Articles should deal with topics applicable to the broad field of program evaluation. Articles may focus on evaluation methods, theory, practice, or findings. In all cases, implications for practicing evaluators should be clearly identified. Examples of contributions include, but are not limited to, reviews of new developments in evaluation, descriptions of a current evaluation study, critical reviews of some area of evaluation practice, and presentations of important new techniques. Manuscripts should follow APA format for references and style. Length per se is not a criterion in evaluating submissions.

The Evaluation of Large Research Initiatives

A Participatory Integrative Mixed-Methods Approach

William M. Trochim
Cornell University
Stephen E. Marcus
National Cancer Institute
Louise C. Mâsse
University of British Columbia
Richard P. Moser
Patrick C. Weld
National Cancer Institute

Abstract: Over the past few decades there has been a rise in the number of federally funded large scientific research initiatives, with increased calls to evaluate their processes and outcomes. This article describes efforts to evaluate such initiatives in one agency within the U.S. federal government. The authors introduce the Evaluation of Large Initiatives (ELI) project, a preliminary effort to explore how to accomplish such evaluation. They describe a pilot effort of this project to evaluate the Transdisciplinary Tobacco Use Research Center (TTURC) initiative of the National Cancer Institute. They present a summary of this pilot evaluation including the methods used (concept mapping, logic modeling, a detailed researcher survey, content analysis and systematic peer-evaluation of progress reports, bibliometric analysis and peer evaluation of publications and citations, and financial expenditures analysis) and a brief overview of results. Finally, they discuss several important lessons and recommendations that emerged from this work.

Keywords: *center grants; concept mapping; evaluating research; federal evaluation; logic models*

William M. Trochim, Director of Evaluation for Extension & Outreach, Cornell University, 249 MVR Hall, Ithaca, NY, 14853; e-mail: wmt1@cornell.edu.

Authors' Note: The views presented in this article are those of the authors and do not necessarily reflect those of the U.S. Department of Health and Human Services or any of the authors' employers.

American Journal of Evaluation, Vol. 29 No. 1, March 2008 08-28
DOI: 10.1177/1098214007309280

The second half of the 20th century witnessed a profound transformation in the organization and management of science (Nye, 1996), beginning in World War II with the Manhattan Project (Lane, 1995; Rasmussen, 2002), and culminating in a crescendo of large scientific enterprises perhaps most aptly epitomized by the Human Genome Project (Nass & Stillman, 2003). The days of the individual scientist working in isolation are rapidly diminishing, increasingly replaced by the collaborative scientific team, the research institution, and the big scientific enterprise. Perhaps nothing better represents this shift to big science than the contemporary research center grant that was introduced in the 1960s and has been increasingly used as a research funding mechanism across many disciplines and fields (Institute of Medicine [IOM], 2004). Although such scientific research initiatives come in numerous forms and varieties—cooperative centers, public-private collaborations, research coalitions, transdisciplinary centers, clinical research networks, multisite projects, science and technology centers—they share some common characteristics. They are large (at least relative to traditional grants to individual scientists), involve collaborative teams or networks of scientists, and are often undertaken to address complex cutting-edge problems in their fields that are not as amenable to individual research grants.

A recent IOM report (2004) documents the rise in such initiatives at the National Institutes of Health (NIH) and the existence of nearly 300 centers in a variety of programs at the National Science Foundation (NSF), and similar trends are reported in the European Union (EU) in connection with the evaluation of Science, Technology, and Innovation (STI) policies (Molas-Gallart & Davies, 2006). These trends suggest that science is getting bigger, in amount of funding per initiative and in numbers of scientists collaborating, and that big science accounts for an increasingly large proportion of total research expenditures, posing new challenges for the management of science and particularly for how to evaluate the processes and effects of these considerable investments in research.

At the same time, the pressure for more evaluation and greater accountability across all programs in the federal government is also increasing (Brainard, 2002a, 2002b; U.S. General Accounting Office, 2000). For example, the Office of Management and Budget (OMB; 1993) instituted the Program Assessment Rating Tool (PART) as part of the 1993 Government Performance and Results Act (GPRA) and now requires every federally funded program to be reviewed on a regular basis, including an assessment of the quality of their evaluation and of the program's functioning and effectiveness.

We know relatively little about how to evaluate the progress and effectiveness of large research initiatives (Nass & Stillman, 2003). Traditional program evaluation approaches are seldom directly applicable for assessing big science, and the development of new methods has been slow in coming. Historically, science has been evaluated by assessing the scientific quality of the work, largely through peer review of research proposals and publications (Godlee & Jefferson, 1999, Kostoff, 1994b, 1995). With the emergence of big science, however, there is a need to assess a broader range of outcomes, including the social impact of research (Smith, 2001). A recent IOM report (Nass & Stillman, 2003) emphasizes that "a set of metrics for assessing the technical and scientific output (such as data and research tools) of large-scale projects" should be developed, "an evaluation of whether the field has benefited from such a project," should be conducted, and "the assessment should pay particular attention to a project's management and organizational structure, including how scientific and program managers and staff were selected, trained, and retained and how well they performed" (p. 196). Leading scientific groups have begun to address the challenges of evaluating large research initiatives (National Academy of Sciences et al., 1996; National Research Council, 1999). Nevertheless, we are still at a very early stage in the development of evaluation methodology and experience. The IOM report (2004) concludes that the NIH "does not have formal regular procedures or criteria for evaluating center programs" (p. 121) despite the considerable financial commitments involved.

The NIH is moving to address the need for evaluation of research center initiatives. In a recent funding request to Congress (National Institutes of Health, 2006) they emphasized accountability for research outcomes and included a specific requirement to report biennially on the performance and research outcomes of each center of excellence. The National Cancer Institute (NCI), one of the largest institutes, created the Evaluation of Large Initiatives (ELI) project to explore how we might improve the capacity for, and quality of, evaluations of large scientific research initiatives. ELI involved looking at what we currently know about evaluating large research initiatives, examining potential evaluation approaches and methodologies, and assessing the challenges and issues that need to be addressed.

This article reports on part of the work of the ELI project, an extensive multi-year pilot study to explore and gain experience with methods for evaluating the Transdisciplinary Tobacco Use Research Centers (TTURC) initiative of the NCI.[1] We present a description of the methods that were implemented and a high-level summary of the findings. In addition, we describe what we learned from this effort and the implications for future large initiative evaluations.

Pilot Evaluation of the TTURC Initiative

The TTURC Initiative

The TTURC initiative is a 5-year $70 million project funded by the NCI, the National Institute on Drug Abuse (NIDA), and the Robert Wood Johnson Foundation (RWJF) (Stokols et al., 2003). This pilot evaluation encompassed information from the first 3 years of its funding. The initiative provides support to multiple research centers to study new ways of combating tobacco use and nicotine addiction, and to help translate the results and implications of this work for policy makers, practitioners, and the public. Each center's research portfolio covers basic and applied research as well as research on policy-relevant issues in studies being conducted at the center. One of the primary goals of the initiative is to encourage and support transdisciplinary (Rosenfield, 1992) research (i.e., research that crosses and integrates theories and methods from different disciplines). Research supported and generated by the initiative is intended to set a new direction in how tobacco-related research should be conducted. Researcher training is a major component of the initiative and includes new and established investigators with the hope of broadening their scope of expertise within tobacco and across disciplines. Specific funds are provided to the centers to help facilitate the translation of basic and applied research into policy and practice. Given these unique characteristics and the information needs of multiple stakeholder groups, the pilot evaluation system was designed to gain experience with potential evaluation methods and tools and provide an assessment of TTURC processes and implementation and a preliminary exploration of short-term and intermediate-term outcomes.

