In the October 1995 issue of EP, *we published the first in what we hope will be a continuing series of exemplary evaluation studies. As noted in that issue, there are alternative forms and focuses that this section could take, and Paul Wortman, contributing editor for this section, and I continue to explore possible directions that may be fruitful. Last time we featured Wortman's description of the Perry Preschool Project Evaluation, which focused much more on its results than on its methods. Here we publish a summary by Len Bickman of his widely heralded Fort Bragg Evaluation, with a brief description of the study's methods, findings, and "lessons learned" by the evaluation. We have also invited a response from John Landsverk. Please let us know what focus you think would be most useful for this section in future issues of* EP.

Implications for Evaluators from the Fort Bragg Evaluation[1]

LEONARD BICKMAN

ABSTRACT

This paper briefly describes an $80 million project designed to evaluate whether a coordinated continuum of mental health and substance abuse services for children and adolescents is more cost effective than services delivered in the more typical fragmented system. The evaluation examined the implementation and quality of the demonstration and the impact on service utilization, client satisfaction, clinical outcomes and cost. Several competing explanations for the results are described and implications for evaluators are discussed. The paper concludes with a discussion of the major evaluation lessons learned in this study.

Leonard Bickman • Center for Mental Health Policy, Vanderbilt University, 1207 18th Avenue South, Nashville TN 37212
E-Mail: Bickman@UANSV5.VANDERBILT.EDU.

Evaluation Practice, Vol. 17, No. 1, 1996, pp. 57-74.
ISSN: 0886-1633

INTRODUCTION

There were no previous evaluations of the cost, quality and mental health outcomes for any of the varieties of mental health systems for children and adolescents prior to the Fort Bragg Evaluation. Stroul and Friedman (1986) outlined the basic principles and components of an accessible, least restrictive, and cost-effective system of care for children. A key aspect of such a system is the full range of mental health services. This range of services has been termed a *continuum of care* (Behar, 1985; Stroul & Friedman, 1986) and includes residential, intermediate, and nonresidential services. The continuum of care is designed to deliver coordinated services on an individualized basis, using case management and interdisciplinary treatment teams to integrate and ease transition between services. The Demonstration and Evaluation[2] were funded because of concerns about large expenditures for mental health services for military dependents, especially for children and adolescents. This article briefly summarizes the results from the Fort Bragg study, and then primarily focuses on the implications of the findings for program evaluators.[3]

The Fort Bragg Demonstration Project

On June 1, 1990, after a ten-month start-up period, mental health and substance abuse services were offered to more than 42,000 child and adolescent dependents of military personnel in the Fort Bragg catchment area in Fayetteville, North Carolina. The Demonstration required that families seeking services for their children use the Demonstration's clinical services, which were free, or seek and pay for services on their own. The Rumbaugh Clinic, an off-post civilian facility at Fort Bragg, was the coordinator, fiscal intermediary, quality assurance monitor, and umbrella organization for all mental health and substance abuse services for CHAMPUS[4] beneficiaries.

Services within the continuum and across other agencies were linked by case managers and interdisciplinary treatment teams led by a doctoral-level staff person. Additional services, such as transportation and other wraparound services, also were provided.

The American Psychological Association's section on Child Clinical Psychology and the Division of Child, Youth and Family Services joint task force considered the Demonstration a model program (Roberts, 1994). The Demonstration was recognized for its continuum of care in a single location and its close ties to other services and agencies.

The Demonstration based its cost control approach not on limiting services or cost per child but on providing a wide range of services designed to permit placement in less restrictive and hypothetically less expensive services. It was assumed that this rational/clinical approach to managing benefits would more effectively control costs than placing fixed and often arbitrary limits on care. The next section describes the standard care children received outside the Fort Bragg catchment area in the comparison site.

Traditional CHAMPUS Services

In fiscal year 1991, CHAMPUS allowed 45 days of inpatient or hospital care, 150 days of Residential Treatment Center (RTC) care, and 23 outpatient visits per year. Unlike many other insurers, CHAMPUS imposes no lifetime benefit or dollar limit (Baine, 1992). Health Management Systems Inc. (HMS) was responsible for providing utilization management for all inpatient and RTC stays and for outpatient visits exceeding 23 visits per year. CHAMPUS coverage is generous. Nevertheless, it does not contain any of the key features of a continuum of care, i.e., no systematic centralized intake, case management and treatment teams, or intermediate level services.

Program Theory

I have defined program theory as "a plausible and sensible model of how a program is supposed to work" (Bickman, 1987). The present theory tested the hypothesis that an integrated and coordinated continuum of care is more cost-effective than a fragmented service system with a limited variety of services.

Figure 1 shows the logic model or program theory of the Demonstration developed by the evaluators. The eligible population can receive mental health services only through a single point of entry, the Demonstration. This feature allows the Demonstration to control access to services. Moreover, it should make it easy for potential clients to know where mental health services can be obtained. Strict standards of timeliness are not part of "services as usual" provided under CHAMPUS.

As indicated in the lower left-hand column of Figure 1, providers and potential clients were informed about the new system of care. These activities were hypothesized to produce a *proximal outcome* of increased access to services. This increased access should have resulted in two *intermediate outcomes:* (a) an increase in the number of clients served by reducing barriers to obtaining services; and (b) an increase in client satisfaction with the intake process through a reduced financial burden and increased ease of access.

