Multiagency Outcome Evaluation of Children's Services: A Case Study

Shirley A. Beck, M.S.S.A.
Pamela Meadowcroft, Ph.D.
Matthew Mason, Ph.D.
Edward S. Kiely, Ph.D.

Abstract

Outcome monitoring has become a focus of accountability for public and nonprofit human service agencies. Besides providing answers to funders' questions about the services' impact, outcome monitoring helps administrators improve program effectiveness. After a three-year development period and a one-year implementation experience, SumOne for Kids represents a technically advanced outcome-monitoring system for children's mental health and/or child welfare services. Initiated, designed, and tested by 31 children's service agencies throughout Pennsylvania, and with state bureaucrats' and policy makers' encouragement, SumOne for Kids represents an effort to create a bottom-up/top-down process for implementing a statewide outcome-monitoring system. This article describes the genesis of this outcome-monitoring system, primary design principles, use of social validation for outcome selection, resolution of methodological difficulties, and reasons for selecting functional over clinical outcomes. The article reviews lessons learned through the development experience instructive to children's service managers, program evaluators, and industry leaders interested in establishing outcome-monitoring systems.

In recent years, measuring and monitoring outcomes has become a critical issue in the child welfare and mental health fields. Children are now entering the service system in larger numbers and with more severe disturbances.[1] Moreover, our most costly forms of mental health treatment (psychiatric hospitalization) and child welfare services (residential care) fail to provide payers or consumers with outcome accountability.[2] But, as the cost of services for these children continues to rise, service providers face increasing demands from legislators and government officials for data that provide evidence of service efficacy. Even the nonprofit, human service sector, through United Way policy directives,[3] is now receiving direct pressure to report on outcomes.

As Mary Jane England argues in Partnerships for Care, "Within the new health care system, the operative word will be value. Care will be provided—and paid for—on the basis of appropriateness and effectiveness as well as cost . . . we need to define, measure, and use . . . outcomes to determine

Address correspondence to Matthew Mason, Ph.D., Director, Center for Research & Public Policy, The Pressley Ridge Schools, 530 Marshall Avenue, Pittsburgh, PA 15214; e-mail: iammason@aol.com.

Shirley A. Beck, M.S.S.A., was a research coordinator, Center for Research & Public Policy, The Pressley Ridge Schools, Pittsburgh, Pennsylvania at the time this article was written.

Pamela Meadowcroft, Ph.D., is deputy executive director, Center for Research & Public Policy, The Pressley Ridge Schools, Pittsburgh, Pennsylvania.

Edward S. Kiely, Ph.D., is a project associate, Center for Research & Public Policy, The Pressley Ridge Schools, Pittsburgh, Pennsylvania.

the effectiveness of . . . care" (p. iii).[4] In addition, L. W. Kaplan of the United Way writes, "[W]e aren't as much concerned about what an agency says it's going to do, or how many people it is going to serve. We want to know how the client (the customer) will be better off: What will this person have learned; what will this person be better able to do" (p. 17).[3]

The general need for good information systems and the specific need for knowledge of results is not new. Requests for increased accountability through good data management began as soon as government expenditures were allotted to the care of dependent children. Yet, the situation today is not markedly different from that in 1923:

> No one in Pennsylvania knows the extent of the expenditures of public funds for the care of dependent children, no one knows the number of such children now being maintained, no one knows the number of families involved, no one knows the extent of the turnover among the clients . . . and no one knows the extent to which some of the expenditures have brought good or ill to the recipients, and certainly no one knows whether better service might not have been achieved for a smaller outlay more intelligently applied. The Children's Commission regards this as perhaps Pennsylvania's most serious situation affecting children. (p. 28)[5]

A software database program, SumOne for Kids,[6] is an outcome-monitoring system that measures the effectiveness of children's services in Pennsylvania.[7] Using SumOne for Kids, one can create a practical method to examine outcomes and, hence, affect the daily operations of services. The system was created with close attention to the data and reporting needs of agencies that provide direct care, supervision, mental health, education, and social services to children and their families. Over the course of three years, Pennsylvanians were surveyed to determine what outcomes to measure, data elements were developed and tested, and the database software was designed and tested.

The experiences gained from this effort should be instructive for any children's service manager, program evaluator, or industry leader who wants to establish an outcome-monitoring system. This article will describe the genesis and progress of SumOne for Kids and will pay special attention to design and guiding principles (both those adopted at the outset and those picked up along the way). SumOne for Kids remains a work in progress until it reaches its ultimate goal of implementation in large numbers of children's service agencies in Pennsylvania and the accompanying creation of a central database for benchmarking results in children's services. Apart from the public policy issues and political realities peculiar to this goal, the development and implementation of a widely used outcome-monitoring system lends itself as a representative case study of the generic issues that must be faced by outcome system designers.

Project Genesis

The focus on outcomes has been part of a sweeping national movement that was keenly felt in Pennsylvania as it struggled to adopt outcome-based education in 1993. Interest among the child-serving nonprofit agencies in Pennsylvania was initiated by John Pierce of the Pennsylvania Council of Children's Services (PCCS) and Clark Luster of The Pressley Ridge Schools (PRS), whose organizations were strategic in launching the effort.