The approach taken in this pilot evaluation is aptly described as mixed-methods (Greene & Caracelli, 1997; Greene, Caracelli, & Graham, 1989) because multiple qualitative and quantitative measures and analyses were incorporated into the design. Many of the individual measures were themselves combinations of qualitative judgmental data and quantitative indicators. For example, we used peer review approaches on several key measures, incorporating the judgments of multiple peer evaluators using their written assessments and their ratings of outcomes on quantitative scales. The approach is participatory in that it sought input from a variety of stakeholders including all of the researchers, center staff, and a number of independent peer reviewers on everything from the conceptual framework to outcome assessments.

Development of Conceptual Framework

There was little program theory available to guide an evaluation of this type. Consequently, the first major evaluation activity was the development of a conceptual framework for data collection and analysis. This framework was developed collaboratively, with active participation by TTURC investigators, funders, and other stakeholders. Concept mapping (Kane & Trochim, 2006; Trochim & Linton, 1986) was used to construct a comprehensive map of the outcome domains that needed to be addressed in the evaluation. The map that resulted was then developed into an outcome logic model (W. K. Kellogg Foundation, 2001) that depicts the hypothesized sequential causal relationships among outcome constructs. The map and outcome logic model were used to guide development of the measurement approaches and the analyses.

To accomplish the concept mapping, TTURC investigators and staff, scientific consultants, and representatives from funding agencies (total $N = 113$) brainstormed 262 potential outcomes that were edited and condensed into 97 final outcome statements. Participants sorted the statements for similarity (Coxon, 1999; Rosenberg & Kim, 1975; Weller & Romney, 1988) and rated them for relative importance. The sort data were analyzed with multidimensional scaling (Davison, 1983; Kruskal & Wish, 1978) and agglomerative hierarchical cluster analysis (Anderberg, 1973; Everitt, 1980), and average ratings were computed for each statement and cluster of statements. These analyses yielded 13 clusters of the 97 outcome statements and five general regions (Collaboration, Scientific Integration, Professional Validation, Communication, and Health Impacts), essentially clusters of clusters that illuminate a higher level of generality. An outcome logic model was developed by arranging the clusters of the concept map in the expected temporal order.

Each shape in the logic model corresponds to a component obtained from the concept mapping analysis (concept map not shown) and represents an outcome domain that encompasses subsets of the 97 multiple relevant specific outcome "markers" brainstormed by the participants.[2] The logic model describes a presumed program theory for the research initiative and conveys the sequence of expected outcomes from immediate to long-term markers.

The model generally flows from the most immediate short-term markers on the left through intermediate markers in the middle to long-term, distal outcome markers on the right. The outcome categories that emerged are consistent with many of those identified in the literature (IOM, 2004; McCullough, 1992) with specific content tailored to the TTURC initiative. Beginning on the left are the short-term immediate basic activities of the centers—Training, Collaboration, and Transdisciplinary Integration—that represent core activities of the TTURC initiative and the earliest, most immediate outcome markers that might be expected.[3] These basic activities are presumed to lead to the development of new and improved Methods and Science and Models. The consequent improved interventions are tested and lead to Publications. The dashed lines suggest that there also will be publications that result from and describe the intermediate products of improved Methods and Science and Models. Publications lead to Recognition and Transdisciplinary Research Institutionalization, which feed back on the overall infrastructure and capacity of the centers resulting in increased support for Training, Collaboration, and Transdisciplinary Integration. Publications also provide the content base for Communication of scientific results to the broader community. Recognition, through the public relations it engenders, provides a secondary impetus for

Communication. Policy Implications result primarily from Communications and Publications whereas Translation to Practice is primarily influenced by Improved Interventions. However, there is a dynamic relationship between Translation to Practice and Policy Implications suggested by the bidirectional arrow between them. Health Outcomes are influenced by the treatments and health practices that have been developed and by the policy changes enacted. In turn, positive or negative health outcomes feed back into new policies and practice. Taken together, the logic model (and the concept map on which it was built) provided an empirically and collaboratively derived conceptual framework for development of the TTURC evaluation measurement system and guided the analysis and aggregation of evaluation results. In addition, because the key constructs are stated without specific reference to tobacco control research, the logic model may be relevant to a broader array of transdisciplinary research initiatives than the TTURCs alone.

Evaluation Approach and Questions

The outcome logic model provided a framework for development of the key questions that guided the evaluation. Each question in turn has subquestions of greater specificity. For example, one question that addressed short-term markers was, “How do researchers assess performance of their centers on collaboration, transdisciplinary research, training, institutional support and center management?” An example of a question for the intermediate markers was, “Does TTURC research result in scientific publications that are recognized as high quality?” And an example of a question for a long-term marker is, “Are models and methods translated into improved interventions?”

Data Sources

The data sources and measures in this pilot evaluation are based on the conceptual framework and are consistent with recommendations made in the literature (IOM, 2004; Molas-Gallart & Davies, 2006), although they constitute only a subset of the potential measures that might be used in an evaluation of large research initiatives. For instance, the current study did not explicitly examine the effectiveness of the center grant mechanism in contrast to other possible ones and did not look at the effects of participation in the center on career paths of researchers. Nevertheless, the measures we did incorporate were focal to this initiative and will be recognizable to those familiar with recommendations of the literature.

This evaluation relied wherever possible on preexisting data sources rather than creating new measures. In particular we drew heavily from standard reports that all federally funded grantees are required to submit on an annual basis to the NIH within 90 calendar days of the last day of the final budget period of the project: the NIH PHS 2590 Progress Report form and the NIH SF269a Financial Report. Data from these sources were further processed using a variety of approaches to generate the data for the evaluation (see Analyses section).

The only new significant data source specific to this evaluation framework is the Researcher Form, an annual survey of TTURC investigators and research staff. A brief overview of the data sources follows.

Progress Report (PHS 2590). The annual Public Health Service PHS 2590 Grant Progress Report is required of all non-competing research grants funded through the Public Health Service, including all such research funded by the NIH and is intended to describe the progress made to date and plans for the following year.[4] For this evaluation, the Progress Report Summary and the Budget Justification provided data that were incorporated in the evaluation framework. The Progress Report Summary requires a brief presentation of accomplishments for the reporting period (usually 1 year) structured into six sections: (a) Specific Aims, (b) Studies and Results, (c) Significance, (d) Plans, (e) Publications, and (f) Project-Generated Resources. The Budget Justification includes two sections: (a) a detailed budget justification for those line items and amounts that represent a significant change from that previously recommended and (b) an explanation of any estimated unobligated balance (including prior year carryover) that is greater than 25% of the current year's total budget.