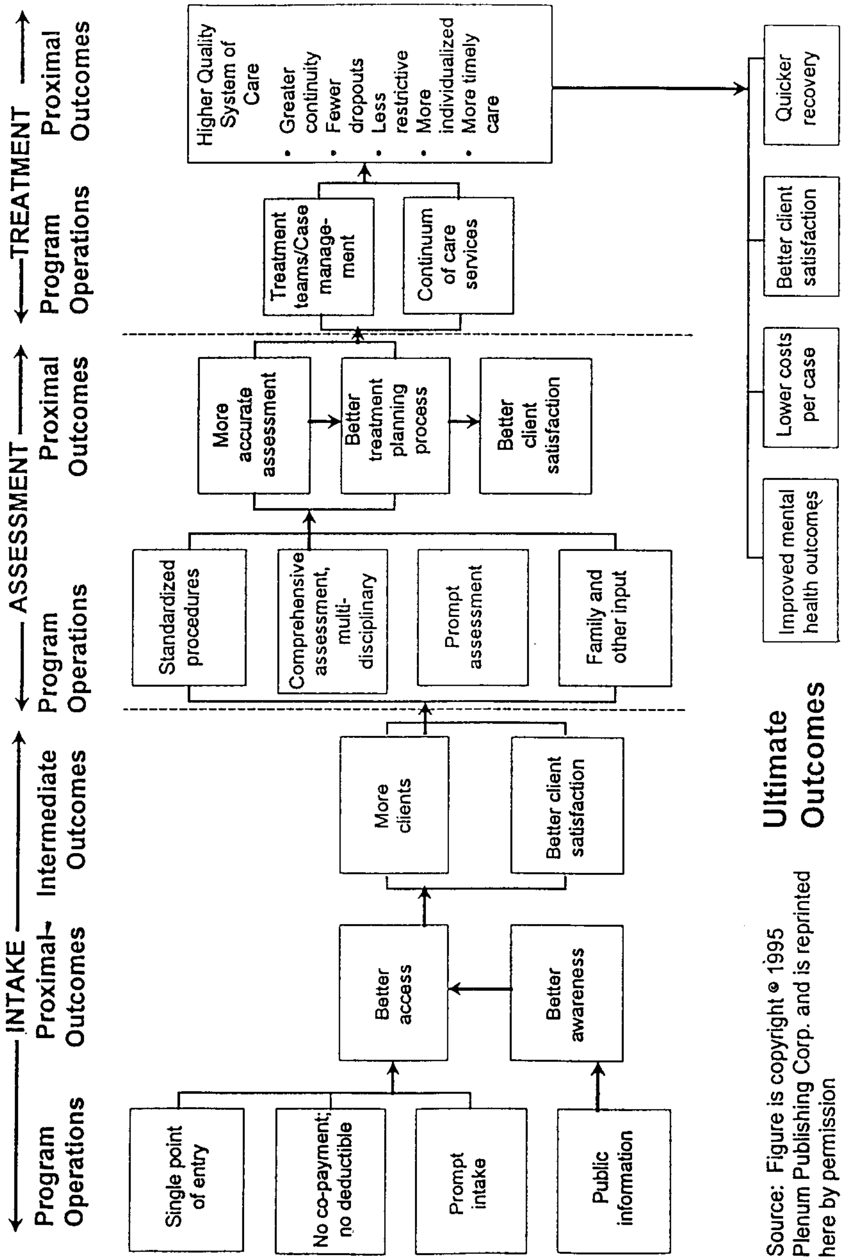
The second major section of Figure 1 illustrates *program operations* during the *assessment* process. All families had a comprehensive evaluation and review by an interdisciplinary treatment team. The Demonstration contract set standards for assessment within three weeks including obtaining input from family and other relevant professionals. These program activities were hypothesized to lead to three *proximal outcomes:* (a) a more accurate assessment of the problem, leading to; (b) a better treatment plan; and (c) an increase in client satisfaction with the assessment process.

The *actual services* provided by the Demonstration should lead to the *proximal outcome* of a higher quality system of care that provided a better match between the services and the needs of the child and family. Quality of system is defined by the terms in that box. The Demonstration should lead to the *ultimate outcomes* of better mental health, lower costs per case, quicker recovery, and more client satisfaction.

METHODS AND RESULTS

Quality and Implementation

The Evaluation consisted of four substudies: (a) The Implementation Study; (b) The Quality Study; (c) The Mental Health Outcome Study; and (d) The Cost/Utilization Study. The Implementation and Quality Studies documented the Demonstration's activities in order to assess the degree to which the Demonstration carried out a high-quality system of care that was faithful to the program model. The Evaluation needed to be confident that the success or failure of the Demonstration could be attributed to the theory and



Source: Figure is copyright © 1995 Plenum Publishing Corp. and is reprinted here by permission.

Figure 1. Fort Bragg Child and Adolescent Demonstration: Program Theory

not to a faulty implementation of the theory. These studies were based on the theory driven and component approaches to program evaluation (Bickman, 1985, 1987; Chen, 1990; Chen & Rossi, 1987) and used a case study approach (Yin, 1986; 1993) to describe the structure and processes of the Demonstration. The case study used multiple methods and multiple sources of information and included a network analysis.

The Quality Study assessed the quality of the intake assessment and case management, and one key administrative function, the Demonstration's Quality Improvement (QI) program. The quality study was an affordable substitute approach to gauge the effectiveness of the individual components of care such as psychotherapy and hospitalization. These studies are described in Bickman, Bryant, and Summerfelt (1993), Bickman, Summerfelt, and Bryant (1996) and Heflinger (1993; 1996).

The Implementation study concluded that the Demonstration successfully executed a coordinated, individualized, community-based and family-focused system of care. Based on a network analysis, services were considered to be better coordinated at Fort Bragg than the services at the comparison site. The Quality Study showed that the services provided at the Demonstration were of sufficient quality. The Demonstration maintained a quality improvement (QI) program that supported quality clinical activities. The two most significant non-therapeutic activities, case management and intake/assessment central to a continuum of care, were also judged to be of high quality.

