5

The evaluator in this chapter provides a realistic account of the actions he would take to provide external evaluation services using the program theory–driven evaluation science approach.

Using Program Theory–Driven Evaluation Science to Crack the Da Vinci Code

Stewart I. Donaldson

Program theory–driven evaluation science uses substantive knowledge, as opposed to method proclivities, to guide program evaluations (Donaldson and Lipsey, forthcoming). It aspires to update, clarify, simplify, and make more accessible the evolving theory of evaluation practice commonly referred to as theory-driven or theory-based evaluation (Chen, 1990, 2004, 2005; Donaldson, 2003, forthcoming; Rossi, 2004; Rossi, Lipsey, and Freeman, 2004; Weiss, 1998, 2004a, 2004b).

This chapter describes in some detail how I would respond to the call from Mary García, principal of the Bunche–Da Vinci Learning Partnership Academy, asking my organization, DGHK Evaluation Associates, for a proposal to provide "an evaluation and recommendations for school improvement." Based on specific instructions from the editors, I have attempted to provide a realistic account of the actions I would take to provide evaluation services using the program theory–driven evaluation science approach. While I have avoided the temptation of simply explicating the principles and procedures for conducting program theory–driven evaluation science again (Donaldson, 2003, forthcoming; Donaldson and Gooler, 2003; Donaldson and Lipsey, forthcoming; Fitzpatrick, 2002), I do provide a limited amount of background rationale in key sections to help readers better understand my proposed actions.

This exercise was a stimulating and useful way to think about how I actually work and make decisions in practice. It was obviously not as interactive and dynamic an experience as working with real evaluation clients and

stakeholders. For example, conversations with stakeholders, observations, and other forms of data often uncover assumptions, contingencies, and constraints that are used to make decisions about evaluation designs and procedures. It was necessary at times to make assumptions based on the best information I could glean from the case description and my imagination or best guesses about the players and context. The major assumptions I made to allow me to illustrate likely scenarios are highlighted throughout my evaluation plan. My goal was to be as authentic and realistic as possible about proposing a plan to evaluate this complex program within the confines of everyday, real-world evaluation practice.

Cracking the Da Vinci Code

It is important to recognize that not all evaluation assignments are created equal. Program theory–driven evaluation science, and even external evaluation more generally, may not be appropriate or the best approach for dealing with some requests for evaluation. The Bunche–Da Vinci case, as presented, suggested that one or more of a highly complex set of potentially interactive factors might account for the problems it faced or possible ultimate outcome of concern: declining student performance. Principal Mary García appears exasperated, and district superintendent Douglas Chase at a loss for how to deal with the long list of seemingly insurmountable challenges for the Bunche–Da Vinci Learning Partnership Academy. Why has such a good idea gone bad? Why is performance declining? Could it be due to:

- A changing population?
- Social groupings of students?
- Student attendance problems?
- The curriculum?
- The innovative technology?
- Language barriers?
- Culturally insensitive curriculum and instruction?
- Staff turnover?
- The teachers' performance?
- Parenting practices?
- Leadership problems?
- Organizational problems?

And the list of questions could go on and on. How do we crack this "code of silence" or solve this complex mystery? "We've got it," say García and Chase. "Let's just turn to the Yellow Pages and call our local complex problem solvers: DGHK Evaluation Associates."

It appears to me on the surface that DGHK Evaluation Associates is being called in to help "solve" some seemingly complex and multidimensional instructional, social, personnel, and possibly leadership and

organizational problems. What I can surmise from this case description, among other characteristics, is:

- There appear to be many factors and levels of analysis to consider.
- Everyone is a suspect at this point (including García and Chase).

Some of the stakeholders in this case may have different understandings, views, and expectations about evaluation, and some may be very apprehensive or concerned about the powerful school administrators calling in outsiders to evaluate program and stakeholder performance.

The conditions listed above can be a recipe for external evaluation disaster, particularly if this case is not managed carefully and effectively. As a professional external evaluator, I do not have the magic tricks in my bag that would make me feel confident about guaranteeing Bunche–Da Vinci that I could solve this mystery swiftly and convincingly. However, I would be willing to propose a process and plan that I believe would stand a reasonable chance of yielding information and insights that could help them improve the way they educate their students. So how would I use and adapt program theory–driven evaluation science to work on this caper?

Negotiating a Realistic and Fair Contract

In my opinion, one of the key lessons from the history of evaluation practice is that program evaluations rarely satisfy all stakeholders' desires and aspirations. Unrealistic or poorly managed stakeholder expectations about the nature, benefits, costs, and risks of evaluation can quickly lead to undesirable conflicts and disputes, lack of evaluation use, and great dissatisfaction with evaluation teams and evaluations (see Donaldson, 2001a; Donaldson, Gooler, and Scriven, 2002). Therefore, my number one concern at this initial entry point was to develop realistic expectations and a contract that was reasonable and fair to both the stakeholders and the evaluation team.

The Bunche–Da Vinci Learning Partnership is a well-established, ongoing partnership program, with a relatively long (more than three years) and complex history. Therefore, I made the following two assumptions prior to my first meeting with García and Chase:

Assumption: There are serious evaluation design and data collection constraints. This is a very different situation from the ideal evaluation textbook case where the evaluation team is involved from the inception of the program and is commissioned to conduct a needs assessment, help with program design and implementation, and design the most rigorous outcome and efficiency evaluations possible.

Assumption: Money is an object. Based on the case description, I assumed that García and Chase desired the most cost-effective evaluation possible. That is, even if they do have access to substantial resources, they would

prefer to save as much of those as possible for other needs, such as providing more educational services.

It is important to note here that I would approach aspects of this evaluation very differently if money were no object (or if the evaluation budget were specified), and there were fewer design or data collection constraints.

Meeting 1. My first meeting with García and Chase was a success. I began the meeting by asking each of them to elaborate on their views about the nature of the program and its success and challenges. Although they had different views and perceptions at times, they seemed to genuinely appreciate that I was interested in their program and daily concerns. They also said they were relieved that I began our relationship by listening and learning, and not by lecturing them about my credentials, evaluation methods, measurement, and statistics, like some of the other evaluators with whom they have worked.

After García and Chase felt that they had provided me with what they wanted me to know about the partnership, I asked them to share what they hoped to gain by hiring an external evaluation team. In short, they wanted us to tell them why their state scores had declined and how to reverse this personally embarrassing and socially devastating trend. It was at that point that I began to describe how DGHK Evaluation Associates could provide evaluation services that might shed light on ways to improve how they were currently educating their students.

In an effort to be clear and concise, I started by describing in common language a simple three-step process that the DGHK Evaluation Team would follow:

1. We would engage relevant stakeholders in discussions to develop a common understanding of how the partnership is expected to enhance student learning and achievement. This is Step 1: Developing Program Theory. (Note that I rarely use the term *program theory* with stakeholders because it is often confusing and sometimes perceived as high brow and anxiety provoking.)
2. Once we have a common understanding or understandings (multiple program theories), I explain that we would engage relevant stakeholders in discussion about potential evaluation questions. This is Step 2: Formulating and Prioritizing Evaluation Questions.
3. Once stakeholders have identified the most important questions to answer, we will then help them design and conduct the most rigorous empirical evaluation possible within practical and resource constraints. Relevant stakeholders will also be engaged at this step to discuss and determine the types of evidence needed to accurately answer the key questions. This is Step 3: Answering Evaluation Questions.

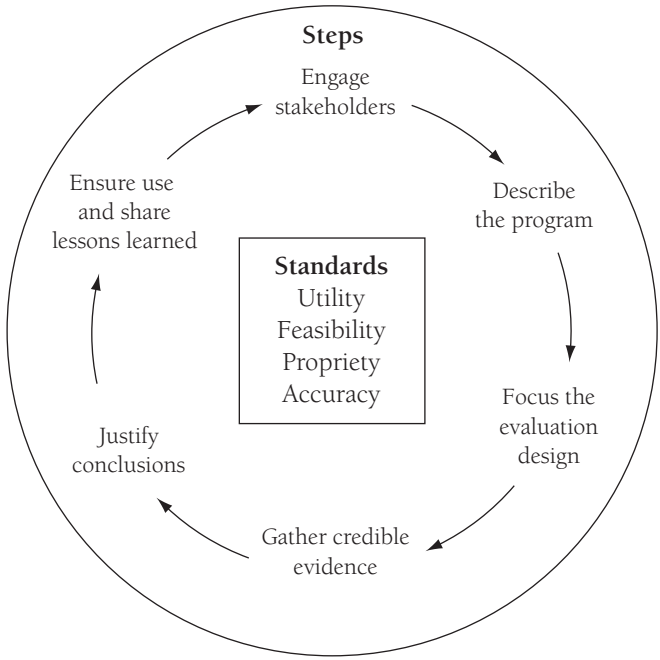
In general, I pledged that our team would strive to be as accurate and useful as possible, as well as participatory, inclusive, and empowering as the

context will allow. That is, sometimes stakeholders may choose not to be included, participate, or use the evaluation to foster program improvement and self-determination. At other times, resource and practical constraints limit the degree to which these goals can be reached in an evaluation. García and Chase seemed to like the general approach, but after they thought about it some more, they began to ask questions. They seemed particularly surprised by (and possibly concerned about) the openness of the approach and the willingness of the evaluation team to allow diverse stakeholder voices to influence decisions about the evaluation. They asked me if this was my unique approach to evaluation or if it was commonly accepted practice these days. I acknowledged that there are a variety of views and evaluation approaches (Alkin and Christie, 2004; Donaldson and Scriven, 2003), but pointed out that the most widely used textbooks in the field are now based on or give significant attention to this approach (examples are Rossi, Lipsey, and Freeman, 2004; Weiss, 1998). Furthermore, I revealed that many federal, state, and local organizations and agencies now use similar evaluation processes and procedures. They seemed relieved that I had not just cooked up my approach in isolation as a fancy way to share my opinions and render judgments. However, they did proceed to push me to give a specific example of one of these organizations or agencies. So I briefly described the Centers for Disease Control's six-step Program Evaluation Framework (1999).

The CDC Evaluation Framework is not only conceptually well developed and instructive for evaluation practitioners, it has been widely adopted for evaluating federally funded programs throughout the United States. This framework was developed by a large group of evaluators and consultants in an effort to incorporate, integrate, and make accessible to public health practitioners useful concepts and evaluation procedures from a range of evaluation approaches. Using the computer in García's office, I quickly downloaded Figure 5.1 from the CDC Web site.

I then proceeded to describe the similarities of the three- and six-step approaches. I explained that the first two steps of the CDC framework (engage stakeholders and describe the program) corresponded to what I described as the first step of the Bunche–Da Vinci Evaluation. The activities of CDC step 3 (focus the evaluation design) are what we will accomplish in the second step I described, and CDC steps 4 to 6 (gather credible evidence, justify conclusions, and ensure use and lessons learned) correspond to what we will achieve in step 3 of the Bunche–Da Vinci Evaluation. In addition, I explained how the standards for effective evaluation (utility, feasibility, propriety, and accuracy; Joint Committee on Standards for Educational Evaluation, 1994) and the American Evaluation Association's Guiding Principles (systematic inquiry, competence, integrity/honesty, respect for people, and responsibilities for general and public welfare; AEA Guiding Principles for Evaluators, 2004) are realized using this evaluation approach. Well, that did it; García and Chase were exhausted. They asked me if I could meet with them again next week to further discuss establishing an evaluation contract.

Figure 5.1. CDC Six-Step Evaluation Framework



Centers for Disease Control (1999).

Meeting 2. I could tell a considerable amount of discussion had occurred since our initial meeting. While it was clear they were eager to proceed, I could sense I was about to be bombarded with more questions. First, García wanted to know who would be engaging the stakeholders and facilitating the meetings to discuss the program and evaluation. My guess is she was concerned that I (a highly educated European American male) would be perceived as a threatening outsider and might not be the best choice for engaging her predominantly Latino and African American students, parents, staff, and teachers. This gave me the opportunity to impress on her that personnel recruitment, selection, and management is one of the most critical components of conducting a successful evaluation. She seemed to get this point when she thought about it in terms of problems she has encountered running her school. I assured her that we would strive to assemble a highly competent and experienced team with a particular emphasis on making sure we have team members knowledgeable about the program, context, and the evaluation topics we pursue. We would also make sure that we hire team members who share key background characteristics such as ethnicity, culture, language, and sociocultural experiences, and who possessed the ability to understand and build trusting and productive relationships with the

various stakeholder groups represented at the Bunche–Da Vinci Learning Academy. Furthermore, we would request funds to support hiring top-level experts to consult with us on topics we encounter that require highly specialized expertise. She very much liked the idea of supporting the assembly of a multicultural team as part of the evaluation contract.

Next, Chase wanted to know if there were any risks or common problems associated with engaging stakeholders. After reminding him of the potential benefits, I described some of the risks related to external evaluation in general, as well as to the evaluation plan I was proposing. For example, it is possible that various stakeholders (for example, the Da Vinci Learning Corporation administration or staff) will refuse to participate, provide misleading information, or undermine the evaluation in other ways. The evaluation findings might deliver various types of bad news, including uncovering unprofessional or illegal activities, and result in serious consequences for some stakeholders. Precious time and resources that could be used to provide services to needy students could be wasted if the evaluation is not accurate, useful, and cost-effective (see Donaldson, 2001a; Donaldson, Gooler, and Scriven, 2002, for more possible risks). Of course, I explained there are also serious risks associated with not evaluating at this point and that we would attempt to identify, manage, and prevent risks or negative consequences of our work every step of the way. He seemed pleasantly surprised that I was willing to discuss the dark side of external evaluation and was not just another evaluation salesperson.

After fielding a number of other good questions, concerns about budget and how much all this professional evaluation service will cost emerged in our discussion. I proposed to develop separate budgets for the conceptual work to be completed in steps 1 and 2 and the empirical evaluation work to be completed in step 3. That is, we would be willing to sign a contract that enabled us to complete the first two steps of developing program theory and formulating and prioritizing evaluation questions. Based on the mutual satisfaction and agreement of both parties, we would sign a second contract to carry out the empirical work necessary to answer the evaluation questions that are determined to be of most importance to the stakeholders.

Our evaluation proposal is intended to be cost-effective and to potentially save both parties a considerable amount of time and resources. During the completion of the first contract, Bunche–Da Vinci stakeholders will be able to assess the effectiveness of the evaluation team in this context and determine how much time and resources they want to commit to empirical data collection and evaluation. This first contract would provide enough resources and stability for our DGHK Evaluation Team to explore fully and better understand the program, context, stakeholders, and design and data collection constraints before committing to a specific evaluation design and data collection plan.

García and Chase seemed enthusiastic about the plan. They were ready to draw up the first contract so we could get to work. However, it dawned

on them that some of their key colleagues were still out of the loop. They began to discuss which one of them would announce and describe the evaluation to their colleagues. At that point, I offered to help. I suggested that they identify the leaders of the key stakeholder groups. After introducing and conveying their enthusiasm for the idea and the DGHK Evaluation Team (preferably in person or at least by telephone, as opposed to email), they would invite these leaders to an introductory meeting where the evaluation team would provide a brief overview of the evaluation plan and invite them to ask questions. García and Chase invited corporate, faculty, staff, parent, and student representatives to our next meeting to learn more about the evaluation plan.

I have tried to provide a realistic account of how I would attempt to negotiate an evaluation contract with these potential clients. As part of this dialogue, I have simulated the types of discussions and questions I commonly encounter in practice. I would assemble a multicultural team (drawing on existing DGHK Associates staff) to introduce the evaluation plan to the larger group of stakeholder leaders. The presentation would aim to be at about the same level as above, with some additional tailoring to reach and be sensitive to the audience.

Evaluation Plan

In this section, I add some flesh to the bones of the evaluation plan proposed. More specifically, I provide a brief rationale for each step, more details about the actions we will take, and some examples of what might happen as a result of our actions at each step of the plan. To stay within the bounds of this hypothetical case and intellectual exercise, I thought it would be useful to use a format that provides readers a window on how I would describe the Bunche-Da Vinci evaluation to prospective evaluation team members. Therefore, I will strive to illustrate the level of discussion and amount of detail I would typically provide to the candidates being interviewed for the DGHK Associates Multicultural Evaluation Team. My goal is to illustrate how I would provide a realistic job preview to those interested in joining the team, as a way to help readers gain a deeper understanding of my evaluation plan. Realistic job previews are popular human resource selection and organizational socialization interventions that involve explaining both desirable aspects of a job and potential challenges upfront, in an effort to improve person-job fit and performance and reduce employee dissatisfaction and turnover (Donaldson and Bligh, forthcoming).

Bunche–Da Vinci Realistic Job Preview

The evaluation of the Bunche–Da Vinci partnership will use a program theory–driven evaluation science framework. It will emphasize engaging relevant stakeholders from the outset to develop a common understanding of the program in context and realistic expectations about evaluation. We will

accomplish this by tailoring the evaluation to meet agreed-on values and goals. That is, a well-developed conceptual framework (program theory) will be developed and then used to tailor empirical evaluation work to answer as many key evaluation questions as possible within project resource and feasibility constraints. A special emphasis will be placed on making sure the evaluation team members, program theory, evaluation questions, evaluation procedures, and measures are sensitive to the cultural differences that are likely to emerge in this evaluation.

Step 1: Developing Program Theory. Our first task will be to talk to as many relevant stakeholders as possible to develop an understanding of how the Bunche–Da Vinci program is expected to meet the needs of its target population. For efficiency, we will work with four or five groups of five to seven stakeholders' representatives to gain a common understanding of the purposes and details about the operations of the program. Specifically, you (interviewee) will be asked to lead or be part of an interactive process that will make implicit stakeholder assumptions and understandings of the program explicit. [See Donaldson and Gooler (2003) and Fitzpatrick (2002) for a detailed discussion and examples of this interactive process applied to actual cases.]

Let me give you an example based on some of the characteristics and concerns I have learned so far. The Bunche–Da Vinci Learning Partnership Academy is an elementary school located in a tough neighborhood. It is a unique partnership between the school district and a nonprofit educational company specializing in innovative school interventions for low-performing students. The school population is characterized by high transience, illegal enrollments from the adjacent district, high numbers of non-English-speaking students, high levels of poverty, a young and inexperienced staff with high turnover, and geographical isolation from the rest of the district. The principal and superintendent are concerned that the partnership program is not an effective way to educate their students. They have shared with me a number of hunches they have about why the program is not working, and the principal has some ideas about how to change and improve the school. But it is important to keep in mind that as we engage other stakeholders in discussions about the program, we are likely to gain a wealth of additional information and possibly hear extremely different views about the program's success and challenges.

After we collect and process the information we gather in the stakeholder meetings, we will attempt to isolate program components, expected short-term outcomes, more long-term or ultimately desired outcomes, and potential moderating factors (Donaldson, 2001b). For example, student performance on state test scores has been the main desired outcome discussed in the conversations I have had with the school administrators. Other stakeholders may strongly object to the notion that the program is designed to improve state test scores and be upset by the No Child Left Behind legislation and

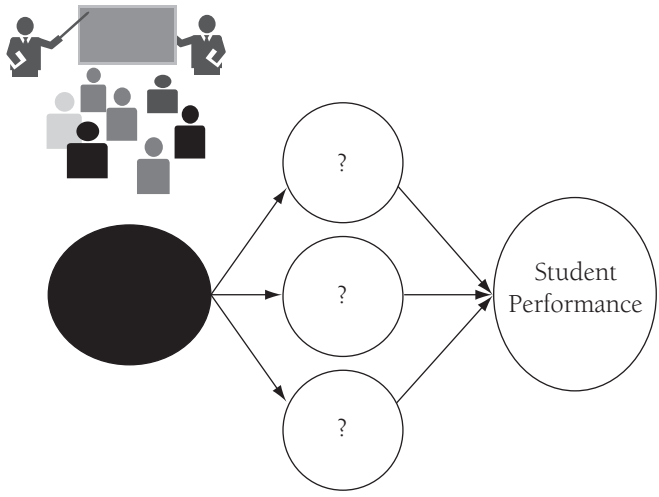
accountability zeitgeist. In fact, I expect they will provide us with a range of other desired outcomes to consider as we try to gain a deep understand of the program. However, the program's impact on student performance will likely end up being one of the desired outcomes we explore conceptually and potentially evaluate.

Figure 5.2 shows an example of how we would begin to diagram and probe stakeholders' views about program impact on student performance. The anchor of this program impact theory is student performance. It assumes that the partnership program compared to no program or an alternative (for example, a typical curriculum in a comparable school) is expected to improve student performance. Our discussion with the stakeholders would attempt to clarify why or how the partnership program is presumed to accomplish this. We may discover that there are the key short-term outcomes or mediating factors (represented by the question marks at this point) that are expected to result from the program, which in turn are expected to lead to improved student performance. Once we have clarified these expected mediating processes, we will begin to probe whether these links are expected to be the same for all students and in all context variations that may exist across the delivery of the program. If not, we will isolate the key student characteristics (such as gender, ethnicity, socioeconomic status, language, acculturation, attendance) and potential contextual factors (such as group or class dynamics, instructor effects, service delivery characteristics, and the like) that could moderate or condition the strength or direction of the arrows in our program impact theory. Our ultimate goal is to work through this interactive process with the diverse stakeholder groups until we have a common understanding about the purposes and expected benefits and outcomes of the program.

Once we have completed this process with the stakeholders, you and the other members of the team will be required to assess the plausibility of the stakeholders' program theory or theories. You will do this by reviewing the available research and evaluation literature related to factors identified. We will specifically look for evidence that may suggest that some of the links are not plausible or that there may be side effects or unintended consequences we have not considered. The findings from the review, analysis, and team discussions may lead us to suggest that the stakeholders consider revising or making some additions to their program theory(ies). I expect it will take us at least three months of full-time work to complete this first step of our evaluation plan.

Step 2: Formulating and Prioritizing Evaluation Questions. Once we have a deep understanding of the program and context, we will focus on illuminating empirical evaluation options for the stakeholders. You and the rest of the team will be asked to frame potential evaluation questions so that they are as concrete and specific as possible and informed by the program theory(ies). The types of questions that will likely be considered in the Bunche–Da Vinci evaluation fall under the categories of program need, design, delivery or

Figure 5.2. Example of Program Impact Theory



implementation, outcomes, cost, and efficiency (see Rossi, Lipsey, and Freeman, 2004).

My best guess, based on what we know so far about the partnership program, is that we will most likely pursue questions about curriculum implementation, program operations and educational services delivery, and program outcomes. For example, the stakeholders may decide that they want empirical data to answer questions such as:

- Are administrative and educational service objectives being met?
- Are the intended curricula being delivered with high fidelity to the intended students?
- Are there students or families with unmet needs that the program is not reaching?
- Do sufficient numbers of students attend and complete the curriculum?
- Are teachers and students satisfied with the curriculum and educational services?
- Are administrative, organizational, and personnel functions of the partnership program effective?

Furthermore, questions about program outcomes will likely include:

- Are the desired short-term outcomes (mediators) being achieved?
- Are the desired longer-term or ultimate outcomes of concern being achieved?
- Does the program have any adverse side effects?

- Are some recipients affected more by the program than others (moderator effects)?
- Does the program work better under some conditions than others (moderator effects)?

It is possible that we will be asked to pursue questions about the partnership program cost and efficiency—for example, (1) Are the resources being used efficiently? (2) Is the cost reasonable in relation to the benefits? (3) Would alternative educational approaches yield equivalent or more benefits at less cost? Furthermore, García does have some ideas for changing the program and may ask us to answer questions about student needs and best ways to satisfy those needs. But I do think it is likely you will be pursuing questions about program implementation and outcomes if both parties decide to enter into a second contract to collect data to answer the stakeholders' evaluation questions.

Once a wide range of potential evaluation questions has been formulated, you and the DGHK Evaluation Team will help the stakeholders prioritize the questions so that it is clear which questions are of most value. You will need to note differences of opinion about the value of each question across the stakeholder groups and factor them into final decisions about which questions to pursue in the evaluation. In an ideal evaluation world, the entire range of relevant evaluation questions would be answered, and the program impact theory would be tested in the most rigorous fashion possible. However, in most evaluations, only a subset of questions and components of a program impact theory can be evaluated due to time, resource, and practical constraints. Prioritizing and identifying the most important evaluation questions can prevent paralysis in evaluation (that is, deciding to wait or not to evaluate at all). Recognizing that some key questions can be addressed, even though some or many components of the program impact theory and other evaluation questions cannot be examined at this time, will help you to facilitate the evaluation process to move forward.

Finally, it is important for you to realize as a prospective employee and team member that the stakeholders may decide that they have learned enough about the program after we complete these first two steps. For example, it is not uncommon to discover that a program is obviously not being implemented as intended. This could lead the stakeholders to focus their attention and resources on fixing the program, or they may determine it is not repairable and decide to terminate the program and replace it with a promising alternative. If this type of situation develops, it is possible we would obtain an evaluation contract to help them develop and evaluate the new initiative, or we are likely to have another interesting evaluation contract in the firm that you could be hired to work on.

Step 3: Answering Evaluation Questions. Assuming the stakeholders decided they wanted to enter into the second contract with us to answer

key evaluation questions, you would be asked to help design and work on an evaluation that would strive to answer those questions convincingly. In many respects, our evaluation approach is method neutral. We believe that quantitative, qualitative, or mixed methods designs are neither superior nor applicable in every evaluation situation (Chen, 1997; Reichhart and Rallis, 1994). Instead, our methodological choices in this evaluation will be informed by program theory, the specific evaluation questions the stakeholders have ranked in order of priority, validity and use concerns, and resource and practical constraints (feasibility). Your main charge at this stage of the evaluation plan is to determine what type of evidence is needed and obtainable, to answer stakeholder questions of interest with an acceptable level of confidence.

As you might know, several factors typically interact to determine how to collect the evidence needed to answer the key evaluation questions such as stakeholder preferences, feasibility issues, resource constraints, and evaluation team expertise. Program theory–driven evaluation science is primarily concerned with making sure the data collection methods are systematic and rigorous, and produce accurate data, rather than privileging one method or data collection strategy over another (see Donaldson and Christie, forthcoming).

It is typically highly desirable to design evaluations so that stakeholders agree up front that the design will produce credible results and answer the key evaluation questions. Participation and buy-in can increase the odds that stakeholders will accept and use evaluation results that do not confirm their expectations or desires. Of course, making sure the design conforms to the Joint Committee Standards (Joint Committee on Standards for Education Evaluation, 1994) and the American Evaluation Association's *Guiding Principles for Evaluators* (2004) as much as possible is also helpful for establishing credibility, confidence, and use of the findings.

The challenge for us in the Bunche–Da Vinci Evaluation will be to gain agreement among the diverse stakeholder groups about which questions to pursue and the types of evidence to gather to answer those key questions. For example, we should be able to gather data from students, teachers, school and corporate administration and staff, parents, and experts in areas of concern. Furthermore, if needed, it looks as if we will be able to collect and access data using existing performance measures and data sets, document and curriculum review, interview methods, Web-based and traditional survey methods, possibly observational methods, focus groups, and expert analysis.

As a member of the evaluation team, your role at this stage of the process will be to help facilitate discussions with the relevant stakeholders about the potential benefits and risks of using the range of data sources and methods available to answer each evaluation question of interest. You will be required to educate the stakeholders about the likelihood that each method under consideration will produce accurate data. You will need to discuss potential threats to validity and possible alternative explanations of findings (Shadish, Cook, and Campbell, 2002), the likelihood of obtaining accurate and useful data

from the method in this specific situation, research with human participants (especially with minors), and informed-consent concerns, and to estimate the cost of obtaining data using each method under consideration. Once the stakeholders are fully informed, your job will be to facilitate a discussion that leads to agreement about which sources and methods of data collection to use to answer the stakeholders' questions. This collaborative process will also include reaching agreement on criteria of merit (Scriven, 2003) or agreeing on what would constitute success or failure or a favorable or unfavorable outcome, which will help us justify evaluation conclusions and recommendations. If you are successful at fully informing and engaging the stakeholders, we believe it is much more likely that the stakeholders will accept, use, and disseminate the findings and lessons learned. If you will allow me to make some assumptions, I can give you examples of how this third step of our evaluation plan could play out in the Bunche–Da Vinci evaluation you would be hired to work on.

First, I must underscore that the information we have so far is from the school administrators. That is, once we hear from the teachers, parents, students, and corporate administration and staff, we may gain a different account of the situation. In our view, it would be a fundamental error at this point to base the evaluation on this potentially limited perspective and not include the other stakeholders' views. Nevertheless, I will make some assumptions based on their perspectives to illustrate the process that you will be asked to facilitate at this stage of the evaluation.

Assumption: The stakeholders have decided that their top priority for the evaluation is to determine why the two separate indicators of student performance (state test scores; Da Vinci test scores) provide substantially different results.

García and Chase have suggested that their top concern is that student performance, particularly on the English–Language Arts components of state standards tests, has declined over the course of the partnership program. In fact, in the most recent testing (year 3), students scored lower than in years 1 and 2. These decreasing state test scores contrasted sharply with their corporate partner's assessments of students' performance. On the company's measures, the percentage of students reading at grade level had doubled over the past three years. Da Vinci staff from headquarters claimed to have heard students say that they were "finally able to read" and were much more enthusiastic learners. García and Chase are greatly concerned about this discrepancy.

Assumption: We will assume that the stakeholders have produced Figure 5.2 and that increasing student performance is one of the main purposes of the Partnership Program.

In this case, we would explore all of the strengths and weaknesses of the feasible options for determining why measures of student performance are

yielding different results. I would expect the stakeholders to decide to have us conduct a systematic and rigorous analysis of the construct validity of each set of measures. We would pay particularly close attention to potential differences in construct validity across our diverse and changing student body. In addition to the expertise on our team, we would likely hire top-level consultants with specific expertise in student performance measurement in similar urban school environments to help us shed light on the discrepancies. It will be critical at this stage of the evaluation to decide whether the performance problems are real or an artifact of inadequate measurement that can be explained by threats to construct validity (Shadish, Cook, and Campbell, 2002). Imagine the implications of the potential evaluation finding that performance is not really declining or if corporate measures are seriously (possibly intentionally) flawed.

Assumption: The stakeholders decided that their second priority is to determine if the partnership program curriculum is being implemented as planned with high fidelity.

We will have learned much about the design of the curriculum and why it is believed to be better than the alternatives during steps 1 and 2 of our evaluation process. It will be your job as a member of the evaluation team to help verify whether the curriculum is actually being implemented as intended. As is sometimes the case in educational settings, school administrators suspect the teachers might be the main problem. They have suggested to us that teachers have not bought into the partnership program, may even resent its requirements and added demands, and may just be going through the motions. They have also suggested to us that the teachers are young and inexperienced and may not have the motivation, expertise, and support necessary to implement the program with high fidelity. Furthermore, there are some doubts about whether groups of students are fully participating in the program and adequately completing the lesson plans. The rather dramatic changes in the student population may have affected the quality of the implementation of the curriculum.

After considering the available data collection options for answering this question, I would expect you and the evaluation team to be asked to observe and interview representative samples of teachers about the delivery of the Bunche–Da Vinci curriculum. You could also be asked to gather data from students and parents about their experiences with the curriculum. For these technology-enriched students, I would expect that a Web-based survey could be designed and completed by many of the students. However, alternative measurement procedures would need to be developed for those too young to complete a Web-based survey. Furthermore, a representative sample of students could be interviewed in more depth. For the parents, I would expect we would need to design a standard paper-and-pencil survey and be able to interview and possibly conduct some focus groups to ascertain their views

and experiences. Finally, interviews of key staff members of both the school district and corporate partner would be pursued to further develop our understanding of how well the curriculum has been implemented and how to improve implementation moving forward. Keep in mind that even if we do find valid student performance indicators in the previous analysis, they could be meaningless in terms of evaluating the partnership program if the program has not been implemented with high fidelity.

The final example of a question that could be pursued in step 3 of the Bunche–Da Vinci Evaluation focuses on determining whether desired short-term outcomes have resulted from the program. For this example, I will assume that the partnership program has been implemented with high fidelity. Figure 5.2 shows that three main short-term outcomes are expected to result from the partnership program. Let us assume that the stakeholders have agreed that the first one is a high level of intrinsic engagement of the curriculum. That is, the program produces a high level of intrinsic engagement, which in turn is expected to lead to increases in student performance.

Assumption: The stakeholders decided their third priority is to assess whether Bunch–Da Vinci students have a high level of intrinsic engagement.

Imagine that after weighing the options, the stakeholders have decided that they would like us to measure and determine whether students at Bunche–Da Vinci have a high level of intrinsic engagement with the curriculum. You and the team would be asked to work with the stakeholders to develop a clear understanding and definition of this construct. Next, we would search and critically review the literature to determine if there are good measures of this construct that could be used for this purpose. Assuming we have identified a strong measurement instrument (we would need to create one otherwise), we would then develop another set of measurement procedures for the students to complete (or include the items on the previous instruments used above, if possible). We would also make sure that we included items about key student characteristics of concern (such as ethnicity, language, acculturation, family, attitudes toward technology, and the like) and characteristics of the context (peer group dynamics and out-of-class study environment, for example), exploring if there may be important moderating influences on the link between the program and this short-term outcome. This will allow us to do more finely grained analyses to estimate whether the program is affecting this short-term outcome more for some students than others.

We will assume that for this initial examination of intrinsic engagement, we do not have a reasonable comparison group or baseline data available. Therefore, prior to implementation, we will gain agreement with the stakeholders about how we will define high versus low levels of intrinsic engagement (that is, establish criteria of merit). Finally, against our recommendation, we will assume that the stakeholders have decided not to survey

or interview parents or teachers about intrinsic engagement due to their lack of enthusiasm about spending additional resources for their third priority evaluation question.

As I hope you can now appreciate, working on the DGHK Evaluation of Bunche–Da Vinci promises to be a meaningful opportunity for you. You will be part of a multicultural team of evaluation professionals engaged in helping to address a socially important set of concerns. The course of these children's educational careers and lives could be undermined if we find that this situation is as bad as it appears on the surface and sound recommendations for improvement are not found and implemented in the near future. I would now like to hear more about why your background, skills, and career aspirations make you a strong candidate for being an effective member of the DGHK/Bunche–Da Vinci Evaluation Team. But first, do have any further questions about the job requirements?

Reflections and Conclusions

The Bunche–Da Vinci Case presented DGHK Evaluation Associates with a challenging mystery to be solved. A highly complex set of potentially interactive factors appears to be suspect in the apparent demise of an innovative partnership program. Whomever or whatever is the culprit in this case seems to be responsible for undermining the performance of a diverse and disadvantaged group of students. In the face of this complexity, DGHK Associates has proposed to use a relatively straightforward three-step process to develop and conduct evaluation services to help crack the Da Vinci Code and to potentially improve the lives and trajectories of these children. The proposed evaluation approach is designed to provide cost-effective, external evaluation services. DGHK Associates promises to strive to provide evaluation services that are as accurate and useful as possible to the Bunche–Da Vinci stakeholders, as well as to work in a manner that is participatory, inclusive, and as empowering as stakeholders and constraints will permit.

In an effort to achieve these promises, I have proposed to tailor the evaluation to contingencies our team encounters as they engage stakeholders in the evaluation process. Obviously, to complete this exercise of describing how program theory–driven evaluation science could be adapted and applied to this hypothetical case, I had to make many assumptions. Examples of the details of each step could be substantially different in practice if different assumptions were made. For example, if I assumed the stakeholders wanted us to propose how we would determine the impact of the program on student performance outcomes using a rigorous and resource-intensive randomized control trial (or quasi-experimental design, longitudinal measurement study, intensive case study, or something else), the particulars of the three steps would differ substantially. However, it is important to emphasize that the overall evaluation plan and process I proposed would be virtually the same.

The sample dialogue with the school administrators during the contracting phase and in the realistic job preview I gave to the potential evaluation team members were intended to be helpful for understanding how I provide evaluation services in real-world settings. Based on the case description, I predicted this evaluation would need to operate under somewhat tight resource and practical constraints and would be likely to uncover intense conflicts and dynamics among stakeholder groups. I tried to underscore the point that a fatal flaw would have been to design an evaluation plan in response to information and views provided almost entirely by one powerful stakeholder group (the school administrators, in this case). It seemed likely that the teachers and teachers' union may have made some different attributions (for example, management, leadership, and organizational problems) about the long list of problems and concerns the administrators attributed to the "young and inexperienced" teachers. It also seemed likely that the corporate leadership and staff could have a very different take on the situation. Based on my experience, I am confident that the failure to incorporate these types of stakeholder dynamics in the evaluation plan and process would likely undermine the possibility of DGHK Associates' producing an accurate and useful evaluation for the Bunche–Da Vinci Learning Partnership.

It would have also been problematic to conduct extensive (and expensive) data collection under the assumption that student performance had actually declined. That is, a considerable amount of evaluation time and resources could have been expended on answering questions related to why performance had declined over time, when in fact performance was not declining or even improving, as one of the indicators suggested. Therefore, in this case, it seemed crucial to resolve the discrepancy between the performance indicators before pursuing evaluation questions based on the assumption that performance had actually declined.

Due to space limitations, there are aspects of this case and evaluation plan I was not able to explore or elaborate on in much detail. For example, during the developing program theory phase of the evaluation process, we would have explored in detail the content of the innovative, technology-enriched curriculum and its relevance to the needs of the culturally diverse and changing student population. During step 3, we would have facilitated discussions with the stakeholders to determine how best to disseminate evaluation findings and the lessons learned from the Bunche–Da Vinci evaluation. Furthermore, we would have explored the potential benefits and costs of spending additional resources on hiring another evaluation team to conduct a meta-evaluation of our work.

In the end, I must admit I encountered strong mixed emotions as I worked on this hypothetical case and evaluation plan. As I allowed my imagination to explore fully the context and lives of these students and families, I quickly felt sad and depressed about their conditions and potential plight, but passionate about the need for and opportunity to provide help

and external evaluation. As I allowed myself to imagine what could be done using external evaluation if there were no time, resource, and practical constraints, I became elated and appreciative about being trained in evaluation and inspired to apply evaluation as widely as possible. However, this was quickly dampened when I realized I have never encountered a real case in twenty years of practice without time, resource, and practical constraints. My spirits were lowered even more when I imagined the risk of using scarce resources to pay the salaries and expenses of well-educated professionals to provide unnecessary or ineffective evaluation services, when these resources would otherwise be used to educate and help these at-risk students and families. Of course, the beauty of this exercise, just like in a nightmare, is that I would quickly elevate my mood by reminding myself I am dreaming. Now that I (and my colleagues in this volume) have walked this imaginary tightrope with you, I hope you have a better understanding of the value, challenges, and risks of external evaluation. I imagine I do.

References

Alkin, M. C., and Christie, C. A. "An Evaluation Theory Tree." In M. C. Alkin (ed.), *Evaluation Roots.* Thousand Oaks, Calif.: Sage, 2004.

American Evaluation Association. *Guiding Principles for Evaluators.* 2004. http://www.eval.org.

Centers for Disease Control. *Centers for Disease Control Program Evaluation Framework.* Atlanta: Centers for Disease Control, 1999.

Chen, H. T. *Theory-Driven Evaluations.* Thousand Oaks, Calif.: Sage, 1990.

Chen, H. T. "Applying Mixed Methods Under the Framework of Theory-Driven Evaluations." In J. C. Greene and V. J. Caracelli (eds.), *Advances in Mixed-Method Evaluation: The Challenges and Benefits of Integrating Diverse Paradigms.* New Directions for Evaluation, no. 74. San Francisco: Jossey-Bass, 1997.

Chen, H. T. "The Roots of Theory-Driven Evaluation: Current Views and Origins." In M. C. Alkin (ed.), *Evaluation Roots.* Thousand Oaks, Calif.: Sage, 2004.

Chen, H. T. *Practical Program Evaluation: Assessing and Improving Planning, Implementation, and Effectiveness.* Thousand Oaks, Calif.: Sage, 2005.

Donaldson, S. I. "Overcoming Our Negative Reputation: Evaluation Becomes Known as a Helping Profession." *American Journal of Evaluation,* 2001a, *22,* 355–361.

Donaldson, S. I. "Mediator and Moderator Analysis in Program Development." In S. Sussman (ed.), *Handbook of Program Development for Health Behavior Research.* Thousand Oaks, Calif.: Sage, 2001b.

Donaldson, S. I. "Theory-Driven Program Evaluation in the New Millennium." In S. I. Donaldson and M. Scriven (eds.), *Evaluating Social Programs and Problems: Visions for the New Millennium.* Mahwah, N.J.: Erlbaum, 2003.

Donaldson, S. I. *Program Theory–Driven Evaluation Science: Strategies and Applications.* Mahwah, N.J.: Erlbaum, forthcoming.

Donaldson, S. I., and Bligh, M. "Rewarding Careers Applying Positive Psychological Science to Improve Quality of Work Life and Organizational Effectiveness." In S. I. Donaldson, D. E. Berger, and K. Pezdek (eds.), *Applied Psychology: New Frontiers and Rewarding Careers.* Mahwah, N.J.: Erlbaum, forthcoming.

Donaldson, S. I., and Christie, C. A. "The 2004 Claremont Debate: Lipsey vs. Scriven: Determining Causality in Program Evaluation and Applied Research: Should Experimental Evidence Be the Gold Standard?" *Journal of Multidisciplinary Evaluation,* forthcoming.

Donaldson, S. I., and Gooler, L. E. "Theory-Driven Evaluation in Action: Lessons from a $20 Million Statewide Work and Health Initiative." *Evaluation and Program Planning,* 2003, *26,* 355–366.

Donaldson, S. I., Gooler, L. E., and Scriven, M. "Strategies for Managing Evaluation Anxiety: Toward a Psychology of Program Evaluation." *American Journal of Evaluation,* 2002, *23*(3), 261–273.

Donaldson, S. I., and Lipsey, M. W. "Roles for Theory in Evaluation Practice." In I. Shaw, J. Greene, and M. Mark (eds.), *Handbook of Evaluation.* Thousand Oaks, Calif.: Sage, forthcoming.

Donaldson, S. I., and Scriven, M. (eds.). *Evaluating Social Programs and Problems: Visions for the New Millennium.* Mahwah, N.J.: Erlbaum, 2003.

Fitzpatrick, J. "Dialog with Stewart Donaldson." *American Journal of Evaluation,* 2002, *23*(3), 347–365.

Joint Committee on Standards for Education Evaluation. *The Program Evaluation Standards: How to Assess Evaluations of Educational Programs.* Thousand Oaks, Calif.: Sage, 1994.

Reichhart, C., and Rallis, C. S. (eds.). *The Qualitative-Quantitative Debate: New Perspectives.* New Directions for Program Evaluation, no. 61. San Francisco: Jossey-Bass, 1994.

Rossi, P. H. "My Views of Evaluation and Their Origins." In M. C. Alkin (ed.), *Evaluation Roots.* Thousand Oaks, Calif.: Sage, 2004.

Rossi, P. H., Lipsey, M. W., and Freeman, H. E. *Evaluation: A Systematic Approach.* (7th ed.) Thousand Oaks, Calif.: Sage, 2004.

Scriven, M. "Evaluation in the New Millennium: The Transdisciplinary Vision." In S. I. Donaldson and M. Scriven (eds.), *Evaluating Social Programs and Problems: Visions for the New Millennium.* Mahwah, N.J.: Erlbaum, 2003.

Shadish, W. R., Cook, T. D., and Campbell, D. T. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin, 2002.

Weiss, C. H. *Evaluation: Methods for Studying Programs and Policies.* (2nd ed.) Upper Saddle River, N.J.: Prentice Hall, 1998.

Weiss, C. H. "Rooting for Evaluation: A Cliff Notes Version of My Work." In M. C. Alkin (ed.), *Evaluation Roots.* Thousand Oaks, Calif.: Sage, 2004a.

Weiss, C. H. "On Theory-Based Evaluation: Winning Friends and Influencing People." *Evaluation Exchange,* 2004b, *9*(4), 1–5.

STEWART I. DONALDSON is dean and professor of psychology at the School of Behavioral and Organizational Sciences, Claremont Graduate University.