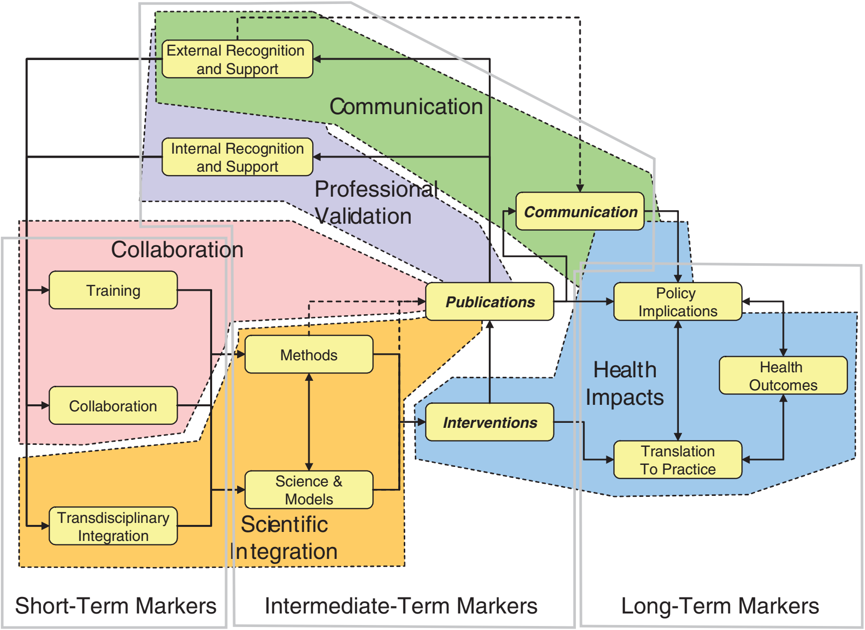
Financial Report (SF269a). The SF269A Financial Report indicates the exact balance of unobligated funds.[5] The key data provided through this form are the total dollar amount authorized for spending (Federal Funds Authorized) and the amount actually spent during the year. The unobligated balance, or carryover, is the total amount authorized minus the amount spent. The SF296A only reports total dollars. Amounts spent and carried over by budget category are not provided.

Researcher Form. The Researcher Form is a self-administered questionnaire designed explicitly for this pilot by the ELI team to elicit the opinions and evaluative assessments of the TTURC researchers regarding the entire range of outcomes in the logic model, including collaboration, transdisciplinary integration, science, models and methods, internal and external support and recognition, communications, and the effects of TTURC research on policy, practice and health outcomes. It consists of 25 closed-ended questions (each with multiple subitems) and three open-ended questions. TTURC funders, consultants, and researchers generated several hundred potential items for this form. These were classified into the outcome categories in the TTURC logic model (Figure 1) and grouped into multi-item questions in the Researcher Form. The instrument went through multiple cycles of structured review and revision with a variety of groups including the TTURC evaluation methodology team, the funders, the TTURC consulting committee, and the TTURC principal investigators to assess the scale's content and face validity. Indices or scales representing each of the dimensions of the logic model were created. Psychometric methods were used to assess the measurement properties of four scales that were developed. Three of the scales related to collaboration: (a) satisfaction with collaboration, (b) impact of collaboration, (c) trust and respect in a collaborative setting, and a fourth scale assessed attitudes about transdisciplinary research. The Researcher Form was designed to be completed on an annual basis by all members of the research team including researchers and research support staff.

Analyses

Researcher Form. A response rate of 92% (216 of 234) was achieved on the Researcher Form while still ensuring confidentiality of responses. How we achieved a response rate of more than 90% in an anonymous survey of extremely busy scientists is worth noting. We asked each center to provide the exact distribution list of all qualified respondents in their center and supplied them with precisely that number of surveys (identified by center, not individual), to be completed by respondents and mailed by them directly to the evaluation office. Each week we announced the cumulative response rates for each center to all center directors so they could see how they and others were doing. Although no one knew who specifically responded because of the anonymity at the level of the individual, it was apparent that a fair amount of friendly competition among the centers led them to encourage their staffs to return the surveys.

Figure 1
Outcome Logic Model for the Transdisciplinary Tobacco Use Research Center (TTURC) Evaluation



Three collaboration scales and one measuring transdisciplinarity were created. All four scales had adequate psychometric properties. The Cronbach's alpha for these scales varied from .75 to .91 and the a priori factor structure of these scales was validated using confirmatory factor analysis. In addition, 26 separate index variables were constructed from different combinations of question items, with each scale and index score linked to an outcome area on the logic model. Finally, basic descriptive statistics and key group differences (e.g., respondent role and center) were computed and tested using the scales and indexes as outcomes.

Content analysis of progress report summaries. A content analysis (Krippendorf, 2004; Weber, 1990) was done on each of the Progress Report Summaries (focusing on the Studies and Results and Significance sections) for each year to determine which of the 14 markers were reported ($N = 269$ reports). In this coding scheme, it did not matter how much a report emphasized any specific marker, only whether it did. Although not as specific as a coding of the degree to which a report addresses each marker, this dichotomous coding is highly reliable, can be accomplished quickly and at relatively low cost, and is capable of demonstrating the general pattern of outcomes across the subprojects. Three methodological substudies were conducted to assess intercoder reliability. In each study, four coders (NCI staff) were provided with the coding instructions and a recording sheet for a sample of six subproject reports. The final intercoder reliability estimate for the content analysis was .94 (kappa = .938, t-value = 8.601, $p < .001$).

Peer evaluation. Eight peer reviewers (Kostoff, 1994a) external to the initiative were used to evaluate the 272 Progress Report Summaries for the subprojects across the seven funded centers. These reviewers were recruited from the same pool of people that composed the original TTURC proposal review team and made up the current TTURC consulting committee. The Peer Evaluation Form assessed several areas: (a) The overall progress of the subproject, (b) progress in each of the outcome areas on the logic model, and (c) the impact of the research to date on four important audiences or constituencies (scientists and the research community, practitioners and clinical practice, policy makers and policies, and clients and consumers of health services). Finally, the form allowed peer evaluators to provide any additional comments or to expound on any of their ratings. The form was designed to be brief so as to not impose undue burden on the peer evaluators. Each of the subproject reports ($N = 272$) was coded by two randomly assigned peer evaluators. More than 80% of the time both evaluators either agreed or differed by no more than one scale unit (e.g., one judge rated a *2* while the other rated a *3*) on the 1-to-5 scale used. Finally, two nonparametric estimates of agreement were computed, Kendall's tau b, and Spearman's rho. For all variables, both measures were positive and statistically significant (for 16 of 18 variables, $p < .01$, and for the remaining, $p < .02$).

Bibliometrics. Bibliometric analysis involves the quantitative assessment of scientific publications, the works they cite, and the citations of them (Osareh, 1996a, 1996b). Citations are made in published scientific work to acknowledge the prior relevant work of other scientists; and, consequently, the numbers and sources of citations can provide important data about the recognition of published work by other scientists (Garfield, 1995). Bibliometric analysis is a critically important source of objective information about the quality and productivity of scientific work (Kostoff, 1995). It can be used to estimate the influence and impact of a single publication, or the quality and recognition of the entire published opus of a researcher, a research journal, or even a field of research. Although there are legitimate and varied criticisms of the use of bibliometric data in evaluating scientific research (Funkhouser, 1996; Seglen, 1997; Skoie, 1999), the rigor and quality of bibliometrics has improved considerably over the time (Hood & Wilson, 2001; Schoepflin & Glanzel, 2001), and this type of analysis is recommended by key scientific advisory groups (IOM, 2004) and administration officials (Brainard, 2002b) as an important potential component of large initiative evaluation.

In this bibliometric analysis a number of index variables were constructed from publication and citation data. Several of these indexes are based on data that enables the centers' results to be compared to external productivity standards (e.g., citation rates of all other articles in the same journal as each publication and citation rates of all articles in the same field or discipline). The indexes used in the analysis include number of citations (total, self, adjusted), number of expected citations, journal impact factor (Garfield 1994a, 1994b), journal and field performance indicators, 5-year journal and field impact factors, statistics on cited and citing journals, and a journal disciplinarity index designed to reflect the multidisciplinarity of cited or citing journals.