Mental Health Outcomes

The Mental Health Outcome Study collected data from a sample of 984 children (574 at the Demonstration; 410 at the Comparison), between the ages of five and seventeen, and their families. The Comparison group consisted of children receiving traditional CHAMPUS services at two comparable army posts. Families were interviewed at the intake to services and six, 12 months and 18 months later. Parents and children provided clinical information through in-person structured clinical interviews and self-report questionnaires. Domains included level of functioning, competencies, pathologies, family environment, and client satisfaction with services. The informants included the children in treatment, their primary caretaker, the mental health provider, teachers, and trained interviewers. The specific instrumentation is fully described in Bickman *et al.* (1995).

Since the study was a quasi-experiment, sample comparability was addressed after recruitment and at each wave of data collection. Analyses of demographic and outcome data showed that the samples were comparable at each wave of data collection. For example, of the 103 measures of mental health status of the children at intake, 14 suggested that the two sites differed significantly: nine showed that children at the Comparison site exhibited greater symptomatology than those at the Demonstration site; five suggested the children at the Demonstration were more severely impaired. None of the effect sizes associated with the mental health status of children was greater than .25 standard deviations.

Two analytic techniques were used to evaluate the effect of the Demonstration on clinical outcomes. Analysis of covariance, with Wave 1 psychopathology data serving as the covariate, was used for Waves 2 and 3; for Waves 4 and beyond, a longitudinal random regression was used. At both sites most of the outcome measures showed a large drop in pathology from Wave 1 to Wave 2 (six months) followed by a smaller drop from Wave 2 to Wave 3. Overall, at Wave 2 there were three significant site differences among the 12 measures, two favoring the Comparison. All three differences had small effect sizes (less than .25 standard deviations). At Wave 3 (12 months), two significant results favored the Demonstration. At Wave 4 (18 months) there was one significant difference in outcomes favoring the Demonstration. The random coefficients analysis showed a slightly better slope of improvement for interview ratings of youth impairment at the Demonstration.

The preceding results can be summarized as six small differences among 36 sensitive tests on a large sample. Four of the six favored the Demonstration, one more than chance. Overall, we conclude that mental health outcomes were equivalent at the Demonstration and Comparison.

Subgroup analyses based on the program theory predicted better outcomes at the Demonstration for:

1. Those who received intermediate care;
2. Children who received more than outpatient services;
3. Children with severe psychopathology and impairment;
4. Children from families with multiple problems; and
5. Children with more than one type of clinical diagnosis.

While overall there were slightly more findings than was expected from chance alone, the subgroup outcomes at Waves 2 and 3 results favored the Demonstration and Comparison sites equally.

Client Satisfaction

Both adolescents and parents were surveyed about their satisfaction with individual service elements (e.g., outpatient) and overall satisfaction with services. Both sites showed a great deal of satisfaction with services but there was significantly more satisfaction expressed at the Demonstration, especially about services unique to the Demonstration.

Cost/Utilization Outcomes

The Cost/Utilization Study compared service use and the costs of those services for all children treated (n = 6033) at both the Demonstration and Comparison sites. The program theory predicted that the continuum should provide better access to services, a greater diversity in the types of services used, less care in the more restrictive settings, and better timing and continuity of services.

The Demonstration served 13.6% of children who lived in the catchment area[5] in contrast to 6.8% served at the Comparison site. It took an average of 17 days between an assessment and the first services at the Demonstration, in contrast to 38 days at the Comparison site. Children also remained in services much longer in the Demonstration. Six months after starting treatment, 41% of the children at the Demonstration still were receiving services compared with 13% at the Comparison site. Greater continuity of care was shown by the finding that more than 93% of the Demonstration children received a follow-up service within 30 days after leaving the hospital, compared with only 47% in the Comparison group. Finally, defining a dropout as someone who received only one visit, 7% of the clients at the Demonstration dropped out as compared with more than 18% at the Comparison.

Only 8.6% of the children at the Demonstration were placed in an inpatient facility, compared with 15.7% at the Comparison site. Children at the Demonstration received an average of two to three different types of services, compared with the one to two at the Comparison site. Parents at the Demonstration also reported that they were more satisfied with the match between their children's and family's needs and the services received.

The Demonstration costs were much higher than at the Comparison for the three-year study period. The Demonstration spent an average of $7,777 per treated child compared with $4,904 at the Comparison site for the three years studied. The costs of treating the average child at the Demonstration were higher because of longer time spent in treatment, greater volume of traditional services, heavy use of intermediate services, and higher per-unit costs. Data from comparable children in the Outcome Study showed the average cost of Demonstration children, depending on the severity of the child's condition, was from 1.5 to 2.5 times more than the Comparison children. The difference in average expenditures per child does not reflect differences in the mix of children treated.

DISCUSSION

Three assumptions underlie the program model: (a) the successful implementation of a system of care (e.g., case management) and clinical services (e.g., psychotherapy); (b) a more basic assumption that these services are clinically effective even when well implemented; and (c) the successful implementation of the evaluation. Evidence concerning these three assumptions should be closely examined when the findings are both unexpected and show no difference.

