FOUR

THE EVALUATION OF THE FT. BRAGG AND STARK COUNTY SYSTEMS OF CARE FOR CHILDREN AND ADOLESCENTS

An Interview With Len Bickman

Introduction: Leonard Bickman is Professor of Psychology, Psychiatry, and Public Policy at Peabody College of Vanderbilt University, where he directs The Center for Evaluation and Program Improvement and serves as Associate Dean for Research. He is a coeditor of the *Applied Social Research Methods Series*, the *Handbook of Applied Research Methods*, and the *Handbook of Social Research* and the editor of the journal *Administration and Policy in Mental Health and Mental Health Services Research.* He has published more than 15 books and monographs and more than 190 articles and chapters. Dr. Bickman has received several awards, including the Secretary's Award for Distinguished Service while he was a senior policy advisor at the U.S. Substance Abuse and Mental Health Services Administration and the Sutherland Prize for Research from Vanderbilt University. He is a past president of the American Evaluation Association (AEA) and the Society for the Psychological Study of Social Issues. He is currently Principal Investigator on

several grants from the National Institutes of Health and the Institute of Education Sciences. His research interests include child and adolescent mental health services, Web-based outcomes measurement systems, and the organizational and psychological factors that influence professionals' practice behavior.

Len Bickman's evaluation studies of systems of care for children at Ft. Bragg and in Stark County have received many awards, including the American Evaluation Association's award for the Outstanding Evaluation of 2000 and the American Psychological Association's Award for Distinguished Contributions to Research in Public Policy. Tom Cook has cited the studies as "among the ten or twenty best evaluation studies ever done in any field by anyone." Carol Weiss called the evaluation "one of the landmark studies of the decade," noting not only its excellent research design but also the integrity of the process and the courage in reporting unpopular results. Michael Patton has noted the success of these evaluators in disseminating the findings; engaging their critics in constructive discussion; and ultimately achieving great import, influence, and utilization for the results. We think readers will learn much from Bickman's comments on the factors that influenced this study.

Summary of the Ft. Bragg and Stark County Evaluations

Len Bickman

Vanderbilt University

The Ft. Bragg evaluation describes the implementation, quality, costs, and outcomes of a $94 million demonstration project designed to improve mental health outcomes for children and adolescents who were referred for mental health treatment. The demonstration, designed to test the systems of care continuum as a means for delivering mental health treatment to children and adolescents, provided a full continuum of mental health services, including outpatient therapy, day treatment, in-home counseling, therapeutic foster homes, specialized group homes, 24-hour crisis management services, and acute hospitalization. Services were provided in civilian facilities.

The evaluation was a quasi-experiment with close to 1,000 families. Extensive mental health data were collected on children and their families over seven waves to evaluate the relative effectiveness of the demonstration. A random-effects regression model for longitudinal data was used to analyze ten key outcome variables that were measured seven times. The results revealed that the outcomes in children treated under the systems of care continuum were no better than the outcomes for children in the comparison group. The systems of care demonstration was also more expensive than the comparison, and there was no medical cost offset of the additional costs.

Given the absence of significant effects for the system of care program implemented at Ft. Bragg, another evaluation of the system of care concept was undertaken to learn if the same absence of effects would be noted in another setting. The Stark County evaluation concerned studying an exemplary, mature system of care designed to provide comprehensive mental health services to children and adolescents. It was believed that the system would lead to greater improvement in the functioning and symptoms of clients compared with those receiving care as usual. The project employed random

assignment to conditions, with a five-wave longitudinal design, and included 350 families. While access to care, type of care, and amount of care were better in this system of care than in the Ft. Bragg demonstration, again, there were no differences in outcomes between those receiving the system of care and those receiving care outside the system. In addition, children who did not receive any services, regardless of experimental condition, improved at the same rate as treated children. Consistent with the Fort Bragg results, the effects of the Stark County systems of care were primarily limited to system-level outcomes, but they do not appear to affect mental health outcomes for children and adolescents, such as functioning and symptomatology.

Dialogue With Len Bickman

Jody Fitzpatrick

Fitzpatrick: Your evaluations of the Ft. Bragg and Stark County mental health systems for children and adolescents have received more recognition in the field of evaluation than any study that I can recall in my 25 years of practice. As I note in the introduction, Tom Cook, Carol Weiss, Michael Patton, and others have praised these evaluations. However, I would first like to ask you which elements of the study give you the most pride?

Bickman: I'm most proud of getting the study done. This was the first study in the field and the largest ever done. There were many people who thought we could never get it done because of its size and complexity. Some of my previous grant experiences made me question whether we could successfully complete the project. It is a compliment to the staff who worked with me for us to have received this recognition.

Another thing I am proud of is that we were able to keep the integrity of the design and the measures throughout the study while under considerable political pressure. The studies that had been done in the past had not even looked at clinical outcomes. They had only examined cost and the amount of services. That's what the Department of the Army wanted in the beginning—to just look at the cost to them. The Army people I negotiated the contract with were not used to dealing with research. They wanted the right to approve anything we published, which I refused. They then wanted to be able to comment on anything we published, which I explained was not under their control. Then, they wanted to lower the price because they argued that the publicity surrounding the evaluation would attract better graduate students to the university! It was a battle with the Army throughout the project to maintain the integrity of the design. In the end, however, they were very supportive of us because they now trusted our independence and our integrity. We actually received a rather large contract from them to conduct additional analyses of the data at the termination of the evaluation.