The PCCS is a statewide association of nonprofit child-serving agencies. PCCS is made up of 85 agencies representing 70% of all private, nonprofit children's service agencies in the state. Collectively, these agencies serve approximately 20,000 children and their families on any given day with the full spectrum of children's services, including both in-home and out-of-home care.

The PRS is a 165-year-old, nonprofit, multistate children's service agency that serves about 1,500 children and their families per day. PRS provides an array of social and mental health services and special education programs for troubled children and their families in Pennsylvania, West Virginia, Ohio, and Maryland. The array of services includes treatment foster care, educational programs, family preservation and in-home services, and residential services. For the past decade, PRS has conducted an annual outcome study to measure the effectiveness of the services provided to kids and

families. Each year, staff members of PRS's Center for Research & Public Policy contact children (and their caregivers) who received at least 30 days of service and were discharged about one year earlier. Through the use of structured interviews with each child and her or his caregiver, outcome information is gathered.

PRS's outcome studies were originally designed to be a simple feedback system on a few socially significant outcomes (e.g., school attendance, restrictiveness of residential and educational environments, frequency of antisocial activities, satisfaction).[8] Demographic information from case records was included to create a "snapshot" of each child. PRS's cumulative efforts represent information from some 2,000 children. The most recent outcome study alone collected information from more than 700 discharged children who received services in the same year. The results of these outcome studies are used to evaluate the impact of PRS's programs, and to plan strategically for program improvement.

PCCS and PRS began discussions regarding the development of an outcome-monitoring system for all providers of children's services in Pennsylvania years before the actual development began. The key to launching the development was securing the approval of the PCCS Board; this approval was not automatic. Although most PCCS member agencies appreciated the merits of outcome measurement, there was no initial groundswell of support for the development of an outcome-monitoring system. PCCS support was nurtured over several years.

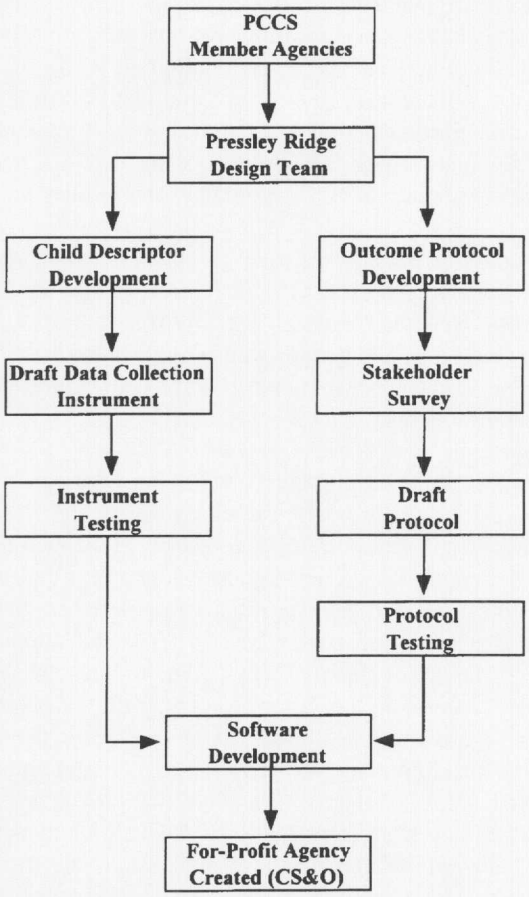
Many PCCS member agencies were unconvinced that the benefit of developing an outcome-monitoring system would be worth the development and implementation costs. Most felt that their mandated data reporting was already too extensive and too fragmented. Another source of resistance was based on the fear that the data collected could be used against member agencies. An outcome-monitoring system designed to report agency service results to funding entities is inherently threatening. Overall, these concerns generated passive, not active, resistance; a strong minority of the agencies were supporters, and most of the rest adopted a "wait and see" attitude. No member agency was diametrically opposed to the development of an outcome-monitoring system.

In 1989, the PCCS Board was finally persuaded to pursue the development of the outcome-monitoring system, with five arguments prevailing: (1) State legislation mandating outcome monitoring for children's services was imminent; the drive for cost control and accountability for the effective use of dollars made it so. Without answers about the empirical data on the actual information related to results of services, such a system would put funding for agency programs at risk. (2) Member agencies realized they would benefit more by participating in the development of an outcome-monitoring system that could be offered for statewide use, rather than have an outcome system designed and then imposed by the state. (3) Outcome monitoring was recognized as an important tool that could assist member agencies to better manage their resources. (4) Member agencies recognized the need for a multiagency system, which would help establish a central database of service outcomes and allow each member agency to compare their own service outcomes to those stored in the central database. (5) PRS had demonstrated, through its outcome studies, that outcome monitoring could be accomplished without using traditional, external, costly program evaluation models. Outcome measurement was a feasible answer to the call for accountability and management of results.