Demonstration Was Not An Implementation Failure

The Demonstration had to implement appropriate components of service and ensure that these components were well integrated. The services had to be available to clients quickly to produce smooth transitions from one service level to another. The service needs of clients had to be carefully monitored, and changes in those needs adapted to quickly. If the Demonstration did not implement these features of the theory, then this study would not be a valid test of the theory. Moreover, both services and their management had to be available and of sufficient quality (i.e., excellent case management cannot affect clinical outcomes unless appropriate mental health services are available to manage). Demonstration management and staff were responsible for the appropriate implementation of the theory at the systems level.

Data from the Implementation (Heflinger, 1993; 1996) and Quality Studies (Bickman, Bryant, & Summerfelt, 1993; Bickman, Summerfelt, & Bryant, 1996) showed that the model was implemented with great fidelity. Services utilization data revealed that access to care increased dramatically, and that children at similar levels of impairment were more likely to be treated at the Demonstration without an institutional stay. These findings are evidence that an implementation failure at the Demonstration site was unlikely to have occurred.

Possible Clinical Services Failure

The second way the continuum could have appeared ineffective was if the clinical services themselves were not effective (Salzer & Bickman, in press). In other words, the Demonstration could have faithfully implemented all the aspects of the theory and still there might be no enhanced clinical impact found if the mental health services (e.g., outpatient therapy, hospitalization) provided by the Demonstration were not effective in altering clients' mental health status. These components of services could have failed to be effective for two reasons: they were themselves ineffective, or the Demonstration did not implement them with sufficient quality.

There is no scientifically valid body of evidence that showed that psychotherapy with children and adolescents is effective in *community settings* (Weisz & Weiss, 1993). Furthermore, no evidence exists that the other forms of intervention (e.g., in home services, hospitalization) used in the Demonstration are effective (Rivera & Kutash, 1994). Thus, the Demonstration may have very effectively delivered seemingly more appropriate services in a higher quality system of care, that were nonetheless themselves ineffective.

Implementation of The Evaluation

A poorly designed and implemented evaluation may also mask a treatment effect. The quality of the evaluation can be assessed from a validity framework that focuses on four different types of validity.

Statistical Validity. The statistical power of an evaluation is crucial, especially when trying to explain non-significant effects. An underpowered design would result in a finding of "no-difference" when, in fact, there may be a difference between the treatment and control groups. The initial planning of the evaluation included a sufficiently large sample to detect an effect size of clinical importance. An empirical estimate of power based on Monte Carlo modeling (Lambert, 1993) showed that the evaluation had sufficient power (.80) to detect effect sizes as small as .25 standard deviations. The evaluation's inability to find statistically significant differences was not likely caused by insufficient statistical power.

Internal Validity. The most plausible threats to internal validity in this study were the initial and subsequent equivalency of the Demonstration and Comparison samples most often called "selection artifacts" and "differential attrition." Bickman *et al.* (1995), Breda (1996), and Foster and Bickman (in press) provide detailed information showing the initial and subsequent equivalency of the Demonstration and Comparison Evaluation samples. Again, confidence in the results would have been improved if a randomized experiment could have been conducted. Randomized designs in children's services at the systems level are difficult to implement but are possible. A randomized study currently being conducted in Stark County, Ohio should provide further clarification about the relationship between system level reform and clinical outcomes (Bickman, Summerfelt, Firth, & Douglas, in press).

Construct Validity. Two different aspects of construct validity are of concern here, the construct of cause and construct of effect (Cook & Shadish, 1994). The Demonstration could be seen as ineffective if measurement of the outcomes or effects was faulty. To guard against the danger of selecting inappropriate or insensitive measures, several steps were taken in planning the evaluation. First, widely used and accepted outcome measures were used. Second, for the most critical outcome (mental health status), multiple measures were obtained from several informants. Third, the evaluation has validated the instruments on the project data (Bickman *et al.,* 1995).

Another measurement factor to consider was the length of the follow-up period. It is not clear how long clients, especially children and adolescents, need to be followed before an effect is manifested. However, since the scores of both Demonstration and Comparison children are close to what is considered normal for the population for the normed measures by Wave 4, it is unlikely that differences will emerge in the future unless the Comparison children start getting worse instead of better.

Finally, some aspects of the model presented in Figure 1 were not adequately measured in the Evaluation. In particular there is no established method for determining whether the clinical assessments in the Demonstration led to better treatment plans as hypothesized. We have some evidence from another study conducted at the Demonstration that clinicians may not be reliable in placing children at the assigned level of care (Bickman, Karver, & Schut, 1996).

The other facet of construct validity, the construct of cause, could have been compromised if the Comparison site system of care was equivalent to the Demonstration, but it is not likely that the equivalent clinical outcomes found could be explained by pointing to equivalent systems of care.

External Validity. The external validity of this study was examined by (a) comparing it with other systems interventions; (b) examining the similarity of the client population to other populations; and (c) analyzing the degree to which client characteristics differentially influenced the success of a continuum of care.

a) *Other system interventions.* Two demonstration projects, sponsored by the Robert Wood Johnson Foundation (RWJF), were conducted contemporaneously with the Demonstration. These resembled the Demonstration in that they examined the implementation and effects of integrating mental health services through a centralized continuum of care. The evaluation of the adult demonstration found it did not affect client outcomes, such as quality of life, when compared with other sites (Goldman, Morrissey, & Ridgely, 1994).
b) *Other populations of children.* A comparison of children in the Fort Bragg Demonstration and other populations (Breda, 1996) showed that the military dependent children who participated in the evaluation were similar to middle and lower middle class children and adolescents treated in the civilian sector. They are also most similar to the majority (about 68%) of the children in this country who are covered by private health insurance (Cutler & Gruber, 1995). However, unlike the RWJF studies, the Demonstration's children were primarily from two-parent families with good education and middle to lower incomes. We do not know if the continuum of care would be more effective with children from poorer environments. The inability to specify how client characteristics would function in a continuum reflects the immaturity of the continuum of care theory. However, the results of the Evaluation show that the impact of the continuum on mental

health outcomes was not different for multi-problem families, or for those clients of different SES or demographic characteristics. The present study provides no indication that the continuum would be expected to work better with clients who were more impoverished.