Financial analysis. The financial analysis integrated data from two separate sources: (a) The annual budget that is completed as part of the annual Grant Progress Report (PHS 2590) and (b) the annual project expenditures as reflected in the Financial Status Report (FSR 269A). The FSR data described actual spending and was collected for each of the funded centers on an annual basis, within 3 months of the completion of each project year. Analysis of this report enabled assessment of spending patterns, whether each center was utilizing all of its allocated funds, and whether there was significant carry-over to the next year. The budget justification data provided a summary of the reasons (as delineated by each principal investigator) for any budget carryover from 1 year to the next. These data were collected approximately 2 months before the completion of a project year and were part of the description of the plans for the subsequent year.

Integrated Evaluation Plan and Analyses

The diverse measures and data sources were integrated into an overall evaluation framework that is depicted in Table 1. There are three primary sources of data: the PHS2590 Progress Report, the Researcher Form, and the SF259a Financial Report. In addition, the Progress Report was further divided into three parts—the Progress Report summary, list of publications, and the budget and justification—that were handled separately in the analyses. Content analysis and peer evaluation were used to assess the Project Report summary narrative. Bibliometrics methods were applied to the list of publications supplied with the Project Report. The financial analysis incorporated data from the Progress and Financial Reports. Standard survey analysis was done on the Researcher Form.

The logic model is the key unifying device for organizing and grouping results from multiple methods for each outcome area and enabling synthesis of the findings. Results were classified according the model into the three broad temporal stages of short-term markers, intermediate markers, and long-term markers. Within each stage, results were examined across the multiple data sources. For example, a key intermediate marker is scientific productivity. Productivity results are available from the bibliometric analysis of publication quantity, quality, and citations; from the assessments of productivity in the Researcher Form; and based on the judgments of peer evaluators. Finally, a type of “pattern-matching” analysis (Trochim, 1985) was used to assess overall progress in the TTURC initiative. The TTURC logic model suggests a sequence of outcomes of the initiative, beginning with the short-term markers and, over time, affecting long-term markers. Over successive years we would expect to observe a pattern of change first in the short-term indicators and subsequently in outcomes from left to right on the logic model. Where we had outcome assessments across all clusters and across multiple years, we were able to overlay the estimates onto the logic model visually to examine whether this expected left-to-right pattern appears to be supported.

Summary of Results

Brief summaries of the results are presented in separate sections grouped into short-term markers, intermediate markers, and long-term markers.[6] We provide this brief summary of results to give the reader a sense of the types of results that were obtained. Within each section, the basic findings are summarized across all data sources and analyses.

Short-Term Markers

The short-term markers emphasize assessment of TTURC activities and immediate outcomes. In addition to the three areas on the logic model of training, collaboration, and transdisciplinary integration, the short-term markers also addressed management-related measures such as financial analysis of expenditures and carryover.

Table 1
Summary of Constructs Extracted From Each Data Source by Analytical Methods

		Conceptual Map Constructs															
		Collaboration			Scientific Integration		Health Impact				Communication						Management*
Data Sources	Methods	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Progress report (summary and budget)	Peer evaluation	X	X	X	X	X	X	X	X	X	X	X	X	X			X
	Content analysis	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	Bibliometric analysis													X			
Financial report	Descriptive statistics														X		
Researcher form	Descriptive statistics	X	X	X	X	X	X	X	X	X	X	X	X	X			
1. Training		7. Translation to practice		12. Communication													
2. Collaboration		8. Policy implications		13. Publications													
3. Transdisciplinary integration		9. Health Outcomes		14. Significant staff changes													
4. Methods		10. Recognition		15. Budgetary expenditures													
5. Science and models		11. Transdisciplinary integration		16. Progress “project on track”													
6. Improved interventions																	

Training. The training of students, new researchers, and staff was one of the highest rated outcome areas according to TTURC researchers. On average, they assessed training good to excellent. Nearly one third of all subprojects reported progress in training outcomes over time.

Collaboration. The results showed that researchers are collaborating across disciplines and value collaboration and transdisciplinarity. Collaboration received the second highest progress rating of the 13 areas rated independently by peer evaluators. In the 3rd year, nearly 50% of all subprojects were coded as reporting progress in collaboration. There are some significant process barriers to collaboration identified, including the difficulties of resolving conflicts, conducting productive meetings, and dealing with the increased time burden required. In addition, there are significant differences in collaboration results by role. Professional research support staff (e.g., [bio]statistician, research associate, research assistant, laboratory analyst, data manager) report relatively more difficulty than researchers in dealing with issues of communication and collaboration. Communication within centers appears hampered by insufficient time and by information overload. With respect to within-center collaboration, evaluations were highest for acceptance of new ideas and the ability to capitalize on the strengths of different researchers. However, the lowest evaluations were given for resolution of conflicts among collaborators and productivity of meetings. In terms of attitudes about collaboration, respondents express strong respect for their collaborators but indicate that collaboration poses significant time burdens in their research. Taken together, these results suggest that though respondents are positive about their collaboration experiences, there are significant barriers to how effectively collaboration is accomplished in practice.

Transdisciplinary integration. The ability to conduct transdisciplinary research was the highest rated performance marker across the centers after publication quality. It was in the top four variables (of 13) in terms of progress ratings by peer evaluators. Researcher attitudes about transdisciplinary research were uniformly high and positive, not surprising given that this was a primary purpose of the initiative.

Financial management. There was significant variability across centers in their ability to spend allocated funds as originally proposed. Several problems were identified including significant difficulties starting up in a timely manner, delays in funding allocations from NIH, significant budget carry-overs from year to year, and significant changes in project personnel. Inability to spend as planned raises questions about whether the centers can achieve their proposed aims in the 5-year framework of the initiative and suggests that expected progress may be slower than expected.

Intermediate Markers

Intermediate markers include the logic model categories Methods, Science and Models, Recognition, Publications, Communications, and Improved Interventions. In terms of peer evaluation, Methods had the highest rated progress whereas Science and Models was third highest. Progress in Methods was reported by peer reviewers for nearly three fourths of all subprojects by Year 3. In addition, nearly one half of the subprojects were rated as showing progress by peer reviewers in Science and Models by Year 3. In the researcher survey results, limited progress is reported overall by the researchers themselves in the development of Science and Models and Methods, although this may be expected at this point in the evolution of the TTURC initiative. On the methods side, "good" progress was reported by researchers with respect to the development of measures. In terms of scientific theory development, "good" progress was reported in "understanding the relationships between biological, psychosocial, and environmental factors in smoking."

As expected, the number of all Publications and of research publications increased each year. Bibliometric analyses indicated that TTURC publications are placed in well-cited journals, and TTURC publication citation rates are statistically significantly higher than for typical articles in the same journals. All statistically significant results reported here were significant at $p < .05$ on one-tailed tests of significance.

In addition, TTURC citation rates are significantly higher than for all articles published in all journals in the same fields/disciplines. The rate at which observed citations exceeds expected rates increased significantly over the first 2 complete years of the initiative. We did not anticipate that analysis of publications and, especially, of citations would show much in the relatively short time frame of the first 3 years of an initiative. The fact that they did is indicative of the productivity of the researchers and of the potential sensitivity of bibliometric approaches even for short-term evaluation.

Communications of research findings was rated on the researcher survey on average as "good" by researchers. And moderately good progress is reported on the development of interventions.