Fitzpatrick: Did you view these studies more as research on psychology and mental health or as public policy studies?

Bickman: I saw it as primarily an evaluation project. It was a policy study, but we embedded several research questions in the evaluation. We have published over 80 articles on this study, many of them in major research journals. We not only had Army funding but a competitive National Institutes of Mental Health (NIMH) grant to extend the project and add additional measures about the families. Every good evaluation has the potential to be good research and to have policy relevance. We thought even if the project [systems of care] doesn't work, we will learn a lot about the mental health of kids. It's a waste not to see the research opportunities as part of an evaluation.

Fitzpatrick: I want to spend some time on the choices you made in designing the studies, but as a political scientist, I'm particularly intrigued with the politics of evaluation and impressed with the attention your study has received by policy makers. Patrick DeLeon, a past President of the American Psychological Association and an administrative assistant to Senator Inouye in Congress, has noted, "Len insisted that all evaluations would represent the most up-to-date level of expertise possible, even when staff within DOD [the Department of Defense] itself strenuously objected. In the end, Len prevailed." Tell us about some of the struggles you endured and what you did that helped you prevail.

Bickman: One important aspect in helping the evaluation prevail was that Lenore Behar, who was the primary contractor, had lobbied heavily for an objective evaluation of the systems of care demonstration. We were a subcontractor, so she shielded us from some of the problems in dealing with the Army directly. But the Army was never supportive of the evaluation until the final report. At first, they thought the system of care project [the demonstration] was not necessary because it was so expensive. So they felt there was no need for the evaluation. The project itself was just too costly. Then, they sent in psychiatrists to visit the demonstration and their reviews were glowing. So now the Army thought the project was good and, thus, no evaluation was needed!

Another problem we faced was developing procedures for identifying and recruiting families. We were new to working with the DOD insurance system and actually new to the whole field of child mental health. We were told that we could use the claims data to locate subjects. However, we were not told that we would not get claims data until 3 to 12 months after the child had the services. We needed to recruit families for the study within 30 days of when they entered services, to collect pre-demonstration data; so obviously, the

claims data were useless for this purpose. We had to develop new ways to recruit subjects. At the demonstration site, this was not a problem since we recruited from just one organization. But in the comparison sites, we had to recruit from every practitioner in each area. The recruitment in the comparison sites required us to identify who provided mental health services and ask them to do the initial recruitment. It also meant that we had to make weekly calls or visits to over 50 providers each week. This work with providers to deal with recruitment of comparison groups for the study not only slowed us down but also cost more than we had budgeted.

A major crisis occurred about halfway through the project. The Army told us that we had enough subjects to complete the project and we should terminate the evaluation. At that point, we had only about 300 cases. We said that they had approved a plan for 1,000 subjects. We went to San Antonio to meet with the Army to discuss this. They hired a consultant to look at the [statistical] power issue. [Bickman's concern was with only 300 cases, they would not have a sufficiently large sample size to have adequate statistical power to analyze the data. Recall that a Type II error, failing to find a significant difference when there really is one, can occur when your sample size is too small. Although Bickman did not know the end results at this stage, his final results showed no significant difference between groups. If the sample size had been 300, he could have been criticized at drawing these conclusions because of low statistical power increasing the probability of a Type II error.] We kept trying to find out who this person was because I know most of the people in this area. We finally got the woman's résumé and she had a Ph.D. in electrical engineering! Our first thought was that they confused statistical power with electrical power! However, it turned out that she also had a master's degree in statistics. Then, we had a site visit during which an Army officer gave me a floppy disk and wanted me to give him the data. I told him we couldn't do that! They weren't happy with me over that. We got into a long battle over the statistical power issue. They hired another consultant who produced a report confirming that we had sufficient statistical power. However, the longitudinal design that we had was now reduced to a one-wave, one-tailed *t* test. We did not consider that analysis to be adequate to answer the questions posed by the evaluation. Our detailed report rebutted their argument. What I was told was that in the end it came down to a general calling and saying we should go ahead with the evaluation as planned. I suspect that there was awareness of the Congressional interest in this demonstration, and it did not pay to alienate

some powerful people in the House of Representatives over this issue. So I concluded the issue was not really statistical power or electrical power but political power!

We didn't have any other major problems with others concerning the evaluation. We kept our interactions with the treatment facility to a minimum. But with the contract business with the Army, almost every year there was a question of whether we would be funded.

Fitzpatrick: Let's come back to the beginning of the study. What prompted it? I know Dr. Lenore Behar, Director of Children's Services in North Carolina's Department of Human Resources, persuaded Congress to fund a study through the Department of the Army to evaluate the "system of care" concept. But whom did you work with in the federal government? What was their interest in the study?