The Demonstration provided the *optimal* conditions under which to test a continuum of care theory. As Palumbo and Olivrio (1989) recognize, "The concern of external validity is generalizing to *future* applications. In doing this, we should be concerned with whether a better program can be put into place in future locations" (p. 342). Future system reform efforts are not likely to have the extensive financial and personnel resources of the Demonstration. If the continuum was not effective under these advantageous conditions (i.e., abundant funds, small interagency problems, intact families), then it is unlikely it will be more successful in more difficult and complex contexts.

Summary

This study has shown that systems reforms can be successfully implemented with sufficient resources that can increase access, treat children in less restrictive environments, and improve satisfaction. However, the notion that costs can be controlled by clinicians and their managers by placing children in what they believed to be the most "appropriate" level of care for the most "appropriate" length of time was not supported. This clinical judgment model was not cost-effective.

The most unanticipated finding of the evaluation is the lack of differences in clinical outcomes between the Demonstration and Comparison sites. Conventional wisdom about what is a better quality system of care is called into question. A fragmented system of care, without these features, did as well clinically and was less expensive.

EVALUATION LESSONS

Several lessons that might be helpful to other evaluators can be extracted from the experience of planning and conducting this evaluation. Nine such lessons are discussed in this section.

1. Use Large Sample Sizes: A Story About Power

It is said that you can't be too rich or too thin (I don't really believe that), but I can add to that list that you can't have too large a sample in conducting an evaluation. We planned the sample size at the outset to be able to detect relatively small effects (.25 standard deviations), since we did not expect the system change to have a large effect. Similar to other field studies, we did find ourselves behind schedule because of the difficulties in recruitment of families. In planning the study we were told we would have a list of those children who just started receiving mental health services, but in fact no such list existed. Thus, we had to use time consuming and expensive recruitment procedures. About two-thirds through the evaluation we had to request more funds and an additional year to meet our goal of about 1000 families. The sponsor did not agree and wanted to complete the evaluation on schedule since they felt that the Demonstration was in fact successful (which was not their initial perception). We had to argue and write reports over a period of six months to convince the sponsor that unless we continued the study the statistical power we needed would not be met, at great detriment to the study.

Since an extension would be expensive, the sponsor hired a consultant with a degree in electrical engineering and statistics, to review our work. We were concerned at first that they may have confused statistical power with some other sort of power. The consultant produced a report that showed that we had met our statistical power goal, given our current sample size using an analysis selected by the consultant. The analysis proposed by the consultant, on this three-wave longitudinal quasi-experiment, was a one-tailed t test on the last wave of data only. We rebutted this design and called in additional experts who criticized this new analysis. The disagreement was actually resolved by congressional intervention, calling into play another type of power, political.

The critical lesson here is that in a field with uncertain measures and effects, large sample sizes are necessary if there is going to be confidence that the possible null effects are true. Moreover, the evaluator has to show that the sample size is a reasonable one based on expected effect sizes. Even with this large sample size, subgroup analyses may be underpowered. From a policy perspective this means that studies of this nature are going to be expensive and time consuming.

2. Develop Program Theory

Given my previous writing, it should not be surprising that program theory was an important aspect of the evaluation. However, we had to develop the theory from scratch, since the field is young and undeveloped in this area. The theory helped us decide what measures were important and allowed us to communicate to the developers of the program and the program staff how we perceived the program and our expectations. The theory warned us about some of the weaknesses of the evaluation, but there was little we could do about some of this, as is pointed out next. More about theory can be found in Bickman (1996a).

3. Be Aware of Limitations and Conduct Auxiliary Studies to Explore Them

Two of the critical assumptions in the theory are that the clinician has the ability to conduct an accurate assessment and to produce a better treatment plan. Theoretically, these are supposed to allow the clinician to place the child in more appropriate services, thus saving money and increasing the chances that the child will improve. We did not give this part of the model much attention when the evaluation was planned, since our focus was on developing the outcomes measurement package. However, once the results showed no effect of the intervention on clinical outcomes, we tried to figure out why. As indicated above we considered the possibility that the services had no effect, but an alternative we could not discard was that children received inappropriate services because we were not able to demonstrate that the assessment process and treatment planning processes were better at the Demonstration, as hypothesized.

If resources are available, it is important to conduct additional studies to try to fill important gaps in the primary evaluation. This approach requires that the evaluator interpret his or her involvement in a substantive area not as a one shot evaluation but as a program of research and evaluation. I believe that this more sustained approach will help ensure that the evaluation contributes to both policy and research. We have conducted one study that supports our concern about the reliability of placement decisions (Bickman, Karver, & Schut, 1996) and see this as an important next line of research and evaluation in this field.

4. Use Multiple Measures From Multiple Sources

The use of multiple measures and sources has long been recommended by methodologists (Campbell & Fiske, 1959). However, many studies still rely on single sources and single methods. The tradeoff we made in taking this approach was to make the research more expensive. In this case the cost was worth it, since it has been much harder to challenge our findings. Once having established the equivalence among the measures we eliminated the most expensive ones (in person interviews) in later follow-ups.

