fact, is subject to a number of judgments in the adoption of assumptions.

Regarding case studies, many programs similar to GRA have found it useful to have 'success stories' that can be presented to stakeholders and funding agents. But these success stories are usually poorly documented. A robust case study (for instance, along the lines advocated by Yin, 1989) can serve the same function but, at the same time, can give useful information about why the success was generated and can provide a better documentation of the success. Case studies would be particularly useful in the GRA context because many of the program elements are integrated and not easily sorted out in quantitative methods. Also, some of the impacts of GRA are as much perceptual as tangible and case studies are useful in analysis of perceptions.

Content analysis would examine the important but always elusive question—what is the impact of GRA and its program on perceptions and images of the Georgia business climate and, in particular, the state as an attractive location for technology-intensive businesses? The latter is important, since elected officials have supported GRA funding at least in part on the assumption that it will raise the state's profile among competing high technology regions. A simple analysis of mass media references to Georgia business can be used as a means of addressing this issue. By examining the number, direction, and favorability of such references it is possible to chart changes in the image of the Georgia business climate. While it is not possible to determine the exact contribution of GRA to changes in image, such an approach at least provides some valuable descriptive information. Furthermore, by examining the content of references to GRA it may be possible to make at least some direct connections.

Peer review provides valid, credible evaluation of scientific research. Peer review has particular applicability to GRA's eminent scholar program, and each major GRA program could be submitted to peer review about every four years. The primary use of peer review should be evaluation of the quality of the scientific work and the peer panels should be comprised of external scientific experts with intimate knowledge of the scientific fields addressed by GRA researchers. Scientific peer review is much less useful in determining the economic potential of scientific work, in which case an 'external review' would be useful in assessing the management efficacy and utility of GRA (and, indeed, has already been performed for some GRA programs). Here, a panel of public-sector, industry, other private-sector, and academic individuals with expertise appropriate to research management would assess GRA's strategic direction, operations and methods of investment. The periodic use of peer review and external review will provide an assessment of scientific quality and the relevance of quality scientific and technical outcomes to the long-term economic development goals of GRA.

To provide a basis for comprehensive evaluation, we recommended that OPB and GRA consider setting aside a small percentage of program funds to devote to evaluation. This is a common practice elsewhere and has led to the production of high quality, highly usable evaluations. Two familiar examples are the resources invested by NIST and state manufacturing assistance programs and for the Department of Energy's Energy Related Inventions Program (Brown, Curlee & Elliott, 1995). With a sum set aside, resources would be available every three to four years for a comprehensive evaluation.

3.3. Relevance to evaluability assessment

As part of our study, we considered how our recommended evaluation regime for GRA related to the issues and challenges identified by the evaluation assessment. This section examines the elements of the routine and comprehensive evaluations in light of the principal issues raised in the evaluability assessment.

We judged that the need for a range of indicators to address differences in stakeholder perspectives about what constitutes success could be accommodated by the adoption of performance indicators. To build consensus about which particular indicators to use, we recommended that stakeholders and program participants should be involved in assessing indicators before adoption. In addition, all elements of the evaluation should be as unobtrusive as possible and mindful of other demands on program participants' time, performed with minimal data gathering responsibilities on program participants.

Many GRA program effects will be realized only in the long term—that is in up to fifteen or even more years. Routine evaluation will not be adequate to capture long-term effects, but the establishment of a baseline set of performance indicators, measured yearly, will provide the ability, in the long run, to measure such change. Also, assumptions and approaches used in cost benefit analysis incorporate approaches to dealing with time-lags and streams of benefit. Nevertheless, the length of time required for benefit will always be an evaluation constraint and should be given considerable attention.

We noted earlier that GRA program outputs occur within a broad environment and many other factors in that environment can be expected to have much greater impacts on such factors as creation of new business, including changes in the interest rate, availability of capital and labor, and other such factors. While the attribution of outcomes is always difficult, case studies are particularly useful for understanding chains of causality. It is widely recognized that changes in perceptions (e.g., state image, business climate) are among the more important contributions of technology-based economic programs, but this vital aspect is rarely evaluated. While the case study approach can be quite useful for analysis of perceptual change (e.g., by determining reasons for business start-up or relocation in case studies performed), the use of content analysis for these purposes could be an important evaluation innovation. Carefully implemented content analysis can reveal differences that occur over time in media perceptions. While such changes probably occur relatively slowly, the three-to-four year evaluation period suggested for the comprehensive evaluation should provide sufficient time for some measurable change.

While job creation and retention data are the most familiar and easily understood measures of economic change, we cautioned against the use of job creation data as the sole indicator of economic change, except perhaps as a minor component of an indicator system. If there is an interest in using jobs as an indicator, then evaluation should focus on audited jobs rather than unsubstantiated self-reports.

In our proposed evaluation approach, we did not recommend that GRA undertake a control-group study or a formal benchmarking process. In the first place, we are not able to identify programs or research projects that are entirely satisfactory comparisons for GRA. In the second place, the primary objectives of benchmarking can be achieved with other methods. Case studies can be used for in-depth (but non-statistical) comparisons between GRA and comparable programs. The set of performance indicators can be used as a de facto benchmark, assuming the availability of comparable information from other states.

4. Conclusions

Our evaluability assessment of GRA highlights the factors that need to be considered when selecting program evaluation approaches. Selection among the array of possible program evaluation methods needs, on the one hand, to consider the methodological strengths and weaknesses of any particular approach, and, on the other, program context, resources, and needs. Program characteristics also greatly affect the ability to conduct an effective evaluation. In the GRA case, these characteristics included differences in stakeholder perspectives, time-lags before realization of the full range of benefits and costs, indirect links between program intervention and desired program outcomes, the influence of the broader industry, economic and technological context, and the difficulty of developing counter-factual evidence and comparisons.

Our approach, in the GRA case, was to recommend the use of selected methods within a two-part evaluation regime. We proposed the use of routine methods that could be readily applied to a research and technology development program. Performance indicator and flow analysis can be performed on a modest budget and, to a large extent, through internal program resources to track and monitor progress and establish how the program should work to produce this progress. In addition, we suggested that more extensive evaluation methods be applied over a longer time frame to gauge the program outcomes—quantitative and perceptual—and to review the program's strategic direction, research, operation, and management. Case studies, survey-based cost-benefit analysis, content analysis, and external and peer review require devoting a share of program resources and the employment of external expertise, but they more comprehensively address various factors that affect GRA's performance and impact.

Although an evaluability assessment addresses constraints on program evaluation and methods useful for addressing these constraints, our evaluability assessment did not go into such details as the specific questions to be asked on survey instruments, the design of management information systems, or the names of outside consultants or review panel members. Our intent was to meet the client and program's need for an initial and broad overview of how they might approach evaluation in a situation where little formal evaluation had been conducted to date. In other contexts, an evaluability assessment might go more fully into the details of specific methodologies.

This point about the GRA assessment suggests that the application of evaluability assessment to technology development initiatives elsewhere will vary depending on the context of the program and its experience in conducting evaluation. Evaluability assessments are typically thought of as being conducted at the start of a new program. In this context, the assessment may be limited to an overview of various evaluation approaches or it may provide more detailed recommendations as to measures, data collection methods, analytic approaches, and reporting. The portfolio of methods examined would depend on the particular nature of the policy or program under review. The GRA case represents a second context in which the program is already in operation, but no formal evaluation system exists. Based on our experience with the GRA, an evaluability assessment of an ongoing program may require exploration of various evaluation approaches and existing program constraints rather than specification of evaluation details, particularly if the party requesting the evaluability assessment is not the program manager. The request for the evaluability assessment of the GRA came from the state Office of Planning and Budget, acting in an oversight capacity and seeking a broad review of possible approaches, rather than from GRA program administrators wanting to implement specific evaluation methods.