Bickman: The impetus was that there is a movement that stresses such values in service delivery as "culturally fair," "community-based services," and others such as being "strengths based." The child mental health system was basically nonexistent before this movement started. Reforming that system became a major political movement. The system of care model caught everyone's attention as a way to deal with many, if not all, of the ills of mental health services for children and adolescents. Lenore was the driving force to develop a demonstration to test this system of care. She developed the term *continuum of care*, and she wanted to test it to prove it worked. Lenore used all her political influence with some important North Carolina Congresspersons to persuade the Army to fund the demonstration. The Army was reluctant to be involved in such a demonstration, but they do listen to Congressional requests by powerful Congresspersons.

Fitzpatrick: How did you decide to focus on the concept of continuum of care?

Bickman: I did not make a decision on that focus. That's the disadvantage of program evaluation. You're given a program, a focus. But I did push for examining outcomes. Most studies of mental health treatment in the community (as opposed to university laboratories) do not show that they are effective. I think not looking at clinical outcomes for services that are intended to affect those outcomes is poor evaluation practice. I did not know if the continuum of care would affect child and family outcomes, but I did know that was what the program claimed it would accomplish. If you are claiming to do policy-relevant research, you must look at what happens to people. Most contemporary mental health services do not measure outcomes.

The problem for evaluators is that most programs are not well designed. But they are the ones given to evaluators to evaluate. I can give you example after example of programs that are not carefully thought out. We should be teaching logic modeling to everyone who thinks they can design a social or educational program. Maybe then we would have programs that are evaluable. Often the goals of the program are not realistic. We spend more time and money on doing the evaluation than on planning the project. Evaluators need to spend more time on the planning end in helping program people plan programs that make sense.

There are many places where you can get an education in evaluation, but I do not know of many places where you can receive a systematic education on how to develop a program. I think evaluators are trained to find the assumptions underlying a program. I feel there is almost an unconscious conspiracy between providers and policy makers. The public complains about a problem. The policy makers allocate funds to deal with the problem. The providers develop a program to get the money. And then, we get funds to evaluate and find it fails. There's no real change for the people who have to deal with the problem.

Fitzpatrick: In these evaluations, however, you developed a logic model to describe the theory of the program. Do you think using such models helps evaluators "to get underneath it" and to prompt program people to think through their program, or do you mean evaluators should undertake other activities to help in planning?

Bickman: I have no doubt that logic modeling helps both the program and the evaluation. It is a logical approach to examining a program. While it is not a substitute for empirical evaluations, I do not think an evaluation should be attempted until at least a rudimentary logic model is developed.

Fitzpatrick: Coming back to the Ft. Bragg and Start County models, who was involved in helping you develop these models, and how did you go about the development?

Bickman: An article we published in *Evaluation and Program Planning* provides a lot of detail on this issue. (See Bickman, 1996b.). But, in brief, there was not just one model. There was a progression of models. When Lenore goaded me with "Where's our program model?" I said, "We can't have one until I find out what you're doing." It was a group effort. We observed what program developers and managers were doing and what they had written. From that, we developed the first iteration of the model. Then, we would sit down with Lenore and the program people and play that back to them and revise.

Fitzpatrick: How did the program people react to your model?

Bickman: Positively. They loved having a concrete representation of what they were doing. But what's uncomfortable to them is that it lays bare some of their assumptions that they weren't aware of. I consider it critical for every evaluation to develop a model for the program.

Fitzpatrick: I like your model partly because of the level of detail it provides. I see some logic models that have so many boxes and loops that it is impossible to summarize the "logic" of the program. Some of these models simply describe multiple steps of a program and provide little clue to the underlying theory. Other models are too superficial and, hence, fail to provide the evaluator with a framework for effectively measuring critical elements of program implementation. What choices did you and your team struggle with in developing the model?

Bickman: You hit the nail on the head. The idea is to communicate that you're not testing the program; you're testing the theory underlying the program. This approach puts the model at the right level of detail. Educational models are well known for being too detailed with seemingly endless objectives and subobjectives. The key issue a model should convey is "Why should this program work?"

Fitzpatrick: How would you assess the utility of the model now that you have completed the study and had time to reflect? For example, you write that your study did not examine the effectiveness of services delivered or, for example, whether treatment plans improved as a result of the system. Your evaluation focused on the system level and that was the focus of the model. Would you change that today?

Bickman: I foreswore never to do a systems-level evaluation again. To me, it misses the interface between the clients and services. If I knew then what I know now, I would have tried to evaluate the effectiveness of the services as well as the systems.

In addition, I think that evaluations have to change focus from studying only clients to studying providers as well. The difficulty in changing practitioner behavior in services delivery is the major problem with most human services programs. In 1989, I wrote a chapter that we called "Program Personnel: The Missing Ingredient in Describing the Program Environment." We said that we treat programs as if the personnel who deliver them are unimportant since we rarely collect any information about them. I should have paid more attention to what I wrote, since I collected very little information about the clinicians in this study. I'm doing that now. We are doing a study with pediatricians

to see how they deliver services for children with ADHD (attention deficit/hyperactivity disorder). The pediatricians are very ambivalent about the study since they are subjects. I think service providers, especially physicians, may consider themselves above the evaluation. However, since programs are usually designed to change practice, we need to know about the barriers and the incentives to practitioner change, especially the ones that will exist once the program is no longer a demonstration.