However, I should point out that a major complication that can occur with the use of multiple measures is the distinct possibility that different measures may produce conflicting results. We have been fortunate that the results to date have been internally consistent. Thus the "story" that we tell about the effectiveness of the intervention is simple and straightforward. It is possible that later waves of results may not all go in the same direction. There are no standards concerning how to balance conflicting results in order to draw a fair and unbiased conclusion about the effectiveness of the intervention. In these more complex situations, it should be possible to turn to the program theory for help. Does it make theoretical sense that the intervention should work differently for different outcomes? In this study it is unlikely that the theory of the continuum of care is going to be helpful given its lack of specificity. Similar complexity can occur when the different results obtain for certain subgroups and not for others.

My approach to apparent inconsistency problems is to apply a preponderance of evidence criteria, given that there is no theoretical reason to expect certain subgroups or variables to differ. If the outcomes are equivalent in importance, and I realize that this too can be a difficult assumption to make, and there is no theoretical reason to expect certain outcome variables to act in a way different from most of the others, then I would not place much importance on these "outlier" variables. Clearly readers with different biases will interpret the findings differently. But the evaluator should provide his or her own interpretation, even if others believe that your judgment is not impartial.

5. Measure Cost of The Program

We included a rather large cost component in our evaluation since the policy question focused on providing better outcomes at a lower cost. However, it is expensive and technically difficult to measure costs accurately. A cost effectiveness analysis cannot be done without sufficient resources (Yates, 1996). If there are not sufficient funds and/or expertise, then focus on program effectiveness first. So few programs are effective that additional funds should be more readily available to conduct a cost analysis of an effective program.

6. Plan Start-Up Time for Program and Evaluation

Some critics of our evaluation argue that the program was not sufficiently mature when it was evaluated (Friedman & Burns, 1996). The program had ten months start up time before it started to deliver services. We did not start to collect most of the baseline data until eight months later. While the program was always in some transition, clearly the first six months were much less stable. Unfortunately, the program theory was not sufficiently developed to indicate when the program was ready to be evaluated. However, the implementation and quality substudies gave us confidence that key features of the Demonstration were well implemented during the data collection period. Moreover, there was no consistent relationship between our outcome measures and when a child entered the evaluation. Nonetheless, when the results do not conform to their expectations, critics may still believe the evaluation occurred too soon, regardless of the timing. The evaluation had a year start-up time that was necessary to develop and pilot test measures and hire and train staff.

7. Understand The Political Environment

One of the features of the field of evaluation that makes it most challenging is its political environment. The U. S. Army did not want to fund the Demonstration and its evaluation until some influential congressmen convinced them that it was in their best interest to fund the Demonstration and Evaluation. The private hospital chains thought the program's emphasis on non-hospital care would harm them (managed care actually took care of that) and fought the project in several ways. Private practitioners in the Fayetteville community thought they would lose money (they didn't). We found that in most cases politics had no effect on the design of the evaluation because the debated issues primarily concerned whether the Demonstration and its evaluation should be allowed to proceed. In this case, data and logic were usually not the issue, but political influence was important. Having strong political allies in high-profile and controversial evaluations is critical. It was also important that the person managing the overall project (Dr. Lenore Behar) supported the evaluation and buffered a great deal of political interference. However, politics of the sort discussed next may depend on data.

8. Deal Directly With True Believers

If the evaluation has important policy considerations, then be prepared to deal with the true believers if the evaluation does not support their strongly held beliefs. In human services, where there are few facts, there are many committed individuals with intensely held beliefs. When there are no empirically established facts, ideological beliefs are often the substitute. In this evaluation, child advocates believed that systems of care can solve most problems. Large federal programs have been funded on this belief. Evaluations that challenge these beliefs, and their funding streams, receive a great deal of scrutiny. However, another reaction I have observed is that the more contradictory data that is presented to a true believer, the more firmly that they hold to their ideology. So do not expect the facts alone to cause change.

Our approach to dealing with the critics of this study has been to invite them to dialog with us. In three special issues of journals (Bickman, 1996a; 1996b; Bickman, Heflinger, Lambert, & Summerfelt, in press) and in an *American Psychologist* (Bickman, 1996c) article my colleagues and I have invited the critics to comment on the study, with the provision that we be able to write a rejoinder. In addition, a book describing the evaluation has been published, and we have made numerous presentations at professional meetings.

The response of the critics has been interesting. There has not been a direct attack on the methodology or me, instead, critics have tried to ignore the study, perhaps in the hope that it would just go away. For example at a recent conference on children services there was a panel discussion of the Fort Bragg study; however, no one from the evaluation team was invited to participate.

Finally, in dealing with advocates one has to understand the motivation for advocacy. I believe that child advocates honestly want to help children who have problems. However, I also believe that these programs have a moral obligation to determine if they are in fact helping children and not just the organization. I am sure that the advocates also want to establish that services are helpful. Thus, I try to point out that holding on to some failed ideology does not accomplish their main objective. The weakness in this argument is that as an evaluator I am unable to tell them what approach they should take to ameliorate children's problems. I can suggest which studies and evaluations need to be conducted but as a researcher I am helpless to direct them to other interventions that should be more successful since none have been found at this time. This is not a satisfying position.