A third context involves situations where there is already an operational program and evaluation system. The evaluability assessment of an ongoing program may take the form of a review of the implementation and usability of existing methods. Under such conditions, evaluability assessment may consider changes in external pressures (budgetary, sponsor expectations) and internal needs for program improvement. In addition, such an assessment would be especially concerned with the validity and reliability of the data generated by existing methods to determine whether existing methods should be modified or possibly abandoned in favor of new approaches. Where there are several related programs with ongoing evaluation systems, an evaluability assessment may take the form of a portfolio evaluation review which assembles valid comparative information.

In summary, evaluability assessment is a concept that offers many benefits in designing and implementing an evaluation system. An evaluability assessment can help to address concerns about conducting an evaluation. Through evaluability assessment, program participants, management, sponsors, and partners can be explicitly involved in dialogue about evaluation at an early stage (or, in situations where evaluation has lagged, to prompt dialogue and action about establishing evaluation while in mid-term). Evaluability assessment can help avoid unnecessarily costly investments or omissions of important methods. Evaluability assessment also forces consideration of the balance of methods in light of their relevance in the context of political, technological, and methodological issues. Applied appropriately, an evaluability assessment can assist programs and sponsors to initiate more effective evaluation efforts or make their first steps into more formal evaluative approaches in cases where evaluation has not previously been conducted.

References

Behn, R., & Vaupel, J. (1982). *Quick analysis for busy decision-makers*. New York: Basic Books.

Berglund, D. (1998). State funding for cooperative technology programs. Unpublished manuscript. Columbus, OH: Batelle Memorial Institute.

Bozeman, B. (1993). Peer review and the evaluation of R&D impacts. In B. Bozeman, & J. Melkers (Eds.), *Evaluating R&D impacts* (pp. 36–49). New York: Kluwer.

Bozeman, B., & Melkers, J. (Eds.). (1993). *Evaluating R&D impacts: Methods and practice*. Boston, MA: Kluwer.

Brown, M. A., Berry L. G., & Goel R. (1991). Guidelines for successfully transferring government-sponsored innovations. *Research Policy*, 20, 121–143.

Brown, M. A., Curlee, T. R., & Elliott S. R. (1995). Evaluating technology innovation programs: The use of comparison groups to identify impacts. *Research Policy*, *24*(5), 669–685.

Capron, H. (1992). *Economic quantitative methods for the evaluation of the impact of R&D programmes. A state of the art*. Brussels, European Commission: Monitor-Spear Series.

Capron, H., & van Pottelsberghe de la Potterie, B. (1997). Public support to R&D programmes: An integrated assessment scheme. In *Proceedings—Policy Evaluation in Innovation and Technology: Towards Best Practices*. Paris: Organisation for Economic Cooperation and Development.

Chubin, D., & Hackett, E. (1990). *Peerless science: Peer review and U.S. science policy*. Albany, NY: State University Press.

Coburn, C., & Berglund, D. (1995). *Partnerships: A compendium of state and federal cooperative technology programs*. Columbus, OH: Batelle Memorial Institute.

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Cosmos Corporation (1996). *A day in the life of the manufacturing partnerships: Case studies of exemplary engagements with clients by MEP centers*. Gaithersburg, MD: National Institute of Standards and Technology.

Evered, D., & Harnet, S. (Eds.). (1989). *The evaluation of scientific research*. Chichester, U.K.: John Wiley & Sons.

Feller, I., & Anderson, E. (1994). A benefit-cost approach to the evaluation of state technology development programs. *Economic Development Quarterly*, *8*, 127–140.

Irvine, J., & Martin, B. (1984). *Foresight in science: Picking the winners*. London: Pinter.

Jarmin, R. (1998). Evaluating the impact of manufacturing extension on productivity growth. *Journal of Policy Analysis and Management*, forthcoming.

Georghiou, L. (1995). Research evaluation in European national science and technology systems. *Research Evaluation*, *5*(1), 3–10.

Kingsley, G., Bozeman, B., & Coker, K. (1996). Technology transfer and absorption: An 'R&D value-mapping' approach. *Research Policy*, *26*(6), 967–995.

Kruytbosch, C. (1989). The role and effectiveness of peer review. In D. Evered, & S. Harnet (Eds.), *The evaluation of scientific research*. Chichester, U.K.: John Wiley & Sons.

Melkers, J., & S. Cozzens, S. (1996). Performance measurement in state-science and technology programs. Paper presented at the Annual Meeting of the American Evaluation Association, Atlanta, Georgia.

Meyer-Krahmer, F. (1988). Evaluation of industrial innovation policy: concepts, methods and lessons. In J. Roessner (Ed.), *Government innovation policy: Design, implementation, evaluation* (pp. 121–133). New York: St. Martin's Press.

Nexus Associates (1996). *Evaluation of the New York Industrial Technology Service*. Report prepared for the New York State and Technology Foundation. Belmont, MA.

Piric, A., & Reeve, N. (1997). Evaluation of public investment in R&D – Towards a contingency analysis. In *Proceedings—Policy Evaluation in Innovation and Technology: Towards Best Practices*. Paris: Organisation for Economic Cooperation and Development.

Porter, A.L., & Detampel. M.J. (1995), Technology opportunities analysis. *Technological Forecasting and Social Change*, *49*, 237–255.

Riall, W. (1991). *Local economic impact: Costs and benefits of development*. Atlanta, Georgia: Georgia Tech Research Corporation.

Rip, A. (1988) Mapping of science. In A. Van Raan (Ed.), *Handbook of quantitative studies of science and technology*, New York: North Holland.

Roessner, J. D., Lee, Y., Shapira, P., & Bozeman, B. (1996). *Evaluation of Iowa State University's Center for Advanced Technology Development*. Atlanta, Georgia: Georgia Tech Research Corporation.

Shapira, P., & Youtie, J. (1998). Evaluating industrial modernization: Methods, results and insights from the Georgia Manufacturing Extension Alliance. *Journal of Technology Transfer*, *23*(1), 17–27.

Shrum, W., & Mullins, N. (1988). Network analysis in the study of science and technology. In A. Van Raan (Ed.), *Handbook of quantitative studies of science and technology*. New York: North Holland.

Southern Technology Council (1997). Benchmarking university-industry technology transfer in the South: 1995–1996 data. Research Triangle Park, NC: Southern Growth Policies Board.

Thompson, C. (1998). Local politics, national policy, and the taxpayer-payback of manufacturing extension. *Journal of Technology Transfer*, *23*(1), 55–64.

Tijssen, R., & Korevaar, J. (1997). Unravelling the cognitive and inter-organisational structure of public/private R&D networks: A case study of catalysis research in the Netherlands. *Research Policy*, *25*(8), 1277–1293.

Wholey, J. (1994). Assessing the feasibility and likely usefulness of evaluation. In H. Wholey, P. Hatry, & K. Newcomer, *The Handbook of Practical Program Evaluation*. San Francisco, CA: Jossey-Bass.

Wilkins, T. (1998). Benchmarking manufacturing extension centers. In P. Shapira, & J. Youtie (Eds.), *Manufacturing modernization: Implications of evaluation results for program improvement and policy development*. *Proceedings of the fourth workshop on the Evaluation of Industrial Modernization Programs*. Atlanta, GA: School of Public Policy and Economic Development Institute, Georgia Institute of Technology.

Yin, R. (1989). *Case study research*. Newbury Park, CA: Sage Publications.

Youtie, J. (1997). Toward a cross-case analysis of outcomes of exemplary engagements by MEP centers. In P. Shapira, & J. Youtie (Eds.), *Manufacturing modernization: Learning from evaluation practices and results. Proceedings of the third workshop on the evaluation of industrial modernization programs*. Atlanta, Georgia: School of Public Policy, Georgia Institute of Technology.