Fitzpatrick: The first two steps in the Ft. Bragg study were evaluations of the implementation and quality of the continuum of care. You argued for these intensive evaluations of process, citing Peter Rossi and your own writings on the need to be able to determine if program success or failure is due to a faulty or weak implementation of the theory or to the program theory itself. Tell us a bit more about why you thought examining implementation was important.

Bickman: The worst outcome that an evaluator can obtain is finding that the treatment produced no effect and that the project was not implemented. That means you wasted your time. I wanted to know that the program was delivered with fidelity—to know that if the program failed to produce the desired effects, it was because it was a theory failure. [If you know the program was delivered as the theory indicated and, then, find that the desired outcomes are not achieved, you have a "theory failure." You have proved that the program theory does not work. However, if you do not know whether the program was implemented as planned, and you find that the desired outcomes are not achieved, you do not know whether you have really tested the theory and it does not work *or* whether the theory, in fact, was not tested because the program was not delivered as the theory would require.]

Fitzpatrick: I like your distinction between studying program implementation and program quality. Often, those are combined. Why did you decide to separate the two issues?

Bickman: There is a distinction between measuring quality and measuring implementation. I believe outcomes are easy to measure. Implementation is relatively easy to measure. Quality is the hardest element to measure. That's because measuring quality involves making a judgment about worth. When I use the word *quality*, I mean a process related to the desired outcomes. Most human services don't have measures of quality. Even laboratory studies have difficulty getting reliable measures of the quality of treatment. But my guess is that it is the quality of the services that accounts for the variance in program success. How much quality is enough? We have no idea of the levels necessary

to produce effects. Evaluators are missing the whole quality bandwagon. We're in some danger of having evaluation taken over by quality managers. I edited a special issue of *Evaluation Review* on quality and mental health in 1997 with the idea of raising evaluators' awareness of the relationship between quality and program evaluation. (See Bickman & Salzer, 1997.)

Fitzpatrick: How did you measure quality in the Ft. Bragg and Stark County evaluations?

Bickman: We looked at the components that we thought were critical—case management and intake. These components were the "glue" that held the program together. On intake, we asked practitioners in the field, "Was the intake conducted centrally helpful? Did it reduce the number of sessions you needed for assessment?" Ultimately, there is no gold standard on treatment plans. To study case management, we had case managers keep logs of their activities, and we analyzed charts, conducted interviews, and reviewed documents. We used a scale that measured program philosophy. We also interviewed parents and did a network analysis. In addition, we developed a "case management evaluation data checklist" that was our measure of quality, based on concept mapping and document reviews. The checklist included such items as parent involvement in treatment planning, client monitoring and follow-up, and linkage and coordination activities. Our evaluation of the quality of case management involved comparing the checklist to the evidence we had collected from multiple sources. Details about this procedure were published in a special issue in *Evaluation and Program Planning* on evaluation methodology and mental health services.

Fitzpatrick: In spite of years of discussion of the need to examine implementation and describe program operations, many evaluators view implementation studies as less prestigious. Similarly, today many organizations and government entities such as schools, private foundations, and the United Way, are very outcome focused and tend to neglect process. Are implementation studies always important? Why do you think evaluators and funding sources sometimes scorn them?

Bickman: I think the bad reaction to implementation is historical. Implementation or input evaluations were the only evaluations conducted by many community agencies. It was rare to find an outcome evaluation in these settings. In addition, we do not have theories of implementation, so it is difficult to study implementation. If implementation studies are not theory based and are just operationally detailed, describing what is happening, implementation

studies are boring. However, in a comprehensive evaluation, we need to know how the program was implemented to learn why it was or was not successful. It is clear we need both in a comprehensive evaluation. But it does add expense to the evaluation.

Fitzpatrick: You note that your implementation study *was* theory driven. Can you tell us more about how theory guided this phase of your study?

Bickman: We developed the implementation plan based on the theory of the program. The program theory guided us not only in measurement issues but also in what aspects of the program it was necessary for us to study. It is impossible to adequately study a whole program, so we had to select those aspects of the program that theory identified as critical to the success of the program.

Fitzpatrick: Ultimately, you concluded that the program did successfully implement the model and that the services were of sufficient quality. Yet, judgments are involved in drawing these conclusions. No program is implemented with 100% fidelity to the model. Similarly, no program is consistently delivered with top-level quality. How did you reach your judgments on fidelity of delivery and level of quality?

Bickman: I think you have identified a significant weakness in this area. Given that program developers usually have no theory related to the quality or amount of services necessary to produce an effect, it becomes the evaluator's judgment of when implementation is sufficient. Moreover, our judgment could be biased since it is in our self-interest to declare implementation a success. There is not much to be gained by evaluating a program that was not implemented. I was always aware of this problem, but I was never challenged on this aspect of the evaluation. You try to describe the evidence as best as you can, and then you make a judgment. Is it half-empty or half-full? It is a value judgment. The fact was that the clients had all these extra services—case management, intermediate services, and so on. This is what the program developers planned.