9. Conduct Additional Evaluations

I noted earlier the need to conduct small auxiliary studies to answer modest questions raised by the evaluation like the ability of clinicians to make reliable level of care determinations. I think it is also important to conduct additional evaluations of similar constructs that address some of the perceived weakness of the original evaluation. I was fortunate to foresee some of the problems that the Fort Bragg study had that could be corrected in another study. The five main concerns were that Fort Bragg was: (1) a quasi-experiment; (2) with military families; (3) a demonstration; (4) a continuum and not a true system of care; (5) was a new program. In designing the Stark county evaluation I wanted to be sure to address these apparent limitations of the Fort Bragg evaluation. The Stark County project used a randomized design, which to my knowledge has never been used in a systems evaluation of children's services. We successfully dealt with the methodological, logistic and ethical concerns that usually accompany these designs (Bickman, Summerfelt, Firth, & Douglas, in press). The Stark County evaluation used a civilian population that was primarily publicly supported. Thus, they matched the population of children that most programs and research deal with in this county, even though they do not represent the majority of children in this country. We also used a system of care that was firmly established over the past 20 years. Moreover the system was developed by the community, it was not a special demonstration. Finally the Stark County system was integrated across all child serving agencies in the community. While the Fort Bragg Demonstration was primarily a mental health sector project, Stark prides itself as a true system of care. We were also fortunate that Stark system has recently been independently evaluated for its systems aspects and has been found to be the second ranked program in the country that is participating in the competitive Center for Mental Health Services children's mental health program (Macro, 1996). If the results of the Stark county project are similar to Fort Bragg, this will strengthen the findings of the Fort Bragg study.

In addition to the Stark County evaluation funded by the NIMH, we have also received a competitive renewal of our original NIMH grant to continue to follow the children in the Fort Bragg evaluation. We proposed a much more economical follow-up that is being conducted by phone and mail for an additional three years. While we do not expect to find differences between the control and treatment children, there are no similar longitudinal studies of children who were in treatment. Regardless of the evaluation issues, the investigation of the long term functioning of these children should be important, especially as they transition to young adulthood.

The U. S. Army was pleased with the objective nature of our evaluation and invited us to submit an additional proposal for a significant amount of funding. The additional funds will be used to further explore the data set to study such questions as the effects of Desert Storm on children and whether premature termination of services affects outcomes. One major primary data collection effort is focusing on an implementation analysis that is exploring the transition from a cost reimbursement demonstration to services managed by a for profit managed care company operating under a capitated contract. A second study is attempting to examine outpatient outcomes on a finer grain basis than the every six-month data collection that we followed in the evaluation. This study plans to collect data from parents, children, and mental health providers on a weekly basis to gauge short term changes. Both of these studies should contribute to our knowledge of child and adolescent mental health and substance abuse services.

Conclusion

In conclusion, the evaluation produced some surprising results that have not been easily accepted by program supporters and child advocates. Instead of calling for more services, our conclusions have led us to question the effectiveness of services now provided, regardless of the system of care (Salzer & Bickman, in press). The evaluation has led us to propose additional studies investigating the quality of care provided to children and adolescents who have mental health and substance abuse problems. This evaluation has contributed to the scrutiny of these services.

Finally, is this evaluation really exemplary? No, if exemplary is construed as "without fault," for every evaluation is flawed or inadequate in some particular. Yet I believe time will prove that this study was conducted carefully and rigorously and that any minor compromises in its conduct did not alter its major findings. Hopefully, it will serve as an example from which colleagues faced with evaluations of important social services can draw useful lessons.

ACKNOWLEDGMENTS

This research was supported by the U.S. Army Health Services Command (DADA 10-89-C-0013) as a subcontract from the North Carolina Department of Human Resources/ Division of Mental Health, Developmental Disabilities and Substance Abuse Services, and a grant to Dr. Leonard Bickman (R01MH-46136-01) from the National Institute of Mental Health.

This five year project was the work of many individuals and organizations. I would like to especially recognize Lenore Behar for her initiative in conceptualizing the Demonstration and her strong support for an independent evaluation and the contributions of the Vanderbilt management team: Pamela Guthrie, E. Michael Foster, E. Warren Lambert, Wm. Thomas Summerfelt, Carolyn Breda, and Craig Anne Heflinger. I would like to thank Mark Salzer and E. Warren Lambert for their comments on this manuscript.

NOTES

1. Parts of this article were adapted from an article in the *American Psychologist* 51(7), 689-701.

2. The capitalized terms, "Demonstration" and "Evaluation" used throughout this paper refer, respectively, to the Fort Bragg Demonstration Project and the Fort Bragg Evaluation described in subsequent sections. "Comparison" will be used to refer to the comparison group of children receiving traditional CHAMPUS services.

3. This article can only provide an overview of the methods and results of this large scale study. More information about the Fort Bragg study can be found in Bickman (1996); Bickman et al. (1995); and Bickman et al. (in press).

4. CHAMPUS refers to the insurer of dependents, the Civilian Health and Medical Program of Uniformed Services.

5. The 13.6% figure is not a true measure of prevalence, but the number of children served divided by the average number of eligible children in the catchment area. Dividing by the average size served is a means of standardizing the measure for differences in catchment area size. A measure of prevalence, however, would use as a denominator the number of children who lived in the catchment area at any point during the Demonstration period. Because families moved in and out of the catchment area during the Demonstration, this figure is no doubt larger than that of the number of children living in the catchment area on a day in a particular year. A measure of the prevalence of mental health services use, therefore, would be less than 13.6%.

REFERENCES

Baine, D. P. (1992). Prepared statement to the select committee on children, youth and families of the United States House of Representatives. In *The profits of misery: How inpatient psychiatric treatment bilks the system and betrays our trust.* (GPO 1992-52-362). Washington, DC: U.S. Government Printing Office.

Behar, L. (1985). Changing patterns of state responsibility: A case study of North Carolina. Special issue: Mental health services to children. *Journal of Clinical Child Psychology*, *14*(3), 188-95.

Bickman, L. (1985). Improving established statewide programs: A component theory of evaluation. *Evaluation Review*, *9*, 189-208.