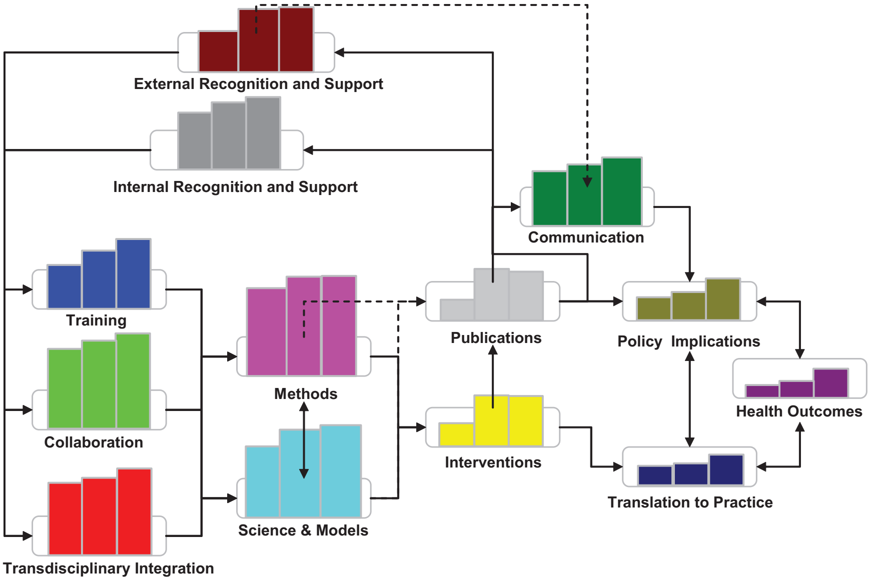
Long-Term Markers

Long-term markers include the effects of the TTURC initiative on policy and practice and, ultimately, on health outcomes. By its 3rd year, the TTURC initiative was only expected to have a moderate impact on scientists and the research community and limited impact on practitioners and clinical practice, policy makers and policies, and clients and consumers of health services. If the initiative is successful, these impact estimates should increase as more work is accomplished. However, even at this early stage in the initiative, TTURC researchers report considerable impact on policies at the state and local levels and on practice with respect to tobacco control. TTURC researchers report considerable optimism that their research will help lead to significant positive health outcomes, especially for consumption and prevalence. Of course, it is important to bear in mind that these are subjective ratings obtained from surveying researchers. As such, they reflect researcher attitudes but do not constitute evidence of long-term impact, an issue that we consider further in the discussion.

Pattern-Matching Results

The TTURC logic model hypothesizes a sequence of outcomes of the initiative, beginning with the short-term markers and, over time, reaching the long-term markers. This pattern makes it possible to examine the degree to which the observed results correspond with the hypothesis suggested in the logic model, a type of pattern-matching design (Trochim, 1985). These results were graphed onto the logic model for the three major data sources (the Researcher Form, Content Analysis, and Peer Evaluation). To illustrate, results for the Peer Evaluation assessment of overall progress on each outcome are shown in Figure 2. In general, as expected, short-term markers (i.e., process measures) show the greatest progress over time with intermediate and longer term markers showing lower progress levels. Formal statistical tests of such pattern-matching hypotheses have not yet been developed; however, the pattern of observed TTURC markers for all three data sources corresponds well visually with the TTURC logic model prediction. The trends over time suggest that the TTURC initiative in general is making progress along the lines that would be expected given the logic model.

Figure 2
PEER Evaluation Results: Average Progress by Year Overlaid Onto Transdisciplinary Tobacco Use Research Center (TTURC) Logic Model



Discussion

This pilot investigation was intended to explore and develop evaluation methods and models and to create products and templates that could be adapted in later evaluation efforts in similar scientific entities. However, as a pilot project it may perhaps be most valuable in identifying important lessons. This discussion concentrates on several general conclusions that emerged from this work and that may be useful in guiding future large research initiative evaluation efforts.

Develop a Comprehensive Conceptual Model

Our experience in this research convinced us that it is essential to develop a comprehensive conceptual model that can guide such an evaluation. A large research initiative is a complex endeavor with a broad range of scientific activities, potential outputs, and outcomes. A comprehensive evaluation typically needs to gather data of different types using a wide range of measurement approaches to address adequately the information needs of the varying constituencies. A conceptual model is essential for identifying the variables that need to be measured and for integrating the various qualitative and quantitative data that typically result. We used a combination of collaborative concept mapping and outcome logic modeling to provide the conceptual and analytic framework that would guide the evaluation. Although these are by no means the only way to approach the modeling task, some type of structured empirical process for creating a conceptual model and using it to guide the evaluation is desirable in work of this kind. The logic model that we developed was useful as an organizing rubric in framing evaluation questions and for synthesizing and reporting results.

Use Participatory and Collaborative Evaluation Approaches

Most large research initiatives have a wide variety of relevant stakeholder groups that are likely to hold differing and sometimes contradictory expectations about what the initiative might achieve. For instance, scientists are often concerned with conducting high-quality research and generating basic knowledge. Administrators may be interested in managing scientific resources effectively and efficiently. Legislators and the public often concentrate on the appropriate use of tax dollars and on results that can be applied to improve the well-being of individuals and society. Evaluation of large research initiatives needs to draw on the deep tradition of participatory and collaborative evaluation (Whitmore, 1998). Such approaches help to ensure that the evaluation addresses the interests of multiple groups and that the research scientists, staff of research centers, and the funders have sufficient buy-in to enable sustainability of the evaluation. Timely and continuous feedback of evaluation findings to stakeholders is essential to maintaining buy-in and commitment to using the information to improve the functioning of the initiative.

Incorporate Integrative Mixed Methods

The pilot evaluation we conducted involved a number of different measurement and analysis methods that were connected to the logic model and needed to be integrated in reporting. For instance, we gathered information about publication quality that included quantitative bibliometric data on citation rates and journal quality, and subjective ratings from center researchers and multiple peer evaluators. These results needed to be summarized and synthesized with the other data collected for other outcomes on the logic model. This could not have been accomplished without use of an integrative mixed-methods approach (Greene & Caracelli, 1997), and it is hard to imagine how any comparable evaluation of large research initiatives could be accomplished without integrating appropriate qualitative and quantitative information.

Integrate Evaluation With Existing Reporting Systems

Research center grants at the federal level typically have regular annual reporting mechanisms. These reports were not developed with evaluating center outcomes in mind, are generally perceived by grantees as a burdensome administrative requirement, are often completed in a perfunctory manner, and are typically reviewed only by a handful of program directors responsible for managing the initiative. One of the major dilemmas we faced in developing this evaluation was whether to create new reporting and measurement systems or improve significantly on the existing ones. We chose the latter because we could not justify creating new annual evaluation data collection burdens if existing annual progress data were not being used in the most effective way possible. The only new measure that we created was the Researcher Form that collected structured attitudinal and behavioral information from the researchers and research center staff. All other data were drawn from existing annual reporting mechanisms. The outcomes on the logic model were used to create more specific instructions to the researchers regarding the structure they needed to use and criteria they needed to address in their reporting. The credibility of using these criteria was enhanced by the fact that the researchers collaborated in generating them. The use of systematic content analysis and peer review of information from the annual reports, and detailed bibliometric analysis and verification of reported publications, sent a clear message to grantees that these reports would be more closely scrutinized and significantly improved the quality of the reported data. In addition, this approach had the value of integrating the evaluation into the existing research management and reporting system so that the funding agency did not need to construct new or additional reporting requirements or data collection, processing, and storage technologies.

Adapt the Evaluation to the Initiative's Stage of Development

A research initiative goes through a number of distinct phases of evolution in the course of its existence. It can take several years for a research center to become fully operational, to get the research and support staff into place, acquire and set up the necessary physical infrastructure, and begin multiple research projects. From an evaluation perspective one expects to see outcomes emerge over time.