There were some challenges to the results we reported. Some said that the Ft. Bragg program was not sufficiently mature. But the program had almost a year and a half of start-up time before we started to collect baseline data. Also, if maturity was an issue, we should have seen improvement in the program outcomes over the three years we collected data. Instead, we did not see any improvement in effectiveness over time.

Fitzpatrick: But did the clients get quality case management? Did they get the services case management recommended?

Bickman: The case managers felt they were doing their job, but the amount of contact with clients was amazingly low. I think this is true of most case management. However, we had claims data to show they got the services that one was supposed to get in a continuum of care. What we could not tell, and I do not believe that it is even possible now, is whether the clinical services were of sufficient quality.

Fitzpatrick: Let me address one specific issue on fidelity and quality. Your studies were evaluations of the system of a continuum of care. But one element of that care is therapy, and since the thrust of the continuum is mental health—your outcome measures focus on changes in psychopathology—one would assume therapy would play an important role. But, in fact, relatively few children in the Stark County study received therapy. By parent report, only 14% of children in the Stark County system of care group received individual counseling during the first six months, and the proportion shrunk to half that in the second six months. Is that amount of therapy sufficiently intense to match the model?

Bickman: The label *counseling* is only one category of therapy. The services they delivered included clinical case management and intensive home-based services. That is where most of the therapy was supposed to be delivered.

Fitzpatrick: Do you find that some audiences misinterpret the results to mean that therapy is not effective?

Bickman: You cannot determine the effectiveness of therapy directly from this study. However, one of my explanations for why the system was not more effective was that the treatment wasn't effective. However, in order to support this explanation, we analyzed other aspects of the data. If the treatments were effective, we would expect there to be a dose-response relationship. That is, the more treatment received, the better the client should be. We examined the dose response in both the Ft. Bragg and Stark County study in several ways, and the conclusion was always the same. The amount of treatment did not matter. Clients who received more treatment did not have more positive outcomes.

Moreover, the reviews of other community-based treatments (i.e., not laboratory studies of therapy) concluded that these treatments show no effect. From these studies, I concluded there was not evidence that treatment in the real world was effective for children and adolescents. Notice that I did not say that therapy was ineffective, just that we did not have evidence to support its effectiveness. But the issue of effectiveness begs the question "What is treatment?" It is whatever clinicians do. Let me tell you how we describe most of

these services. They are in-home services, day treatment, hospitalization, private office visits, and so on. Would you buy a car if I just told you where it was located? We're describing services by location. We're assuming that they are different because they occur in different settings, but that is an assumption.

But I do worry that what I am saying demoralizes clinicians. Is there anything beneficial about me saying all this? Clinicians are working in a very difficult, emotional area. They do not work in this area for the big money. We also suspect that to be a good clinician, you have to believe in your efficacy. I don't know whether my criticism is constructive for these people. But I strongly believe in what I am doing and that our first concern is to help identify effective services for these children. If the services are not effective, then it is the evaluator's responsibility to say so.

Fitzpatrick: Your outcome studies, of course, have probably received the most attention. In both the Ft. Bragg evaluation and the Stark County evaluation, you found that receiving a continuum of care made no difference in children's symptoms or functioning. Of course, when no difference is found, critics attempt to identify problems in data collection or the design that may have resulted in a failure to identify real changes, a Type II error. A major strength of your design and methods was its resilience in dealing with such criticisms. You collected extensive data on many different constructs from many different sources. You conducted power analyses and subgroup analyses to test various hypotheses and explore extensively for possible effects. For example, you examine whether the continuum of care was more effective for children with different demographics and different diagnoses. You also explore whether differences exist between children who actually received treatment. Tell us a bit about how you planned your design and analyses. Did you identify all possible analyses in the design phase? Do you think it is better to explore the data thoroughly after they are collected to consider subgroup tests and the like? Did you seek input from others on exploratory analyses?

Bickman: There are certain main analyses that we had planned. We also had to learn how to deal with the thousands of variables. We identified 10 key outcome variables; two of them were individualized and represented that child's progress in his or her specific problem area. We did some subgroup analyses, but we didn't have a theory about who should benefit most from the continuum, so we didn't go on fishing expeditions. I am concerned that such analyses creep into our literature and result in the inconsistent results we often see. I think it is important for investigators to report on all the analyses they

conducted and explain when they have a large amount of data collected but only a few results reported. I assume they did a lot of other tests that were not significant and that they are primarily reporting only the significant ones. I consider that kind of publishing as biased. Not only does it exploit chance, but it also is not driven by any theory or concept of the program.

Fitzpatrick: While your study is too extensive for me to ask every question I would like, let us explore a few issues. As I mentioned above, you tested for differences in children not simply based on which group they were in but on whether, in fact, they actually received treatment or not. What treatment variables did you examine when comparing outcomes for children receiving treatment with those who did not? For those who did receive treatment, what were the typical treatments, and how much did they receive?

Bickman: We analyzed treatment by amount, such as sessions or days or dollar amount, and what we called negligible treatment, such as only one or two outpatient visits. But all the treatment measures are correlated. A session is a session is a session. We have published several studies on patterns of treatment and improvement as well as dose response based on these studies. A major weakness of our approach was that we had no measure of the *quality* of the treatment and how it varied. In a recent study, we had children tell what they did during treatment. We had nothing like that in these earlier studies. I was not aware that quality was such an important issue until Ft. Bragg.

Fitzpatrick: Your study at both sites used many different measures. Sources included the children themselves and primary caregivers, generally parents. The Ft. Bragg study also collected data from the mental health provider and teachers. Baseline interviews were conducted in the home, but other data collection methods included computer-assisted telephone interviews and mailed questionnaires. Your measures generally focused on the constructs of psychopathology and family functioning and were standardized measures used in mental health research. Your large sample sizes at both sites gave you sufficient power to preclude a Type II error but would have made the cost of collecting qualitative data from all families prohibitive. What issues did you struggle with in making choices about constructs, sources, and methods in data collection?

Bickman: Well, we had no choice in the design. That was given to us. We did have a lot of choices in instrument collection. This was a whole new area for me. We did a lot of research on existing measures. We selected what we thought were the best in the field. We found out that some of the best really were not too good.

Fitzpatrick: Why do you lean more toward existing measures?

Bickman: Because I know what it takes to develop new measures. However, we did have to develop a measure of functioning because none existed at that time. Since then we have developed a whole new system of measurement that better fits the real-world environment than some of the research instruments we used.

Fitzpatrick: Another major strength of these evaluations was your selection of a second site, Stark County, to see if the results would be replicated with a different population and system of care. Researchers and evaluators rarely replicate their own work so systematically, nor do they take such care to select a site whose characteristics strengthen the external validity of the findings. Stark County's clients were nonmilitary, lower-income, publicly funded youth. The system of care was a full, mature system more typical of systems of reform in many communities than the demonstration in Ft. Bragg. So the Stark County results, which did replicate those of Ft. Bragg, added greatly to the external validity of your results. However, replicating a study brings the opportunity for some change as well, though, of course, too much change can threaten the replication. What changes did you consider in planning the design and data collection in Stark County?

Bickman: Basically, we changed what we had been criticized for not doing in Ft. Bragg. The Stark County study was all NIMH funding. They had turned me down twice for this study. The first study included seven cities that were in the Robert Wood Johnson Foundation study of systems of care demonstration. This evaluation did not plan to collect child outcomes, and I thought it was a natural to try to work with them and build in an outcome design and outcome measures. The reviewers said it was ridiculous to study seven cities. So we resubmitted it with one. Then, that city withdrew a month before the study was to start. I called NIMH and they said, "Go find a site." So I picked one that would meet my need for a random assignment design. I also picked a site that was proud of their program and was nationally recognized as a leader in the field.

Fitzpatrick: While your external validity is greater than most evaluations because of your use of two contrasting sites, what limitations do you see in generalizing the results from these two sites? Would you say you had pretty good external validity?

Bickman: The way the ideologically committed critics dealt with the results and the methodology of the study was to say, "We think you did an excellent evaluation, but we don't do that program anymore." That happened

with DARE (Drug Abuse Resistance Education). Evaluations found that DARE did not decrease alcohol and drug abuse, so the DARE people argued that the evaluations of the earlier program were irrelevant because the program had changed.

I am not as concerned about external validity as I am with construct validity. I'm not really testing the Ft. Bragg demonstration. I'm testing the concept of a continuum of care. That is why the implementation analysis is so important. It helps me decide if the program is a good representation of the construct. I did not see the generalizability limited by the population of children served or the region of the country. What I was looking for were excellent examples of the theory represented by these projects. If we could not obtain the hoped for outcomes with these excellent and well-funded sites, then it was unlikely that sites with fewer resources would be successful.

Fitzpatrick: Your ultimate interpretation of your results concerns the quality of care delivered. That is, you suggest that rather than focusing on system-level issues, which may be too removed from patients to impact outcomes, policy makers should focus on improving the training of treatment providers and, ultimately, improving the quality of services delivered. How did you reach that conclusion since your study did not directly address the quality of services provided?

Bickman: There were only a few alternative explanations for the results we obtained. First, was the evaluation critically flawed? No one has been able to demonstrate that we did a flawed evaluation. Second, was the demonstration well implemented? We presented evidence that implementation was fine. If the evaluation was good and the program was well implemented, then what do we have left? The theory underlying the continuum model is wrong. There are several factors within the theory that could be wrong. For example, the theory assumes that clinicians are able to assess children's needs and match them to the appropriate services. We showed in a separate study that this did not seem likely. But the key theoretical assumption was that the services were effective. Our dose-response studies, plus the meta-analyses of children's therapy delivered in the community, convinced me that we had to take a better look at the quality of treatment. This conclusion has led me to three areas for my future work in this area.

First, we need to be able to measure outcomes in the real world. My colleague Ann Doucette and I have developed a new measurement system that we believe will allow community service providers to learn what is effective.

Second, we need to look at the process of care to determine which mediators are important in affecting outcomes. We have seen that therapeutic alliance, the relationships between the provider and the client, can be very important. Third, I have started some studies that examine how to change practitioner behavior, so that when we *do* have something that is effective, it can be adopted in the community.

Fitzpatrick: Have your evaluations brought about the changes you desired? In particular, have they helped to change the focus from system-level variables to the effectiveness of the treatment itself?

Bickman: Yes and no. Critics are now talking differently. They are saying we need to consider treatment and services as well as the system. It is not a brilliant insight, but it is a big change. On the other hand, the Center for Mental Health Services is still funding projects like Ft. Bragg to the tune of $80 million a year. To be a good evaluator, you need to be skeptical. I characterize myself as a skeptical optimist, even when it comes to my own work. My skepticism helps me to be a good evaluator, and my optimism motivates me to stay in this field. However, providers should also maintain some degree of skepticism about what they do. I think the continuous-quality-improvement approach has a lot of appeal, but if the providers are sure that they are already delivering the best services, there is no need for continuous data collection or evaluation.

Fitzpatrick: Your results have implications for so many different organizations and systems: any system using a continuum of care or emphasizing a full system of care; schools, and licensing organizations that educate, train, and oversee therapists of different types; managers and supervisors in practice settings; managed care; researchers; and others. Who do you see as your primary audiences? How do you attempt to reach them?

Bickman: I have a lot of audiences—basic researchers in psychopathology, clinicians, and policy people. I've tried reaching them through writing. I don't know how else to do it. Some people read. Other people hear about these studies from others. My latest contract is with a state that is under court order to provide better mental health services for their children. They have been at it for about seven years and have spent about a billion dollars. They just contacted me because they read in their newspapers about the studies I did, because of a legal problem someone associated with the Ft. Bragg study is having. I offered them an alternative way to deal with their problem. However, being the skeptical evaluator, I didn't promise any quick solution or even lots

of confidence that what I am proposing will work. I tell them that whatever they do, they should also do an evaluation.

Fitzpatrick: In speaking to mental health researchers, you note the need for them to do more testing in real-world environments so that the changes made in the field can be more readily research based. You call for their help in developing research-validated practice standards. Do you think your studies and the attention they have received have prompted some change in the norms and practice of these researchers?

Bickman: It has already changed radically in mental health funding. NIMH is promoting building bridges with the clinical world very hard. They are seeing more clearly that their responsibility is to improve the mental health of children and that this was not occurring through just the publication of research. Getting into the real world is now a priority. I would like to think that our work helped push that along.

Fitzpatrick: Finally, I like to close with asking the exemplars I interview what they have learned from the evaluation they have done. What might you do differently if you were beginning on these evaluations today?

Bickman: There are things that I would have done, but I did not have the power to do it then. I would have liked to collect data more frequently, not every six months but at least monthly. I would have liked to have more information on the quality of treatment. I'm doing that now.

I think about why I do evaluations. I used to tell students that it was purely hedonistic. I enjoyed it. But that has changed. I think we *can* help kids get better through our work. Often evaluators don't get that opportunity. I feel grateful that I had the opportunity to continue to work in the same field for over a decade. I am hoping that our work here can improve outcomes for children who have mental health problems.

Fitzpatrick's Commentary

Bickman's evaluations of the Ft. Bragg and Stark County systems of care for children have received many well-deserved accolades in the communities of evaluation and mental health. Bickman and others have written in more detail about the measures, design, and results of these evaluations in other venues. (See References.) For this interview, I focused my questions to help us learn more about the challenges that Bickman and his colleagues faced in planning and implementing these evaluations and his reflections on the evaluations.

As in each interview I have conducted, we learn much about the exemplar's approach to evaluation by the choices they make. Bickman's studies are summative. The audience is broad: policy makers and thinkers in the field of children's mental health. His purpose is not formative; his audience is not program managers or staff. And, as such, he need not involve them intensively as users, though he certainly involves them in the development of the logic model. In such evaluations, the distinction between program evaluation and policy analysis is somewhat blurred, but we needn't quibble over terms. Instead, we should simply note the different focus.

Bickman's methods come from the quantitative tradition, but they match the purposes of the study. To clearly establish outcomes for summative decisions, they needed large samples, experimental and quasi-experimental designs, and multiple measures from different sources. But as I note below, he also takes a useful focus on process, examining implementation and quality to enhance his interpretation of outcomes. He selects measures primarily from the research literature, but having learned the strengths and weaknesses of these, he has now moved on to developing measures for use by community practitioners. Though research based, his goals concern learning about what is actually going on in practice and include a genuine commitment to improving mental health services in the field. These evaluations and his integrity and ability in research and evaluation have permitted him to argue effectively for major changes in the way people study children's mental health. As such, his work and writings make a bridge between research and evaluation.