Bickman, L. (Ed). (1987). *Using program theory in evaluation.* San Francisco: Jossey-Bass, Inc.

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

Bickman, L. (Ed). (1996). (Special Issue) *Journal of Mental Health Administration*, *23*(1).

Bickman, L. (Ed). (1996b). The evaluation of the Fort Bragg Demonstration. (Special Issue). *The Journal of Mental Health Administration*, *23*(1).

Bickman, L. (1996c). A managed continuum of care: More is not better. *American Psychologist.*

Bickman, L., Bryant, D., & Summerfelt, T. (1993). *Final report of the quality study of the Fort Bragg evaluation project.* (Unpublished manuscript) Vanderbilt University Center for Mental Health Policy.

Bickman, L., Guthrie, P. R., Foster, E. M., Lambert, E. W., Summerfelt, W. T., Breda, C. S., & Heflinger, C. A. (1995). *Evaluating managed mental health services: The Fort Bragg experiment.* Plenum Publishing: New York, NY.

Bickman, L., Heflinger, C. A., Lambert, E. W., & Summerfelt, W. T. (1996). The Fort Bragg managed care experiment: Short term impact on psychopathology. *Journal of Child & Family Studies*, *5*(2), 137-160.

Bickman, L., Karver, M. S., & Schut, L. J. A. (Submitted). *Clinician accuracy and reliability in judging appropriate level of care.*

Bickman, L., Summerfelt, W. T., & Bryant, D. (1996). The quality of services in a children's mental health managed care demonstration. In L. Bickman (Ed.) The evaluation of the Fort Bragg demonstration. (Special Issue). *The Journal of Mental Health Administration*, *23*(1), 30-39.

Bickman, L., Summerfelt, W. T., Firth, J., & Douglas, S. (in press). The Stark County Evaluation Project: Baseline results of a randomized experiment. In D. Northrup & C. Nixon (Eds.), *Evaluating mental health services: how do programs for children "work" in the real world?* Newbury Park: Sage.

Breda, C. S. (1996). Methods for evaluating mental health outcomes: In a children's mental health managed care demonstration. In L. Bickman (Ed.) The Evaluation of the Fort Bragg Demonstration. (Special Issue). *The Journal of Mental Health Administration*, *23*(1), 40-50.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, *56*(2), 81-105.

Chen, H. (1990). *Theory-driven evaluations.* Newbury Park, CA: Sage.

Chen, H., & Rossi, P. C. (1987). The theory-driven approach to validity. *Evaluation and Program Planning*, *10*, 95-103.

Cook, T., & Shadish, W. (1994). Social experiments: Some developments over the past fifteen years. *Annual Review of Psychology*, *45*, 545-580.

Cutler, D. M., & Gruber, D. (1995). *Does public insurance crowd out private insurance.* National Bureau of Economic Research, Inc. Working paper No. 5082. Cambridge, MA.

Foster, E. M., & Bickman, L. (in press). An evaluators guide to detecting attrition problems. *Evaluation Review.*

Friedman, R. M., & Burns, B. J. (1996). The Evaluation of the Fort Bragg Demonstration Project: An Alternative Interpretation of the Findings. *Journal of Mental Health Administration*, *23*(1), 128-136.

Goldman, H. H., Morrissey, J. P., & Ridgely, S. M. (1994). Evaluating the Robert Wood Johnson Foundation program on chronic mental illness. *Milbank Quarterly*, *72*(1), 37-48.

Heflinger, C.A. (1993). *Final report of Fort Bragg evaluation: The implementation study.* Unpublished manuscript: Vanderbilt University Center for Mental Health Policy.

Heflinger, C. A. (1996). Implementation analysis of a children's mental health managed care demonstration. In L. Bickman (Ed.), The evaluation of the Fort Bragg demonstration. (Special Issue). *The Journal of Mental Health Administration*, *23*(1), 16-29.

Lambert, E. W. (1993). Monte Carlo modeling. In Bickman, L. (Ed.), *Response to the directorate of health care studies and clinical investigations final report [revised].* Unpublished manuscript: Vanderbilt University Center for Mental Health Policy.

Macro, Inc. (1996, April). *Report from Year One: The National Comprehensive Community Mental Health for Children and Families Program Evaluation.* Macro International, Inc. and University of South Florida.

Palumbo, D. J., & Olivrio, A. (1989). Implementation theory and the theory-driven approach to validity. *Evaluation and Program Planning*, *12*, 337-344.

Rivera, V. R., & Kutash, K. (1994). *Components of a system of care: What does the research say?* Tampa, FL: University of South Florida, Florida Mental Health Institute, Research and Training, Center for Children's Mental Health.

Roberts, M. C. (1994). Models for service delivery in children's mental health: Common characteristics. *Journal of Clinical Child Psychology*, *23*, 212-219.

Salzer, M., & Bickman, L. (in press). Delivering effective children's services in the community: Reconsidering the benefits of system interventions. *Applied and Preventive Psychology.*

Stroul, B. A., & Friedman, R. (1986). *A system of care for children and youth with severe emotional disturbances [rev. ed.].* Washington, DC: Georgetown University Child Development Center, CASSP Technical Assistance Center.

Weisz, J. R., & Weiss, B. (1993). *Effects of psychotherapy with children and adolescents.* Newbury Park, CA: Sage.

Yates, B., (1996). *Understanding & Improving Costs, Processes & Outcomes.* In L. Bickman & D. Rog, Applied Research Methods Series. Thousand Oaks, Sage.

Yin, R. K. (1986). *Case study research.* Newbury Park, CA: Sage.

Yin, R. K. (1993). *Applications of case study research.* Newbury Park, CA: