

Dialogue with Stewart Donaldson

JODY FITZPATRICK

Editor: You've selected your evaluation of the California Wellness Foundation's Work and Health Initiative (WHI) as representative of your approach to theory-driven program evaluation. Why do you think this particular evaluation is most illustrative of your approach?

Donaldson: I've been conducting theory-driven evaluations for close to a couple of decades now, and one of the things that I've noticed is that some people think this approach is very academic—an ivory tower approach almost like laboratory research. I would say that probably most published evaluations are what I would call efficacy evaluations and I've done these types of evaluations. This is when the investigator sets the conditions for the study and has complete controls. If theory-driven evaluation is good only for that purpose, it's pretty limited. That's not a robust way to practice evaluation.

I chose this application for our interview because it illustrates how the theory-driven approach is much more robust than the efficacy trial. It can be used to examine how an intervention works out in real organizations, in the real world where there's not a lot of control over who gets into the program and who doesn't. This is evaluating in a real world context that is similar to what most evaluators face. So, I wanted to focus on this evaluation because I wanted to show that theory-driven evaluation can be applied in the circumstances most evaluators work in.

Editor: Can you tell me a bit more about what you mean by efficacy evaluations?

Donaldson: Let's take a prototypical case. A university researcher has a theory about something, say how to prevent drug abuse. The researcher marshals the resources to develop a curriculum or program based on her or his theory, gets funding from the federal government, and quickly moves into what I would see as a research study or efficacy trial. What they're really testing is "under ideal conditions can this program have an effect?" This is a science project. If you get good results, that doesn't mean you can take the program out into an agency and deliver it on a mass scale and necessarily replicate the effects. Before it can have a big impact on society, it has to be effective in the field. I would argue most evaluations in the literature are efficacy trials, not effectiveness trials. We need effectiveness trials to help us learn if interventions or programs make a difference in the real world.

Editor: So you see the evaluation literature dominated by efficacy trials and these not reflecting the real work of programs or evaluators. Why do you think the literature differs so much from actual practice?

Jody Fitzpatrick • Graduate School of Public Affairs, University of Colorado, Colorado Springs, CO 80907, USA; Tel: (1) 303-393-8567; E-mail: jfitzpat@carbon.cudenver.edu.

American Journal of Evaluation, Vol. 23, No. 3, 2002, pp. 347–365. All rights of reproduction in any form reserved. ISSN: 1098-2140 © 2002 by American Evaluation Association. Published by Elsevier Science Inc. All rights reserved.

Donaldson: To gain acceptance as a legitimate scientific endeavor, program evaluation or intervention research most often follows the well-respected experimental paradigm. Recent developments in intervention research synthesis, including meta-analysis and meta-meta analysis (meta-analyzing meta-analyses), reveal a vast published literature using experimental or quasi-experimental designs to determine the impact of programs or interventions. It is generally known that it is more difficult to publish evaluations that don't conform to these scientific standards. Ideally, a program should first pass the test of an efficacy trial. But, my main point is it is not safe to assume the same effects will occur when the program is delivered under "real world conditions," nor is it realistic to believe that methods and procedures used in efficacy trials are always feasible and appropriate in effectiveness trials. Theory-driven evaluation is one of a few evaluation approaches providing a framework for evaluators to operate effectively in both worlds.

Editor: Coming back to the WHI, what aspect of this evaluation did you think was the most exemplary?

Donaldson: Given the large number of people who were involved and the differences across sites, I think our ability to build the continuous improvement function into the evaluation was the most exemplary part. By continuous improvement, I mean what many traditionally consider formative evaluation. In doing evaluation for continuous improvement there's a lot of detailed work. It was gratifying, toward the end of the project, to see how the evaluation had been used to improve the programs along the way, and to find out that many of the stakeholders believed that the evaluation process created a community of learners and "learning organizations."

When we first began working with these same stakeholders (e.g., program designers, managers, trainers, Foundation staff, etc.), they seemed to have had a very different view of evaluation. Some saw it as something they would be punished with and believed it wouldn't have any value to them. At most of the sites and programs we were able to make the evaluation timely and useful. We tried to help them mount programs that had a chance to be implemented. At the same time, we let them know we were going to do a summative evaluation at the end.

Editor: The four programs that emerged from the WHI were quite different. What similarities were there in developing the program theory?

Donaldson: The similarities are in the process and role of the evaluator. Even though you're trying to engage stakeholders in this process, the evaluator really has to be a strong leader. Related to that, you don't use evaluation language when you develop program theory. Most grantees or stakeholders don't understand it. Theory has kind of an academic, negative connotation. Instead we say something like, "Before we do any evaluation, we really want to understand your program. For us to design a good evaluation, we really need to be sensitive to what you're trying to do." As they're talking about it, we're trying to conceptualize it.

One thing I've learned over time is that it's really important not to treat this as a one-directional process. It's important to move right to left *and* left to right, really probing the assumptions in somebody's program theory. I now instruct my students to start at the right (with desired outcomes), to ask "Where are you going with all this?" Once they are clear about where they believe they want to go, we explore how they plan to get there. Sometimes it's really helpful to take just one participant. How will this participant be different or change after completing this program? How does this change or transformation occur? In other words, start with what the client will be like at the end and move backwards. Most people do exactly the opposite. Most people want to start on the left side of the model, with the details of the

program. If you start with “What are the components of the program?”, you can quickly get lost in details that aren’t central to the program theory. So, instead start with defining what the program is trying to achieve, what the client will ideally be like in the end.

For example, in one of the programs, Computers in Our Future (CIOF), stakeholders were concerned with describing all of the ways they were going to have participants interact with computers. Instead of starting with that level of detail, we started by asking them to define the purpose of the program. They collectively viewed the purpose to be improving the health of the participants. On the surface, most people would not find it reasonable to believe that learning computer programs is a major determinant of health. However, after we worked for a considerable amount of time helping the stakeholders conceptualize their program, we were able to articulate a concise program theory and help them design a program that made this link more plausible.

This bi-directional approach to making implicit program theory explicit has several other important advantages. It is very common to discover the program under investigation is poorly conceptualized. For example, when you take a presumed outcome and examine it carefully, it is common to find that the program as designed has very little chance, based on what we know from prior research and evaluation, to account for a significant portion of the variance in the presumed outcome. It is also common to discover that the relationship between a program and its presumed outcomes may be conditional or better explained by taking moderating factors into account. All of this can be useful for improving the design of the program in advance of implementation and evaluation, which can lead to substantial savings in time and resources. In addition, once the presumed program theory has been formulated, it is important to conceptualize alternative program theories and potential unintended consequences. This step can be used to identify plausible competing program theories and potential side effects that should be examined and/or minimized.

Editor: What differences occurred in developing program theory across the project?

Donaldson: Throughout this project, some groups were much more anxious and hostile about the evaluation than others. Their feelings, of course, make frequent interactions somewhat stressful and unpredictable. Some groups were happy to be part of the process while others were very nervous about trying to articulate a program theory because they were skeptical about the evaluation itself. It’s easier to develop program theory when you have a motivated group that has a well-formulated program. Winning New Jobs (WNJ) was the easiest because this was a program that had been well developed and tested at the University of Michigan through randomized designs and under a number of conditions, though it was initially developed for auto workers who were downsized in Detroit. So, the real question with WNJ was “Can you take that program and implement it in California with a very different population?” And at least the Michigan group was very clear about what they were trying to accomplish and were open to the evaluation.

For two other projects in the Initiative defining program theory was much more difficult. As I mentioned, CIOF, unlike WNJ which began with a previously formulated and tested model, began with a generally defined problem—kids and young adults in the inner cities were being left out of the computer and information revolution and, in order for them to have good employment situations, something needed to be done about that. So, the grant was to develop programs that made sense and seemed culturally appropriate to the local communities to meet this issue. There was no curriculum in place. Many of the people we dealt with really had not thought much as to how they were going to get from Point A to Point B. So, really the exercise of developing program theory itself had a huge impact on program development.

The other difficult project was the Future of Work and Health (FWH) program. It was difficult to get the first grantee to articulate program theory and to work with the evaluation team. There was a lot of resistance and no real buy-in. For a number of reasons, the grantee decided not to continue with the grant. Then, the Foundation decided to take it inside, to do the program themselves. The Foundation had always been the client we were evaluating *for*. Now, they began to have a dual role. They seemed to become less enthusiastic about the evaluation of the FWH program. They felt the same level of evaluation that was going on in the other programs wasn't necessary.

In the end, we could only get to a really basic kind of model arguing, "You have to at least try to tell us what you're trying to accomplish with this program." That program theory really didn't emerge until close to the end of the project.

Editor: Those examples are very helpful in showing an array of different circumstances. With CIOF you indicate developing a program theory which didn't exist yet had a huge impact on the program. Can you tell us more about that?

Donaldson: In the CIOF program there were 11 diverse grantees responsible for developing 14 community computing centers. The process of developing program theory eventually led to a group vision of acceptable, desired short-term outcomes (e.g., technology skills, career development knowledge, job search skills, basic life skills), desired long-term outcomes (e.g., increased participation in future education, internship, and employment opportunities), and identified curricula that needed to be developed and implemented in order to achieve these goals. Without a common program theory, it is very likely grantees would have done very different activities that would have been ill-defined, unmeasurable, and not tightly linked to desired outcomes and the goals of the Initiative.

Editor: The resistance of the initial grantee with the Future of Work and Health project is also interesting and a not uncommon problem. Can you tell us a bit about how you attempted to work with that grantee? What do you think went wrong?

Donaldson: In my view, the primary problem was the grantee was not comfortable working with Foundation staff and the evaluation team to clearly define the program. It seemed like the grantee did not want to be held accountable for a set of pre-defined activities. Instead, the preference was to do things and see how they turned out (a trial and error approach). The tension around disagreements about how to manage this program between Foundation staff and grantee staff eventually led to the termination of the grant. The pressure that was imposed by initiating discussions of program theory forced the problem to be addressed sooner rather than later and saved the Foundation from additional costs in terms of time and valuable resources.

Editor: You've talked about the differences between working with these groups. What else makes developing program theory difficult or relatively straightforward?

Donaldson: Three things are important. The phase in development of the program is a key issue. If a program has been well-tested and been running and is clearly defined, it's easier to develop program theory as opposed to when they have the money and sort of know what they want to do, but aren't that sure. Then, you're helping them develop it. A second issue is some programs are more complex than others. Some are pretty molar, monolithic, and have clear direct effects. Most human service programs in this Initiative were complex and multi-faceted. The effects are subtle. They are diffuse. There is a lot more complexity. So, to get a good conceptualization is more time consuming and difficult. The third issue is the group dynamics. If the two groups, the evaluation team and the program team, can set up a good relationship with clearly defined roles, the program people understand why we're defining program theory at this time. That makes it go pretty fast and it's pretty enjoyable. On the other hand, if they

don't like the evaluation team, if they don't like being evaluated, or there's dissension within the group itself, it's much more difficult. There was one group that I think really assigned roles. One person from the program was always attacking the evaluation team, so the program leader would remain in good standing.

Editor: Did you consider developing an overall program theory for the WHI to show how choices were made about directions in funding and options?

Donaldson: One of the things we signed up for in our original proposal was evaluating the four components and then evaluating the entire Initiative. One thing that was nice was we were one of the earliest grantees. Usually evaluators come in downstream. So, in the first year when many sites weren't even selected yet, we were spending a lot of time trying to develop an overall framework for the Initiative and to develop, I don't know if I would call it a program theory, but at least a conceptual framework because from an outside perspective these are really four very different programs. We started to make some progress. We had a couple of big meetings, but once we started defining the work, it became really clear that people did not have anything in their budget to work on cross-cutting issues. We began reporting this to the Foundation—that the program people are drowning in the detail of their own programs and that we would need additional resources to get at the whole macro level. The Foundation wasn't willing to put more resources in it. The Foundation felt if they pushed people into cross-cutting issues, it would hinder the development of their own programs.

Editor: Perhaps we could focus on WNJ for a minute. The theoretical model for WNJ is essentially this: WNJ builds job search confidence, job search skills and problem-solving strategies. Each of these immediate outcomes then contributes to both reemployment and better mental health, the ultimate outcomes. Can you tell us a bit about how you developed the program theory for the Winning New Jobs program? What controversies or questions emerged?

Donaldson: Fortunately, there was an extensive theoretical, research, and evaluation literature to draw upon. We also traveled to Michigan to observe some of the training for those who were going to implement the program, to interact with the participants, and we were able to observe the program in action. This helped us to develop a grounded understanding of what the participants would experience. We had discussions with program managers and trainers at each of the implementation sites in California and had extensive discussions with the Manpower Research Demonstration group, who were primarily responsible for making sure the program was implemented according to plan in the California organizations. By observing the program in action and from all of these discussions that is, from talking with trainers (from the California sites) and with the trainers of the trainers (from Michigan), it became pretty clear what the core model was.

A central issue that came up was whether they should adapt or try to remain true to the original model. There were some different views on that, to the point that the decision was made to do it only in English which, in California, cuts out a significant portion of the population. For the most part, the Foundation supported the Michigan team's view that sites should strive to remain true to the model. A question for the evaluation became: if you deliver the program as it's supposed to be delivered, do you get the immediate outcomes? The curriculum was adapted somewhat so that it was relevant to diverse California workers, but they stayed with the same conceptual model—if you can get those immediate outcomes—improved confidence, job search skills, better problem-solving strategies—the participants should have better employment and mental health outcomes. So, a real issue was would this work with an ethnically, educationally, and occupationally diverse population. These were questions that emerged. The developers saw this as a relatively robust program that should work across the

world. Journals had published much on the success of this model. But, some of the people in California felt they needed to give it a California flavor.

The Michigan team seemed very interested in issues of local adaptation, but was concerned that too much deviation from the model could lead to null effects. At the end, we saw there were clearly some things they could have done to make WNJ more effective in California. For example, language and cultural adaptations would have enabled them to serve much more of the population in need, and additional services to remove non-psychological employment obstacles that many Californians faced would have helped improve reemployment outcomes. The issue of maintaining program fidelity versus making significant local adaptations was a balancing act and seemed to produce creative tension throughout the project. In the end, after hearing all of the arguments, the Foundation staff made the decisions on this issue because they controlled the resources.

Editor: Did you test to see if the program theory's assumptions concerning unemployment and its causes were the case with the California population?

Donaldson: The Michigan group presumably had done a needs assessment in Detroit. They believed that psychological factors were limiting reemployment efforts. That was the basic assumption. There are people in California who fit that profile. They really do need to be motivated and this would work well. But, if the clients have needs beyond that, as we found that many of the program participants did, more service is needed. As we went around watching the program being implemented we saw these other needs. We went into Fresno and there's an exercise dealing with why employers don't hire you. One of the first answers from a participant is, "My criminal record hinders my being hired." The next participant says, "It's all my tattoos and body piercing." The next one says, "I don't have a driver's license." Then, we went to the Silicon Valley site and the answer is, "They just want to hire young people. They want people right out of college." These were higher-level professionals dealing with age issues. The characteristics of the workers, occupations, and economy seem to make a big difference.

Editor: I can see how these sites differ from each other, but how do they differ from the model?

Donaldson: Well, if you give them job search skills, the program works as intended in achieving its immediate goals. They get job search skills, but they still don't have driver's licenses or child care. It's not going to work in achieving the ultimate goal—getting them employed—because these other needs aren't met. We had a lot of discussions with the Foundation and program staff about whether psychological factors are a major barrier to employment for the California worker we were serving, and it always came back to the empirical question—can we use this model with different populations? No matter what your issues are, this should have some effect. Again, out in these effectiveness situations they don't have much control over who comes into the setting.

They also struggled with whom they should try to reach. Do they go after the Latino population? Do we limit it to the "ideal" participant? The thing that the project team did agree on, based on the research base from other work, was that depression was an appropriate factor to use as a screen. If a person is showing signs of depression, they can ruin the group dynamics that occur during the program and are necessary to gain job search confidence and problem-solving skills.

Editor: So people running the programs weren't very involved in these discussions?

Donaldson: Not early on because the sites were just being selected then. The discussion continued when we brought the program people on board later. But, the program theory was already well in place at that point.

For the most part, we did see the immediate changes they were expecting. Most participants seemed to gain confidence, job search skills, problem-solving strategies, and 65% of the participants reported being employed 180 days after the program. Of course, another external factor relevant to reemployment is whether there are available jobs in their community that match their skills.

The model for WNJ represented a mid-level of complexity among the four Initiative programs. The model for another program (HIPPP) was a little more detailed and for the FWH, where we had difficulty getting the model articulated, the final model consisted of only one step.

Editor: How complex should a model be to adequately convey the theory of a program?

Donaldson: In trying to critically understand a program, to develop program theory, you can think and think and think until you get this massively complex model because many things are very complex. That tends to be a real hindrance in an evaluation. I strive for a fairly parsimonious model. If a group of stakeholders wants to do some logic modeling of key links, that's OK, but that's different than developing the program theory. Program theory clearly represents the presumed active ingredients of the program and how they lead to desired outcomes. You're trying to consider whether the paths leading into this outcome have any chance at all for accounting for a significant portion of the variance in the desired outcome. Often when you do that, you see there's not a chance that these actions could lead to a significant change in the variance of that outcome. I take each path and really think carefully about the timing, when it should occur and whether there are key moderators of that path. But, to put that all in the model is really counterproductive; however, as evaluators we really think carefully about these paths. I try to address complexity in the evaluation war-room. If we think there's a real key moderator, we really want to measure and test that.

Editor: In your writing on program theory, you discuss four sources of information for program theory: prior theory and research, implicit theories of those close to the program, observations of the program, and exploratory research to test critical assumptions. To what extent did you rely on each of these in your development of theory for WNJ?

Donaldson: The extent to which you can do all of these in any one program theory is limited. In practice, the program dictates which method you're more able to use. With WNJ, we relied very heavily on looking at prior theory and research around this whole model. With most programs I come across in the effectiveness area, there's very little prior research, but with WNJ there was. We went to Michigan. We could see that this prior research really influenced their model. We didn't do exploratory research in this case, though we did observe the program in action. When we were in the process of developing program theory for WNJ, we were more observing the program being implemented in other settings and the trainers from Michigan training the trainers that would implement the program in California. I can't emphasize how important it is to pull away from the conceptual jargon and go see what's happening. I'm a strong believer in trying to observe the program in action and doing it in a way that you're not changing the way the program is delivered when you are not present. You want to get a realistic view of what's happening.

Editor: Let's move away from program theory to the development of the evaluation questions. You note in your report that the evaluation was participatory. Typically, evaluation questions were developed through discussions with staff and managers from the sites as well as program managers from the Foundation. Can you tell us a bit about how you worked with these stakeholders to develop the questions?

Donaldson: When we initially set up the contract for the evaluation the Foundation was very interested in being involved in the process and having all the key grantees involved. So,

in developing the program theory and the evaluation questions we identified all the people we thought would benefit from being included at this stage. And, then we used a number of ways to communicate, the most personal being face-to-face meetings which we would have fairly regularly, not at Claremont, but at the Foundation or at their site so we could also do observations. We had a lot, probably in the hundreds, of conference calls because people were geographically dispersed. So, it would be common to have eight to ten people involved in a conference call across the state. We also used a lot of e-mail.

Every year the Foundation pulled all the grantees together to review the Initiative. So, at least once a year, usually more often, we had face-to-face contact. We had to have everybody's buy-in before we moved forward. Grantees ranged from the management team at the Foundation to people at all the sites and sometimes sites had sub-contracts so there were representatives at all levels.

Editor: Did you involve clients in your planning?

Donaldson: The clients were represented through the actual trainers. Also, from time to time, we would observe and get to interact in a meeting format with selected clients. It wasn't feasible to engage the clients at the same level as the other stakeholders though.

Editor: How did you decide on the final focus? For example, with WNJ, you focus on five core areas—implementation, service (population served), short-term outcomes, reemployment outcomes (longer-term outcomes), and sustainability and replication. These five core areas are pretty comprehensive. Were there areas you considered and rejected?

Donaldson: Once we had agreement on the conceptual theory, we moved to step two, formulating and prioritizing the evaluation questions. The evaluation team is really the facilitator and leader. We're not saying "What do you want to do?" But, "Here's what we could do. These are the things that are most feasible." Then, sometimes, others would say, "No, no, what about this?" The Foundation played a real hands-on role here; they were very involved. Others, because of the funding relationship, would want to follow the lead of the Foundation. As the Foundation people were involved, we decided on these core areas. WNJ was moving faster than Computers in our Future program. When we got to the point of developing the program theory for CIOF, the Foundation staff often wanted to look at WNJ as a model. The major group decisions—and this is not always ideal—were largely influenced by the Foundation representatives, but with a lot of input from everyone else. In a sense, they were paying for all of this, so being a key client, they would come down on certain sides of issues and influence the process.

What did we consider and reject? There were a number of things we could have done and didn't. One of the big tensions in the project was at the method level and another tension involved measuring mental health outcomes. This is a health foundation; we're working on employment to improve health. We strongly encouraged looking at some outcome measures of health and had a fair amount of support from the group for this. But, about half of the program team and many of the sites felt very uncomfortable asking mental health questions of participants. They felt it would send the wrong message. We had to compromise and just focus on the reemployment outcomes. Others argued the ultimate purpose is health. But both the sites and the Foundation said we had to really know what happens with re-employment. Looking at health outcomes is something that would have been nice to do, but we didn't.

The big issue was the evaluation team and the Michigan people who had developed the model really felt we should pursue conducting a randomized trial or at least a quasi-experiment. But, there was intense resistance to this approach from others who would have to implement

this type of design. The counter position was that this was a demonstration project and we already had the experimental data from earlier studies.

Editor: I noticed across both WNJ and CIOF you have similar types of evaluation questions, for example, some questions involving implementation (describing program delivery), service (describing recipients), and, then, assessing various outcomes. Is this a common pattern that emerges in theory-driven evaluation?

Donaldson: This is a common pattern. It's a simple framework. Each program has an action theory. It will achieve something, an immediate outcome. And the program has to be delivered well to achieve that outcome. So, each of these, the critical program actions and the immediate outcomes, needs to be examined. Then, we can move to the conceptual level. If we achieve the immediate outcomes (e.g., job search confidence and skills) do they in fact lead to the desired long-term outcomes (e.g., reemployment)? You usually have at least three levels being addressed—actions, immediate outcomes, and long-term outcomes.

Editor: You note your evaluation was both formative and summative. Who were your primary audiences for this study? How did these audiences (and others) use the information?

Donaldson: We started with the formative piece. Remember, this was a six-year process. Many of these programs had a lot of implementation and design issues. For four to five years we were in this heavy formative evaluation phase, but still building databases for the summative piece. We produced over 200 evaluation reports and we tried to give rapid cycle feedback to the relevant grantees. And then the Foundation, program management team, and sites were also reading the reports and trying to make decisions with respect to how to implement the programs. For the formative phase, the audiences were the Foundation, the program management team, the trainers—all up and down the organizations involved.

People were getting reports and sometimes reacting with concerns. Few like to be evaluated, but they were getting regular evaluation feedback from us from the start. Most of the grantees would agree that this feedback influenced how far they were able to get with the program. That is, regular evaluation feedback and discussion about program accomplishments and challenges dramatically enhanced the quality of program implementation.

For the summative report, the first set of users was the Foundation Board and staff. They've invested somewhere around \$22 million on this Initiative. Now that we've been evaluating the programs for about six years, we need to lay it all out. They want to know whether this was worth their money. So, they're the first stop—the President, the staff, the Board. Then, we arranged with them, once that process is completed, to disseminate the findings as broadly as possible so others can learn about the Initiative and we distributed summative evaluation reports to the grantees. So, we involved everyone in the summative evaluation as well.

Because sustainability is a big issue, the sites can now use these numbers to try to get more funding and learn how to move forward. We have pretty massive databases. We get e-mails and calls asking us to look at certain things for them, so we're still answering questions for them.

Editor: Often formative evaluations are most directed to program managers and staff because they're the closest to the program, the people most able to make changes. What would the Foundation do with the formative evaluation?

Donaldson: There was a large range of stakeholders and grantees. In addition to a formal report, we did have a relationship with the sites where we would give feedback in between the formal reports. We also allowed them to have input into our formative report, not to change major findings, but they could make factual corrections or append a challenge to our findings. The Foundation had a senior program officer who ran each of these programs. They would

receive copies and be involved in discussions of the reports. In some ways they would sit back and listen to discussion by the sites, program management, and evaluation teams about issues contained in the reports.

Someone from outside might say, "Boy, this foundation was micro-managing," but the general feel was they wanted to be hands-on and be a part of it. The Foundation would say to everyone publicly that the evaluation was just one form of input. The evaluators don't always see things the Foundation staff see. There are other forms of input. All of the sites sent reports to the Foundation. If some issue came up, the players could build a case against it in their reports. The Foundation also told the sites, "You're funded now. The evaluation won't influence that funding." At first no one believed it, but as we modeled it over time, people became quite comfortable. Oftentimes, the Foundation would side with the grantees.

There is one thing that I thought was critical to this whole relationship with these various audiences and this was our use of 360° feedback. When you step back and look at some of the literature on stakeholder-evaluator relationships and the psychology of evaluation, when you get into this frequent interaction, you can quickly get into view that the evaluators are always the ones criticizing. The program people can think the evaluators aren't doing anything, just critiquing others doing the "real" work. It became clear to us that we needed to let all the players evaluate us and the Foundation. So, we set up a process where they could give us formative feedback on how the evaluation was going.

Editor: The 360° feedback sounds like an interesting way to change the environment. How did that work? How did the grantees give feedback? Do you think they felt safe in what they could say?

Donaldson: Each year the grantees were asked to provide feedback about the strengths and limitations of the management of the Initiative by the Foundation and the evaluation of the Initiative by CGU. We experimented with methods for doing this. For example, one year the evaluation team received the feedback and reported about it; another year the Foundation team received the feedback. In the end, we decided an ideal procedure would be to have someone outside of the Initiative run the 360° process and make it as confidential as possible. Of course, it seems impossible to create a situation where the grantees would feel completely comfortable writing a negative evaluation of the Foundation that is funding them. As some of the Foundation staff would joke, "You have many 'friends' and there are few barking dogs when your job is to give away money." Nevertheless, many of the grantees reported liking the opportunity to evaluate instead of always being evaluated.

Editor: What was the most useful feedback the evaluation team received?

Donaldson: The most useful feedback about the evaluation itself was the observation that early on the evaluation reports were primarily focused on program challenges. I think most evaluators assume that is where they add value and forget about the psychology of evaluation. Grantees are looking to see if you have seen and reported all their program accomplishments. Based on this feedback, our reports began by detailing program accomplishments before we noted the challenges. It was amazing the positive effect this had on the process. Grantees were much more open to addressing challenges if they believe we were aware, and made others aware, of their accomplishments. A little sugar helps the bitter pill go down.

The Foundation staff reported liking and learning from this process as well. Evaluation of the Foundation program officers by grantees during the first three years of the Initiative revealed several factors that contributed toward successful management of the overall Initiative. These included Foundation staff's sensitivity and responsiveness to the concerns and needs of grantees and demonstrated interest and involvement of program officers in grant

programs. In addition, grantees viewed TCWF program officers as very accessible, approachable, and hands-on. Program officers were also commended for being solution-focused, rather than problem-focused. Grantees suggested several areas for improvement which the Foundation attempted to address including: (1) clarifying roles and responsibilities of Foundation staff, program management teams, and sites, (2) providing more education on how Foundations operate, (3) communicating clearly about changes in Foundation policies or program expectations, and (4) hiring a professional facilitator for the convenings of grantees.

Editor: You note you used a process of continuous program improvement. What were some changes that occurred as a result of that process?

Donaldson: As noted above, approximately 200 evaluation reports were written and discussed with the grantees throughout the project. I believe there were numerous changes that occurred as a result of these reports and discussions. For example, eligibility criteria, recruiting strategies, and target numbers were modified in WNJ; report content, report format, and dissemination strategies were changed in HIPP; and the content of the core curriculum in CIOF was developed and changed in response to this continuous improvement process using formative evaluation. However, when you are involved in this dynamic process with over 40 organizations it is risky to claim you can establish cause and effect here. That is, the Foundation and grantees often internalize and use the evaluation feedback (which is a good thing) and see this as part of the program management and development process, as opposed to attributing these changes solely to the evaluation. That is, it was common for the evaluation team members to feel like others were taking credit for their findings, ideas, and insights.

Editor: Do you always have a focus on continuous program improvement?

Donaldson: No, because sometimes the evaluation budget won't cover the expense. I do evaluations where that is just not in the budget. But, when you have complex human service programs, I feel that's where the big payoffs are. It's very rare that I find a program that is so sound that we just do the summative piece. I'm sure such programs exist, but they're few and far between.

What I focus on in continuous improvement is the speed of the feedback. We look for early warning signs that occur between formal reports. If you look at the literature and writings on utilization, many people argue there is more use for enlightenment than direct utilization. But, when you get into this continuous improvement model, you couldn't be further from the truth. I guess you could get into a position where people aren't using what you're given, but on this project we had a big influence. I would love to do a project where half receive continuous improvement feedback and the other half do not and examine the differences. But, in this project, when we weren't evaluating something, we could see a year where no progress was going on. If no one's looking, nothing may be going on.

Editor: In quite a few cases things didn't work out as planned. For example, with WNJ the target number to be served is ultimately reduced to half the original goal due to recruitment problems and the eligibility criteria are broadened. Also, you suggest skill development was a greater need than building self-confidence and job search skills which had been the focus of WNJ. How did the different stakeholders react to these findings?

Donaldson: I'll make a general statement first. When you get into these feedback cycles and continuous improvement, as you deliver what we might think of as negative feedback on areas that need improvement, the stakeholders are in a position where they can go one of two routes. One route is, "What we've been doing so far hasn't worked. How do we improve on that? What different strategies can we use?" The other main route is, "Maybe that's not a realistic goal." Instead of changing activities, they change their goals. Oftentimes, they feel

if they have the opportunities and support to change goals, they would much rather do that. With WNJ, the original vision was 10,000 participants. As we began going through the process and selecting sites and looking at sites' capacity and population there was discussion by the sites, program management team, and the Foundation that led to the feeling that they couldn't achieve 10,000. So, the Foundation cut it down further to be about 6,500. Then, as we were in the continuous improvement cycle, some sites were really struggling and were off track in reaching their goals. So, rather than fail and have a big gap in the end, there was a decision made by the Foundation, which we weren't necessarily supportive of, to bring the target more in line with where we would end up. The revised goal was 5,000, so the sites have met their numbers. Now, the Foundation in retrospect really questioned whether they should have changed the goal. There's nothing wrong with having 6,500, and so you're a little short. We don't expect perfection, but keeping the number at 6,500 might have motivated the sites to close that gap.

We handled it in this way. When you're writing a summative report, sometimes those responsible for delivering the program want you to leave out the history, but we didn't. We said, "Here's where we started. Here are the changes along the way". We reported the changes in numbers so, we're not leaving things out. That may not present as rosy of a picture. Some questioned whether we really need all that in there. We said, "Your input is important, but at the end of the day when we have all the evidence, it's important for us to own the report," and the President of the Foundation was strongly behind that. We were able to retain our objective, independent view of things and we believe that benefits everyone. These Board members are very smart and pretty tough. If they think we're just doing public relations (PR) for the program officers, it won't be credible.

Editor: I'm also interested in your conclusion that more of the focus for job training needed to be on skill building. More than half of the population you served had education beyond high school. Did this group still need skill building or were you referring to other potential audiences? How did you reach the conclusion that they needed skill building?

Donaldson: When you go out and look at how the program is being implemented, there are three very distinct communities. It's kind of hard to see it's the same program because of the populations involved. I referred to how different the participants were in Silicon Valley from those in Fresno. The way the exercises went was very different. If we were looking in Silicon Valley, for example, where many participants had been let go from a tech company, part of what we would hear was they hired young people right out of college with up-to-date technology skills, so someone who had been working for a while and was not on the cutting edge would be replaced by a younger person with more current skills and also at a cheaper salary. They would put these people through a motivational workshop (WNJ) and they would get through and be really geared up to go, but they still wouldn't have the high-tech skills that companies want. So, the motivation wasn't enough. And, that same process was going on in Fresno though it looked a lot different. If someone doesn't have a drivers' license and didn't have the skills for the job they want to go for, the motivation and what was in the curriculum wasn't enough. This concern was something that we raised from the start with the program. Is there a way to do more tailoring and needs analysis? But it's really difficult once a Foundation decides what they're going to do, for us to say, "Let's go back and consider if this is the right program." Michael Scriven was on the Board of Advisors for the evaluation and when he first came on he did a needs assessment for CIOF, but it had no impact on the project. It was just too late for this type of input to influence program design; the train had left the station.

A big goal for the Foundation was that once they pulled their money out these programs would still be part of the agency. This didn't happen with the one exception of one of the

organizations. That organization in Fresno has participants first deal with basic things (e.g., having transportation, child care, skills), then they put them through WNJ to gain the confidence and job search skills. That makes a lot of sense to me.

Editor: But how did you arrive at the decision that skills were what was needed? Did that emerge from data you collected?

Donaldson: Job-relevant skills were one of the additional needs that some of the participants identified, and we noted this in our observations of the program in action. This finding was confirmed by data collected during interviews of the WNJ trainers and program managers and was consistent with other research on employment issues in California.

Editor: Your data collection makes use of both quantitative and qualitative methods, though with the large numbers of people in the two service delivery programs your emphasis there is primarily quantitative. With WNJ you measure implementation through structured observation forms and participant reaction forms. You collect data on participants' demographics and employment history and pre-post paper-and-pencil data on some of the immediate program goals, for example, e.g., self-efficacy, self-mastery. Then, you conduct telephone interviews for four periods following the workshop to describe employment outcomes. How did program theory contribute to your selection of measures? How do you decide what measures to use?

Donaldson: As I look across all the evaluations I've worked on using this approach, the best pay off for spending all that time up-front really trying to understand the process you're trying to evaluate and articulating the program theory is that it helps you to design a much better evaluation. It's so easy to miss things if you don't understand the program in a grounded way. It's easy to miss things about how the program is supposed to lead to certain outcomes. That up-front work helps you identify the key concepts that you should measure and the timing of those measures.

If you don't collect data at the right time, you're going to miss it. There's a large literature suggesting that most evaluations are very insensitive. I'm referring to meta-analyses and some of Mark Lipsey's work. By using program theory to help you consider the timing of your measures, you're in a position to collect information on the right things at the right time. This is where the interplay of going back and looking at other research on the topic is very useful. You may find another program looked at changes in participants at 3 months and found nothing but at 6 months found a whopping effect, so that really helps you in timing your measures. The program theory of WNJ helped us decide when to measure reemployment as well as the intermediate outcomes.

The other issue is that when you really get down to it, what this Initiative was about was improving something way down the road, health. A lot of people who don't go through the exercise of developing program theory say, "This program is about health and that's 20 years down the road, so it's not worth evaluating." But, program theory shows you the intermediate steps to measure.

Editor: Which measures that you used on WNJ did you find to be the most meaningful or informative?

Donaldson: What I have found very informative, although not completely conclusive, was comparing our numbers to those found with the efficacy trials in Michigan and other places around the world. They had roughly the same programs and had pure control groups and experimental groups. We modeled our measures after those used in the efficacy trials. When we compared our group to their experimental group and control group, considering how much more disadvantaged our population was, our numbers still looked pretty good. I thought that was very informative. In the other trials they were dealing with someone who had been

recently unemployed, auto plant workers in Michigan. Those people were devastated and the program was trying to keep them from going downhill. The control group received written materials on job information—real standard information. It does not seem surprising to me that this program is better than a booklet. But, in our evaluation, the reemployment numbers looked pretty good especially considering participants, on average, had been unemployed for 12 months. However, it is important to note reemployment programs are constrained by the available jobs in the local community. For the most part, this program was implemented during good economic times.

The implementation data—both qualitative and quantitative—were very useful. There are so many things that don't happen the way they plan them to in the program design. Data bring that to surface and you can deal with it.

Re-employment outcomes—all the sites really see the game as reemployment. We tried to compare our numbers to critical competitors in the communities, other programs for the unemployed, and that worked against us. Few other programs kept data and those that did collected it for PR or promotional uses. We had pretty rigorous data, so we're always going to look worse than PR data. So, the re-employment data finally gave a benchmark to how it was working. On the other hand, when people were trying to get it adopted in their organizations, others would say, "This looks much worse than our other programs." The key to the goals of replication and institutionalization is you have to be able to demonstrate they're better than existing services. We weren't able to do that because we weren't able to study existing services. We strongly encouraged the Foundation to consider requiring that the control group be the best existing program on future evaluations.

Editor: Which types of data collection were most persuasive to your different stakeholders?

Donaldson: The trainers and program managers really liked qualitative data. They liked anecdotes and stories of success in particular for obvious reasons. People with a stake in the success of the program often assume that if you do qualitative measurement you will get a more accurate and positive picture. That's interesting to think about. Is there some artifact to the qualitative approach? Or, is it that the quantitative is insensitive to changes?

Editor: Were your qualitative results more positive than the quantitative?

Donaldson: No, we're aware of these issues. So, we made sure we weren't using weak qualitative measures. A lot of the qualitative measures were extremely informative.

People with an obvious stake in outcomes like qualitative success stories. Outsiders, such as Board members and representative of the Foundation, want the numbers. That's really clear to me. For me personally, I get a lot out of the mixed design approach. Numbers give me one piece of the picture, but I get a lot out of interviewing people. It's the interplay between the qualitative and quantitative that is really informative. In increasing our depth of understanding, it's the qualitative things that push us forward. But, the quantitative is necessary and does give an important piece.

You can really get bogged down in methods issues. A huge advantage of the theory-driven approach is we avoid getting bogged down in any of these issues until we have a really developed program theory. There is a tendency to want to start thinking about methods before the program theory or conceptual framework is developed.

Editor: Can you give an example from WNJ of some findings based on qualitative data that were really useful for you?

Donaldson: I discussed some of the qualitative findings about participant needs and the value of observing the program in operation earlier. We also gathered a wealth of qualitative

data about program implementation by interviewing the trainers and program staff. This was very useful for understanding program implementation challenges, as well as how similar the WNJ program implementation was to the original JOBS model. At the end of WNJ (and all the WHI programs), we interviewed most of the key people involved in the project to understand their views on program success and challenges. These data were also quite informative for identifying key lessons learned and for generating ideas for how to improve this kind of effort in the future.

Editor: This was really a massive evaluation that continued for four to six years depending on the program. How did things change as time passed?

Donaldson: Obviously the Initiative in each program went through phases. Initially, we were just trying to develop an evaluation framework that everyone would buy into. With some programs, for example with CIOF, there was a lot of discussion about just what a program would look like so the kinds of things the evaluator does at that stage are very different from working with a program that is more well-defined and developed. For example, the discussion is much more focused on testing program theory, as opposed to program development, when a program is clearly defined.

For a long time, we really didn't have anything that might be considered recommendations for improvement because we weren't collecting data, but we were giving feedback. As soon as there was data, the whole dynamic changed dramatically. Now, they could, with some degree of confidence, reject evaluation findings. The key is how you model this process. If they react and you react or get defensive, you're not modeling what you want this to become. You're trying to model that everyone is trying to learn. Then, as they see their funding is not being affected, they become more comfortable with hearing and accepting evaluation findings and conclusions. It takes some cycles for people to believe you're doing what you say you will do.

Then, the other major transition is at the summative stage. With the continuous improvement model, at some point you have to notify everyone that you're no longer involved in just improving, but need to move to summative evaluation. As you can imagine, it has taken a while to get them to buy into the continuous improvement, learning organization model, so that then when you move to a summative phase the anxiety and tension rises again. Now, we're saying we are going to judge the merit, worth, and significance of the work. They think, "You're doing what we thought you would do in the beginning!" Some of them get very nervous. It changes the nature of relationships.

Editor: If you were to do the evaluation again, what would you do differently?

Donaldson: I like this question, not just for the purposes of this interview, but I love to step back and think about this. The dark side of this project is we sure used a lot of staff hours and killed a forest with our reports. So, a number one concern is writing so many reports. If I were to do it again, I would argue the key is how effectively you communicate evaluation findings with grantees not how many documents you produce. The Foundation really pressed us to produce many reports. I see a lot of RFP's with unnecessary reporting requirements. Too many written reports slow down the communication process. What I would do differently is really look at how to streamline how we communicate, discuss, and act on evaluation findings. In most situations, it seemed like conference calling (vs. written reports, face-to-face meetings, e-mail, or web communications) was most effective in this project. People seemed to prefer to talk for an hour on the phone to the other methods we used.

The other issue that really played out in this project is that the stability of the workforce has dramatically changed. Throughout all of these projects there was significant staff turnover. You spend all this time building a relationship, developing program theories and getting everyone

on board and then a couple of years later it's entirely a new staff. This can undercut the entire process unless you find a way to plan for it and prevent disruptions. The new people don't know about the development of the theory. You have to start over. In your design you have to plan for turnover by putting extra time into educating and orienting new staff and getting their buy-in.

Finally, we feel really fortunate that we were able to get into the project when we did, but I guess you always want more. If we had been there one step earlier, when the committee designed the Initiative, we might have been able to make the Initiative more effective by conducting rigorous needs assessment. The needs were already determined by someone else when we were hired. The passions and commitment were so strong to those programs, As I've said, the Foundation didn't want to reconsider these issues.

Editor: How do you think a new needs assessment might have helped?

Donaldson: I can really only speculate at this point. I would like to believe that systematic needs assessment might have expanded the content of some of the programs and possibly led to the creation of different and more programs. If so, the WHI might have been able to produce even stronger and more long-lasting effects. Of course, evaluators always seem to want more involvement (and budget). In the end, I feel very fortunate that the TCWF had the foresight to involve us early, invest in evaluation focused on continuous improvement, and to create an environment where we could remain objective and present negative findings in a timely manner when appropriate.

Editor's commentary: In 2000, this column took a new approach making a change from interviewing people about award-winning, or exemplary evaluations, to interviewing exemplars, or leaders in evaluation, about their own work. Donaldson's interview and evaluation work for the California Wellness Foundation on their WHI serves that purpose well. Through his discussion of the evaluation we learn more about his approach to theory-driven evaluation, an approach he has written about widely in other settings. Through this interview, we see how he uses program theory to gain a greater understanding of the program and the link between program actions and outcomes. Developing program theory with those connected to the program and the evaluation helps him to begin a vital process of communication with the stakeholders and to gain a greater understanding of the assumptions behind the program. Of equal importance, the theory helps him in determining what to measure in the course of the evaluation. That is, rather than developing evaluation questions targeted to information needs of a specific group or groups or to the stage of the program, Donaldson conducts a comprehensive evaluation of the program theory. He explains that most theory-driven evaluations take this comprehensive approach, examining key elements of process and outcomes. Thus, in this example, theory-driven evaluation helps Donaldson to understand the program, communicate with stakeholders, and select appropriate types and times for data collection. It differs from many theory-driven evaluations in that it does not use multivariate techniques to assess the extent to which the identified mediators in the theory effect the outcome, but instead uses the theory as a tool for planning and communication.

Donaldson describes the program theories he develops as parsimonious and, indeed, they are. In my work, I've come across some program theories that have so many boxes and links that theory is lost. I agree with Donaldson that models which are *too* intricate can fail to communicate the theory for *why* the program activities are supposed to achieve their outcomes. In order for program models to be useful for planning and evaluation, they must be more than flow charts of the progress of clients through a program. They must convey the assumptions

upon which the program is based. I prefer models that are more detailed than Donaldson's written models to convey those assumptions, but Donaldson is using the models to stimulate discussion, questioning, and understanding with his stakeholder groups. As such, these written models may serve well as heuristic devices among the stakeholders to identify key elements of process and outcome.

Like Bickman (Bickman & Fitzpatrick, 2002) who was an early advocate for program theory in evaluation, Donaldson observes that developing program theory, helping program staff and administrators to articulate their assumptions concerning the program and how it works, can reveal problems in conceptualization. We see such problems emerge in the WNJ model with the focus on self-esteem and self-confidence; it is successful with the unemployed in Michigan, but is insufficient for the California populations. While the groups in Fresno and Silicon Valley differed in their educational and employment histories, they both needed certain work skills in addition to self-esteem to help them find jobs. (Note: The effect of theory-driven evaluators in helping program people recognize problems in the conceptualization of programs is reminiscent of an earlier evaluation approach, evaluability assessment. Joseph Wholey developed this approach to determine if programs were ready for evaluation, but practitioners found that the evaluability assessment approach itself was useful for program personnel in identifying problematic links between program actions and outcomes. See Smith, 1989.)

Donaldson's interview also illustrates the struggle evaluators sometimes face between testing an existing model well, primarily for purposes of external validity, versus making adaptations for differing characteristics of a new population (Latinos and other unemployed in California). While CIOF required the development of a new theory because it was a new idea, WNJ had an existing, tested theory. While Donaldson and his evaluation team raised questions about how the theory would work in California, they also were excited by the prospect of testing the Michigan model. We can envision here the pressures from different stakeholder groups. The people who developed the model in Michigan were very interested in the model being implemented in the same way. The Foundation ultimately decided that was the way to go as well. Program people were not involved as early in the process, but one would expect that stakeholders closer to the clients, and the clients themselves, would have argued for adaptations. Donaldson notes that some California people pushed for "a California flavor." In the end, the evaluation helps document that some adaptations are necessary for the program to succeed in this new setting, one very different from Detroit. As such, the results add to knowledge about the model itself, but knowledge the Detroit researchers may not have welcomed. That is, the model doesn't generalize to settings as different as this one and, as Donaldson notes in his comments concerning what he would do differently, local needs assessments can be critical in determining whether a program theory will work in a new setting.

This evaluation was extensive—six years and over 200 reports. Unlike my previous interviews with exemplars where the focus was more purely formative or summative, this evaluation addressed both issues. Donaldson's comments illustrate the different strategies he takes in these approaches. Communication with program staff at the sites is extensive during the formative phase. In fact, Donaldson makes some excellent suggestions for creating a good environment for communication: acknowledge accomplishments as well as challenges (don't feel your job is just finding things to change), let them critique you and your evaluation, when they do criticize you *model* the learning behavior you are asking them to show when you talk about their program. In the end, Donaldson observes that they may have overdone written reports, to meet Foundation requirements, and urges evaluators and funders to make more extensive use of oral communication through conference calls and face-to-face meetings. This type of

communication encourages the give-and-take that is necessary to stimulate use. (I'm reminded of Michael Patton's introducing himself and his approach to users by saying he will do an evaluation, but not a written report. His emphasis prompts the user to recognize that what they're paying for is the advice, not the report.) Yet, when Donaldson moves to the summative phase, his focus is on the Foundation and its board who want to judge the ultimate worth of the Initiative. Other sources receive the reports, but at much later stages.

While Donaldson sees the evaluation as participative, as he frequently notes, the Foundation is the major player and he believes other stakeholders frequently deferred to them. The Foundation staff note "there are few barking dogs when your job is to give away money." The Foundation wanted to be a key player and was. Their role may have been very collaborative, working closely in a partnership fashion with grantees so that all could learn and the gains of the programs would be maximized, as many foundations do today. Nevertheless, it is difficult for an evaluation to achieve a truly participative environment when the stakeholders have such uneven status, as [House and Howe \(1999\)](#) have noted in their model of deliberative democracy. But, clearly, many different stakeholders are involved in the evaluation and benefit from its results.

The focus on continuous improvement builds on the learning organization models that are prominent today, though its distinction from formative evaluation is unclear. This evaluation, and those we have learned about in other interviews, have clearly shown the changes from the early stages of evaluation when the definition was purely judging merit and worth and the focus was seen as primarily summative to today when most evaluations have a formative emphasis and [Mark, Henry, and Julnes \(2000\)](#) view questions of judging merit and worth as only one of the four purposes of evaluation. In fact, Donaldson's evaluation addresses each of those four purposes: judging merit and worth (the board's summative assessment of the Initiative), program and organizational improvement (the continuous improvement component), oversight and compliance (formative evaluation information provided to Foundation staff), and knowledge development (testing of the Michigan model for external validity). As such, Donaldson's evaluation becomes both comprehensive in the scope of questions addressed and in the purposes it served.

Donaldson begins this interview with his desire to show that theory-driven evaluation is not "an ivory tower approach almost like laboratory research," but how it can work in the real world. We see his use of theory working to improve programs and his observations in the field giving him a real sense for what is happening and how programs have to adapt and change. We see his extensive communications with stakeholders and his suggestions for how to improve their receptivity to evaluation results. I think he succeeds in his goal of showing how his approach works in the real world. Where Donaldson and I differ is in whether most evaluations are efficacy evaluations. I think our different perceptions of the evaluation literature may come from reading different literature. At least in my reading, most evaluations are not efficacy evaluations but are evaluations of programs developed by administrators, policy experts, or elected officials who hire the evaluator to address their concerns. These evaluations do occur in the rough and tumble world of real clients, real providers, and the unpredictable changes that occur in program delivery, and, subsequently, in evaluations. Each of the interviews I have conducted with exemplars and with evaluators who have won awards have illustrated evaluations in such settings. Their approaches to adapting to this real world differ, but all are struggling with making choices to conduct informative evaluations in real world settings. Donaldson succeeds in illustrating how his approach to theory-driven evaluation makes that transition and how he uses research and theory development to enhance the evaluation.

REFERENCES

- Bickman, L., & Fitzpatrick, J. L. (2002). Evaluation of the Ft. Bragg and Stark County systems of care for children and adolescents: A dialogue with Len Bickman. *American Journal of Evaluation*, 23 (1), 69–80.
- House, E. R., & Howe, K. R. (1999). *Values in evaluation and social research*. Thousand Oaks, CA: Sage.
- Mark, M. M., Henry, G. T., & Julnes, G. (2000). *Evaluation*. San Francisco, CA: Jossey-Bass.
- Smith, M. F. (1989). *Evaluability assessment: A practical approach*. Boston: Kluwer Academic Publishers.