While probably all audiences saw the purpose as summative, Bickman successfully argued that studies of children's mental health must, of course, examine the outcomes on children and their families rather than the process and cost measures originally considered by the funder. This adjustment of focus, in and of itself, is a commendable outcome of his evaluation. But he does not neglect process. As he advocates in his writing, he makes use of logic models to define program theory and identify critical elements of the process to monitor. He then attempts to measure not just the implementation of critical program elements but the quality of those elements as well. The delineation of implementation and quality is an important element because quality is often neglected in process studies. Yet, as Bickman argues, in all likelihood, quality is the most important element in ensuring successful outcomes.

In addition to the many oft-cited qualities of Bickman's work on these evaluations, one of the elements I admired most was his conscious effort to

build and learn from previous evaluations and to, then, take the next step in his subsequent studies in order to learn how to improve mental health outcomes for children. This effort is seen in his selection of Stark County for the replication and in his current work. Having focused on system-level variables in the Ft. Bragg and Stark County evaluations and found no effect, he begins to consider what other elements could or must have an effect to bring about change. In current studies, he is examining providers' behavior and becoming curious about the incentives and barriers to changing their behaviors. Having used primarily existing research measures for the Ft. Bragg and Stark County evaluations, Bickman is now developing new measures that can be used by community service providers to provide more effective, practical, and immediate feedback on outcomes. In other words, he doesn't just find problems, he goes on to explore solutions.

Bickman correctly notes that we evaluators must evaluate the programs we are given. But his frustration with ill-conceived programs compels him to argue for expanding evaluators' roles in planning. Like Reichardt (1994), Patton (1997), Preskill and Torres (1999), and others, Bickman sees that our future may be at the beginning of programs rather than the end.

DISCUSSION QUESTIONS

1. Is Bickman assessing the merit and worth of the system of care? How does he do so? What are the strengths and weaknesses of his methodological choices?

2. Bickman's evaluation randomly assigns children and adolescents to different levels of mental health treatment. The developers of the system of care believe that it will result in better mental health services to children. Is it ethical for him to randomly assign children to a treatment that is not thought to be the best (the old treatment)? Why or why not?

3. Bickman argues with the Army over what to measure, the size of the sample, and other methodological issues. Shouldn't he be respecting stakeholder wishes on these issues rather than arguing for his own preferences? What would you do?

4. Would you characterize Bickman's study as participatory? Why or why not? Do you agree with the choices he made regarding the level and depth of stakeholder participation desired for the study? Why or why not?

5. Bickman argues that many programs are poorly designed. To remedy that problem, he says, program managers and staff and others who develop programs need more skills in developing logic models. He also indicates that evaluators should become more involved in program development because they have those skills and can help uncover the underlying assumptions of a program. Think about a program that you know well. Is it well designed? What is its logic model? Can you think of a program or policy that failed because its logic model was not carefully thought through?

Do you think evaluators should be involved more heavily at the stage of program development? What should be their role? What expertise do they bring to the process? What expertise or other attributes do they lack?

6. How does Bickman disseminate his results? Contrast his dissemination process with Riccio's dissemination. How do they differ? Are there elements about the context of each of their evaluations that make their different choices in dissemination appropriate? Or would you have had one of them use more of the ideas of the other in disseminating results? Why or why not?

7. Bickman believes that it is advantageous for evaluators and program providers to be skeptical. He describes himself as a skeptical optimist. To what extent are you a skeptic? Do you agree that evaluators need to be skeptics? Why or why not? Should program providers be skeptics? Why or why not?

FURTHER READING

An article to read that summarizes the results of the evaluation: Bickman, L. (1996a). A continuum of care: More is not always better. *American Psychologist, 51,* 689–701.

Bickman, L. (1996b). The application of program theory to a managed mental health care evaluation. *Evaluation and Program Planning, 19*(2), 111–119.

Bickman, L. (1997). Resolving issues raised by the Ft. Bragg findings: New directions for mental health services research. *American Psychologist, 52*, 562–565.

Bickman, L. (2000). Improving children's mental health: How no effects can affect policy. *Emotional & Behavioral Disorders in Youth, 3*, 21–23.

Bickman, L., & Salzer, M. S. (1997). Measuring quality in mental health services. *Evaluation Review, 21(*3), 285–291.

REFERENCES

Bickman, L. (2002). The death of treatment as usual: An excellent first step on a long road. *Clinical Psychology: Science and Practice, 9*(2), 195–199.

Bickman, L., Noser, K., & Summerfelt, W. T. (1999). Long-term effects of a system of care on children and adolescents. *Journal of Behavioral Health Services & Research, 26,* 185–202.

Bickman, L., Sumerfelt, W. T., & Noser, K. (1997). Comparative outcomes of emotionally disturbed children and adolescents in a system of services and usual care. *Psychiatric Services, 48,* 1543–1548.

Bryant, D., & Bickman, L. (1996). Methodology for evaluating mental health case management. *Evaluation and Program Planning, 19,* 121–129.

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text.* Thousand Oaks, CA: Sage.

Preskill, H., & Torres, R. T. (1999). *Evaluative inquiry for learning in organizations.* Thousand Oaks, CA: Sage.

Reichardt, C. S. (1994). Summative evaluation, formative evaluation, and tactical research. *Evaluation Practice, 15,* 275–282.