An evaluation needs to be flexible enough to emphasize different types of information at different stages of initiative development. Not all of the potential evaluation questions identified can be addressed in the early stages of a center grant initiative. Early in the development of a center one might expect that more qualitative rapid-feedback processes and implementation evaluation would dominate. As the initiative and the measurement of outcomes associated with it become more stable, it becomes increasingly possible to look at change in outcomes over time, qualitatively and quantitatively. With more stable baselines, more formal comparisons and controls become feasible (such as our comparisons with normative publication standards in relevant substantive fields or comparisons of scientific quality with other research initiatives). Evaluation needs to be built into the entire life of the initiative, from design through initiative completion, and needs to be able to change and adapt in its focus as the initiative evolves. Doing so helps evaluation to be seen by all stakeholders not as a disciplinary after-thought but rather as an integral and essential part of the conduct of science.

Develop Standardized Cross-Initiative Evaluation Systems

Standardized systems that have common elements and approaches across multiple research initiatives need to be developed to increase efficiencies, lower overall costs of evaluation, and enable cross-initiative comparison on appropriate outcomes. There are signs that the federal government is moving in this direction, albeit slowly and with somewhat mixed results. One of the most ambitious efforts is that of the OMB (2007c) that developed the PART review for assessing all federal programs on a variety of dimension including program purpose and design; program management and strategic planning, performance measurement, and evaluation of program outcomes. Although outcome evaluation is not conducted within the PART review, the process does attempt to assess the quality and results of such evaluation for all federal programs, including large research initiatives. The results of the PART assessments are published on the ExpectMore.gov Web site (OMB, 2007a). For example, the ExpectMore.gov site reports on the assessment of the NSF's Federally Funded Research and Development Center initiative (OMB, 2007b). Although the OMB PART effort has not been without legitimate critics, it represents a major effort to encourage more standardized cross-initiative evaluation and is relevant to the evaluation of large scientific research endeavors (U.S. General Accounting Office, 2005).

The need to develop generalized evaluation systems is increasingly also recognized at the federal agency level. For instance, an IOM report (2004) concluded that,

> A program evaluation plan should be developed as part of the design and implementation of new center programs, and data on indicators used in the evaluation plan should be collected regularly and systematically. Data should be collected from the centers according to a common format. Most of the indicators should also be useful for program monitoring and progress reporting. (p. 122)

We can see such suggestions beginning to be incorporated into Requests for Applications for research grants. For instance, the recent Institutional Clinical and Translational Science Award from the NIH is intended to fund 60 research centers over a 5-year period. The instructions to research applicants call for center and cross-center evaluation, including the development of “a strong evaluation and tracking plan for all research education, training and career development activities” (U.S. Department of Health and Human Services [USDHHS], 2007, n. p.). The plan should “include the review of the effectiveness of all aspects of the program” and explicitly requires a logic model that can guide the evaluation (n.p.).

One way to systematize evaluation so that it is more routinely expected and funded would be to simply require it for all research-based initiatives, or at least those that exceed a certain size. This appears to be the approach taken since the inception of ELI by the NCI’s Executive Committee that implemented a policy that all scientific concepts involving financial commitments require that evaluation plans be approved. Such approaches encourage improved integration between science and management, including the financial management of large initiatives.

The central challenge remains determining what is needed to ensure that appropriate standardization occurs across large research initiatives, in terms of common outcomes and measures where possible, but certainly with respect to common evaluation policies that detail which initiatives should be evaluated, when evaluation should be done, how it should be financed, key components that are required (e.g., logic models), timing and integration in the large initiative life cycle, and reporting to the public. In the future, moving multiple initiatives into a similar common evaluation framework opens up new possibilities for cross-initiative comparison groups and measures. For example, it would be possible to compare across initiatives at the same point in time or to look at the evolution of initiatives that began at different time points. Of course, there are challenges to standardized evaluation systems and policies. A “one-size-fits-all” approach may fit all without doing a good job of evaluating any. And evaluation systems perceived by researchers as irrelevant, obtrusive, or inappropriate run the risk of alienating the researchers and ultimately jeopardizing the quality of the results. Nevertheless, judicious evolution of a common approach to evaluation for large initiatives appears to offer promise for making more coherent sense of the large federal investment in research.

Utilize Peer Review Approaches

Peer review is central to the idea of rigor and quality in scientific research (Kostoff, 1994a, 1995). In the culture of NIH, qualitative peer-review processes are often utilized for assessing the progress, impact, or the effectiveness of a scientific initiative, and to judge plans and outputs based on scientific relevance and merit. Our process formalized and structured peer review in the evaluation, yielding qualitative and quantitative assessments that could be assessed for reliability across multiple reviewers, and linked the criteria by which peers judged performance and outcomes directly to the logic model that the researchers developed.

Systematic research initiative evaluation using peer review is a relatively new endeavor. Peer review of research proposals and subsequent publications tends to dominate the landscape (Brainard, 2002b; Kostoff, 1994b). Attempts to create a structure outside this system, and especially methods for looking at the implementation of scientific research, are likely to continue to be challenging. One of the major challenges is in identifying a sufficient pool of prospective peer reviewers who are knowledgeable enough to make informed assessments of advanced scientific research and do not have conflicts of interest with funded centers being evaluated. Many of the most appropriate potential peer-review candidates in any given field are either funded by the initiative, have previously applied and been denied funding, or are prospective grant applicants. Despite these challenges, in our work incorporation of structured peer review of progress reports as a major component of the multimethod approach to evaluation helped ensure buy-in of the grantees, at least in part because of the strong normative bias that emphasizes peer review in science.

Address Issues of Causation and Control

One of the central methodological challenges in evaluation of large research initiatives is assessing the degree to which the initiative can be said to have a causal effect on outcomes. The primary challenge in assessing causation in this context is that it is difficult to separate the contributions of the initiative from all of the other potential factors that might affect outcomes (Cook & Campbell, 1979), especially for the longer term outcomes. In addition, it is unclear how much time is needed to significantly change the long-term markers. The research designs usually recommended for causal hypothesis testing, such as randomized experiments and quasi-experimental designs like the regression-discontinuity (Trochim, 1984) design, are typically not feasible in this context. The approach to causal inquiry taken here in addressing the overall effectiveness of the initiative was to use a pattern-matching variation of the Nonequivalent Dependent Variables Design (Cook & Campbell, 1979; Trochim, 1985) that compares patterns of outcomes across multiple constructs with the hypothesized or expected pattern of outcomes, in this case as reflected in the sequence of outcomes in the logic model. Nevertheless, we would do well to keep in mind the challenges involved in causal assessments of such complex initiatives (Molas-Gallart & Davies, 2006):

> The main problem here is that we are trying to measure the exact extent to which specific outcomes can be attributed to policy measures, which are likely to play a small role among the many other factors that will emerge in a systemic model. Such detailed attribution requires comprehensive modeling and measurements that are not currently available. (p. 78)

Another type of control that would be possible and potentially useful would involve estimation of changes in individual researcher performance. For instance, one might assess the quality and quantity of publications quasi-experimentally through pre- and postinitiative or time-series measurement and analysis.

Improve Funding and Organizational Capacity for Evaluation

There is not yet sufficient financial support and organizational capacity (Compton, Baizerman, & Stockdill, 2002) to integrate evaluation into the planning of large research initiatives, although this has been changing in the past few years. For example, the NIH Office of Evaluation provides some "set-aside" funds for evaluation activities. That office, as well as the NCI Office of Science Planning and Assessment and a newly formed Evaluation Committee within NCI's Division of Cancer Control and Population Sciences, provide consultation and assistance in the development of evaluation projects. However, these mechanisms tend to be initiative specific and typically require considerable evaluation proposal development before funding can be allocated, delaying the commencement of evaluation activities well beyond the start of the initiative, and posing barriers to more effective and timely initiative start-up.

Evaluation capacity building also needs to be addressed. The ELI team that led this pilot consisted of three NCI employees providing approximately 25% of their time, a half-time external evaluation expert, and a team manager. The process also involved time commitments from the NCI program director and initiative staff (particularly the center principle investigators and External Consulting Committee). Although this comprehensive team commitment was justifiable in a pilot project, implementation of this kind of effort on a more sustained basis would require institutional and structural commitments of resources that are not currently in place.

The funding for this evaluation came from multiple internal sources as the project evolved and needs arose. We estimate that the total cost of this pilot evaluation was between US$400,000 and $500,000; however, much of this can be attributed to start-up and one-time costs involved in development of new measures, methods, analyses, and data structures that could be reused or adapted in subsequent large initiative evaluations. Given the pilot nature of this effort and the likely costs of sustaining such endeavors, it seems reasonable that an allocation of at least 1% of initiative funding for such evaluation is a reasonable benchmark, especially if cross-initiative efficiencies could be achieved. Although NIH has a 1% set-aside fund for evaluation, it requires a separate and considerably demanding proposal and review, and there is currently no direct routine allocation of funds to an initiative evaluation that goes into place when it begins. Streamlining this process and ensuring automatic allocation of 1% of initiative funds would greatly enhance the capacity for and timeliness of evaluation.

Address Management Issues in Large Initiative Evaluation

Large research initiatives require a constructive interplay of scientific and management skills. It was apparent in our research that evaluation issues are inseparable from the management of the research initiative. As noted in a recent IOM report (Nass & Stillman, 2003) on large-scale biomedical science, it is the gaps in the management of large scientific endeavors that most often lead to problems in implementation and accomplishments from these efforts.

One of the most important management issues involves setting expectations from the inception of the call for proposals that evaluation will be required of all grantees so that evaluation is viewed as an integral part of the initiative and not as an add-on imposed on grantees after awards are made. Detailed expectations for cross-center and cross-initiative evaluation should be built into the initial call for proposals. Grant applicants should be required to describe their evaluation plans and how they will be integrated into larger cross-center evaluations, as done in the recent NIH Clinical and Translational Science Awards (U.S. Department of Health and Human Services, 2007) described earlier. Grantees should be expected to collaborate on evaluation from the inception of the award to develop a clearer sense not only of what their own research is trying to accomplish but also of how it addresses the outcomes that the initiative is collectively trying to achieve. When these things are done, applicants will be better able to anticipate how they will need to integrate their scientific work with the evaluation effort.

A critical management issue is determining the right balance between external and internal evaluation for large center-based science. How much responsibility for evaluation should be undertaken by the centers themselves who often have experienced researchers and evaluators and are closest to the phenomenon but have conflicting interests about how evaluation results might reflect negatively on their centers? How much of the evaluation should reside in the funding agency responsible for oversight, but also having potential conflicting interests and needs? In many instances, though centers or funding agencies may have talented researchers who could be drawn on for evaluation, it will likely be necessary to provide specialized training in evaluation to prepare them adequately for the unique challenges such work entails. To what extent should evaluation of such initiatives be conducted by external independent contractors who may have relatively less understanding of the purposes and functioning of the initiative but be in a better position to draw on a wider base of evaluation experiences? The pilot conducted here used a team that resided within the federal agency (NCI) with the support of the program director and center researchers. Although that may have been sufficient for a pilot study that required considerable coordination with the existing federal reporting system, it is probably not desirable as a routine mechanism for evaluation because of the vested interest that the funder and grantees have in how major investments in research dollars appear to perform.

An alternative arrangement to situating evaluation within the funder, in the centers themselves, or some combination of both would be the use of a third-party evaluator external to the funder and grantees who is given primary responsibility to manage and implement the evaluation. Third-party evaluators could reduce the burden for center scientists and the potential for conflicts of interests and can potentially bring a broader range of evaluation experience to the task and a broader systems-level perspective to the evaluation. It may be sensible for third-party evaluators to measure aspects of the larger environmental context in which the initiative operates. Such a systems-based approach to evaluation (National Cancer Institute, 2007; Williams & Imam, 2006) would be likely to identify incentives and barriers to success outside the control of the investigators and funding agencies. For example, university tenure policies can be examined to see if they appropriately take into account the potential impact of additional time demands of participating in collaborative, transdisciplinary research on investigator productivity. Similarly, the number of peer-reviewed journals that invite submissions of interdisciplinary manuscripts clearly has the potential to affect productivity and, therefore, should be factored into any evaluation of such initiatives.

Moreover, though issues will still remain regarding the ability of an external evaluator to understand the nuances of a complex research center initiative, these can probably be mitigated, at least to a great degree, through structured input from the scientists and funders throughout the evaluation. Perhaps the greatest challenge to using independent evaluators will be to the management of the initiative, within funded centers and in the funding agency. Both of these entities will be concerned with how the evaluation will make them look and the degree to which it could imperil future funding. Continuing pressure and increased independent funding from higher levels of the federal government will be critically important to ensuring the success of large research initiative evaluation.

The research described here was a preliminary attempt to address a complex problem of increasing importance in contemporary scientific research. It contributes to the growing literature on how such evaluations might be accomplished. The nature of how science is managed is changing, and there is every reason to expect that the trend toward larger and more collaborative research endeavors will continue for the foreseeable future. As greater investments are made in large research initiatives, the public and the Congress will increase their calls for evidence that the funding is being well managed, is contributing to research goals, and is affecting the problems of our society. A more concerted effort is needed to develop appropriate evaluation models and methods before we will be ready to answer such calls effectively.

Notes

1. This 3-year project resulted in the compilation of a set of measurement instruments, database structures, templates, and other outputs, and in the production of a number of detailed project reports. Please contact the first author for information on these resources and their availability.

2. The term *marker* is used throughout this article and should be considered a synonym for the term *outcome*. It proved to be a more accessible term to the biomedical research community in conveying, in particular, the idea of a short-term and intermediate-term outcome.

3. Throughout this article, the labels for clusters of outcomes identified through concept mapping and used in the logic model will be distinguished in the text with capitalization.

4. For more detailed information and the complete form and instructions, see http://grants1.nih.gov/grants/funding/2590/2590.htm. Although the form itself has seven pages, a completed form is likely to be considerably longer than that. For instance, for the Transdisciplinary Tobacco Use Research Center (TTURC) initiative, grantees are expected to generate a Progress Report Summary (PHS 2590, p. 5) that is two to four pages long for each subproject.

Because each of the seven centers has as many as 10 to 12 such reports, this section of the annual report alone can be considerable.

5. The form and instructions for the SF 269a can be found at http://www.whitehouse.gov/omb/grants/grants_forms.html.

6. The results presented here are necessarily brief summaries of the full pilot evaluation results that may be obtained from the first author on request.

References

Anderberg, M. R. (1973). *Cluster analysis for applications*. New York: Academic Press.

Brainard, J. (2002a, March 29). New science measures released by OMB. *Chronicle of Higher Education, 48,* p. A25.

Brainard, J. (2002b, June 7). New NIH director asks researchers' help to provide accountability. *Chronicle of Higher Education, 48*. Available at http://chronicle.com/daily/2002/06/2002060702n.htm.

Compton, D. W., Baizerman, M., & Stockdill, S. H. (Eds.). (2002). *The art, craft, and science of evaluation capacity building* (Vol. 93). San Francisco: Jossey-Bass.

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

Coxon, A. P. M. (1999). *Sorting data: Collection and analysis*. Thousand Oaks, CA: Sage.

Davison, M. L. (1983). *Multidimensional scaling*. New York: John Wiley.

Everitt, B. (1980). *Cluster analysis* (2nd ed.). New York: Halsted Press, A Division of John Wiley.

Funkhouser, E. T. (1996). The evaluative use of citation analysis for communication journals. *Human Communication Research, 22*(4), 563–574.

Garfield, E. (1994a). The impact factor. *Current Contents, 25*, 3-7.

Garfield, E. (1994b). Using the impact factor. *Current Contents, 29*, 3-5.

Garfield, E. (1995). Citation indexes for science: A new dimension in documentation through association of ideas. *Science, 122*(3159), 108-111.

Godlee, F., & Jefferson, T. (1999). *Peer review in health sciences*. London: BMJ Books.

Greene, J. C., & Caracelli, V. J. (Eds.). (1997). *Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms. New directions for program evaluation* (Vol. 74). San Francisco: Jossey-Bass.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11*, 255–274.

Hood, W. W., & Wilson, C. S. (2001). The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics, 52*(2), 291–314.

Institute of Medicine. (2004). *NIH extramural center programs*. Washington, DC: National Academies Press.

Kane, M., & Trochim, W. (2006). *Concept mapping for planning and evaluation*. Thousand Oaks, CA: Sage.

Kostoff, R. N. (1994a). Assessing research impact—Federal peer-review practices. *Evaluation Review, 18*(1), 31–40.

Kostoff, R. N. (1994b). Quantitative and qualitative federal research impact evaluation practices. *Technological Forecasting and Social Change, 45*(2), 189–205.

Kostoff, R. N. (1995). Research requirements for research impact assessment. *Research Policy, 24*(6), 869–882.

Krippendorf, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage.

Lane, N. F. (1995). What is the future of research—The science policy of the USA. *Interdisciplinary Science Reviews, 20*(2), 98–103.

McCullough, J. (1992). *Draft report of the NSF/Program Evaluation Staff Workshop on Methods for Evaluating Programs of Research Centers.* Washington, DC: National Science Foundation.

Molas-Gallart, J., & Davies, A. (2006). Toward theory-led evaluation: The experience of European science, technology, and innovation policies. *American Journal of Evaluation, 27*(1), 64-82.

Nass, S. J., & Stillman, B. W. (2003). *Large-scale biomedical science: Exploring strategies for future research*. Washington, DC: National Academies Press.

National Academy of Sciences, National Academy of Engineering, & Institute of Medicine. (1996). *An assessment of the National Science Foundation's Science and Technology Centers Program*. Washington, DC: National Academies Press.

National Cancer Institute. (2007). *Greater than the sum: Systems thinking in tobacco control* (Smoking and Tobacco Control Monograph Series). Bethesda, MD: U.S. Department of Health and Human Services, National Institutes of Health.

National Institutes of Health. (2006). *FY 2007 president's budget request*. NIH Office of Budget. Retrieved August 17, 2006, from http://officeofbudget.Od.Nih.Gov/fy07/supporting%20information.Pdf.

National Research Council. (1999). *Evaluating federal research programs: Research and the Government Performance and Results Act*. Washington, DC: National Academies Press.

Nye, M. (1996). *Before big science: The pursuit of modern chemistry and physics, 1800-1940*. New York: Twayne Publishers.

Office of Management and Budget. (1993). *Government Performance Results Act of 1993*. Washington, DC: Executive Office of the President of the United States. Retrieved August 17, 2006, from http://www.whitehouse.gov/omb/mgmt-gpra/gplaw2m.html.

Office of Management and Budget. (2007a). *Expectmore.Gov*. Retrieved August 12, 2007, from http://www.whitehouse.gov/omb/expectmore/.

Office of Management and Budget. (2007b). Program assessment: NSF's federally funded research and development centers. Retrieved August 12, 2007, from http://www.whitehouse.gov/omb/expectmore/summary/10004401.2005.html.

Office of Management and Budget. (2007c). Program Assessment Rating Tool (PART). Retrieved August 12, 2007, from http://www.whitehouse.gov/omb/part/index.html.

Osareh, F. (1996a, September). Bibliometrics, citation analysis and co-citation analysis: A review of literature I. *Libri, 46*, 149-158.

Osareh, F. (1996b, September). Bibliometrics, citation analysis and co-citation analysis: A review of literature II. *Libri, 46*, 217-225.

Rasmussen, N. (2002). Of "small men," big science and bigger business: The Second World War and biomedical research in the United States. *Minerva, 40*(2), 115–146.

Rosenberg, S., & Kim, M. P. (1975). The method of sorting as a data gathering procedure in multivariate research. *Multivariate Behavioral Research, 10,* 489–502.

Rosenfield, P. L. (1992). The potential of transdisciplinary research for sustaining and extending linkages between the health and social sciences. *Social Science and Medicine, 35*, 1343–1357.

Schoepflin, U., & Glanzel, W. (2001). Two decades of "scientometrics"—An interdisciplinary field represented by its leading journal. *Scientometrics, 50*(2), 301–312.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *British Medical Journal, 314*, 498-502.

Skoie, H. (1999). Bibliometrics—Some warnings from the north. *Scientometrics, 45*(3), 433–437.

Smith, R. (2001). Measuring the social impact of research: Difficult but necessary [Editorial]. *British Medical Journal, 323*, 528.

Stokols, D., Fuqua, J., Gress, J., Harvey, R., Phillips, K., Baezconde-Garbanati, L., et al. (2003, December). Evaluating transdisciplinary science. *Nicotine and Tobacco Research,* (5, Suppl. 1), S21–S39.

Trochim, W. (1984). *Research design for program evaluation: The regression-discontinuity approach*. Beverly Hills, CA: Sage.

Trochim, W. (1985). Pattern matching, validity, and conceptualization in program evaluation. *Evaluation Review, 9*(5), 575–604.

Trochim, W., & Linton, R. (1986). Conceptualization for planning and evaluation. *Evaluation and Program Planning*, *9*(4), 289–308.

U.S. Department of Health and Human Services. (2007). *Institutional Clinical and Translational Science Award*. Retrieved August 12, 2007, from http://grants.nih.gov/grants/guide/rfa-files/RFA-RM-07–002.html.

U.S. General Accounting Office. (2000). *NIH research: Improvements needed in monitoring external grants* (No. GAO-HEHS-AIMD-00–139). Washington, DC: Author.

U.S. General Accounting Office. (2005). *OMB's part reviews increased agencies' attention to improving evidence of program results* (No. GAO-06–67). Washington, DC: Author.

Weber, R. P. (1990). *Basic content analysis* (Vol. 49, 2nd ed.). Newbury Park, CA: Sage.

Weller, S. C., & Romney, A. K. (1988). *Systematic data collection*. Newbury Park, CA: Sage.

Whitmore, E. (Ed.). (1998). *Understanding and practicing participatory evaluation* (Vol. 80). San Francisco: Jossey-Bass.

Williams, B., & Imam, I. (Eds.). (2006). *Systems concepts in evaluation: An expert anthology*. Point Reyes CA: EdgePress.

W. K. Kellogg Foundation. (2001). *Logic model development guide: Using logic models to bring together planning, evaluation and action*. Battle Creek, MI: W. K. Kellogg Foundation.

The Evaluation of Large Research Initiatives: A Participatory Integrative Mixed-Methods Approach

William M. Trochim, Stephen E. Marcus, Louise C. Mâsse, Richard P. Moser and Patrick C. Weld

American Journal of Evaluation 2008; 29; 8
DOI: 10.1177/1098214007309280