When the idea of creating an outcome-monitoring system was presented at a PCCS general membership meeting, 60 member agencies volunteered to participate in the development of such a system. Of these, 31 immediately pledged to commit staff time to the tasks involved in development and testing. With this show of support, PCCS leadership decided to create an outcome measurement system. Figure 1 summarizes the developmental process, which is explained in detail.

PCCS asked PRS to lead the design, testing, and implementation of a prototype outcome-monitoring system on the behalf of PCCS member agencies. PRS was charged to create an outcome-monitoring system that would (a) guide the collection of data in each participating agency and (b) aggregate the data into a central database. The outcome-monitoring system needed first to

Figure 1
Development Flow of Outcome-Monitoring System



provide useful information to program managers, and second to meet the information objectives of the multiagency outcome system.

Since the ultimate sanction for a comprehensive statewide system would come from the state government, it was essential to involve state-level children's service officials from the beginning. A bottom-up/top-down model for planning and implementation was consequently adopted. This approach began by assembling a group of state-level bureaucrats, policy makers, educators, researchers, and practitioners to inform them of the purpose of the outcome-monitoring system, gain the support of their offices, and involve them in articulating a vision of what the system might look like. Those present gave high praise to the concept, pledged their full support, and asked to be kept informed of progress. With this endorsement, the development of the outcome-monitoring system was launched in 1990 (initially called the Pennsylvania Outcome Project for Children's Services, and later renamed SumOne for Kids).

Development began by first learning about the 31 participating agencies and building a working relationship with them. The hope was that the agencies that volunteered for the development of SumOne for Kids would represent the diversity of child-serving agencies within the state. Such

representation would maximize the generalizability of the outcome-monitoring system. Structured interview instruments were used to gather information at the agency and program level about history, funding sources, demographic makeup of the client constituency, average length of stay, admission and discharge criteria, program offerings, and services offered within each program.

These 31 agencies, in fact, represented the desired diversity. Geographically, they are located in 13 of Pennsylvania's 67 counties, covering all four regions of the state. The agencies range in age from 15 to 163 years. Most began by offering a single service or were an orphanage or other residential facility and now offer several programs with an average of 4.8 programs per agency. The most frequently operated program types are residential treatment facilities, foster family care, in-home and family preservation services, community-based group care facilities, partial hospitalization, special education (approved private schools for seriously emotionally disturbed children), and day treatment. More than two-thirds (67%) of the agencies provide mental health services in addition to their more traditional child welfare services. The number of Pennsylvania children served daily by the 31 agencies ranges from 44 to 595, with an average of 165. Although many agencies accept both boys and girls, the combined client population is approximately two-thirds male, one-third female. Nearly two-thirds (64%) of the agencies serve school-age children, whereas the remainder (36%) serve children of all ages. Funding and referrals for program services come primarily from the fields of child welfare, mental health, juvenile justice, and education.

System Design

Guiding Principles

From its inception, SumOne for Kids was guided by five basic principles. The first of these was to involve individuals in design and development who would eventually use the management system. The system development relied on these persons to articulate the technical specifications of the system, test the products, examine how system requirements could be built into ongoing program operations, and, ultimately, ensure that the system was user friendly and useful.

The second principle was to build on existing products. For example, PCCS had for some time required its member agencies to submit pen-and-paper information using its 31-item Child Profile Form. PRS was using its own pen-and-paper instrument, the Child Descriptor at Entry Form. An early decision in the life of the project was to build from these two instruments to create the project's child descriptor component.

When designing the system, project staff discarded such traditional program evaluation approaches as point-in-time measures of a particular program's or intervention's success, usually conducted by an external, objective evaluation researcher. Instead, the third design principle was to base the system on a self-evaluation model.[9] In this model, evaluation is not done to the program; rather, it is done with the program. Self-evaluation assumes that providers of services want to do a good job and that the purpose of an outcome evaluation system is to help them do so. According to Mathison, self-evaluating organizations are

> peopled by reflective individuals who have no particular expertise in program evaluation, but characteristically are interested in self-monitoring and self-improvement. . . . Proponents of this type of internal evaluation perceive managers and service providers to be capable of and prepared to conduct evaluations of their own programs. The logic is that those who develop programs and provide services are in the best position to know what their intents are and have the greatest interest in improvement. (p. 160)[10]

The fourth principle involved articulating what a useful outcome evaluation system would include: (a) a detailed description of the client population; (b) a description and count of the types of services delivered; and (c) measures of outcomes, or results, of the services on the lives of the clients.[11]

The fifth and final principle was the emphasis on end or functional outcomes rather than on process or clinical outcomes.

Process outcomes reflect how services are delivered, or other key operational indicators of service. Examples of process outcomes might include length of stay, number of therapy sessions, cost per unit of service, and so on. Process outcomes are clearly important from a service delivery standpoint, but they offer no insight into the effect of services on consumers or communities. Rosenblatt and Atkisson[12] provide a conceptual framework for discussing various types of outcomes, including clinical and functional.

Clinical outcomes reflect psychological and physical changes related to signs or symptoms of disorders. Measurement of clinical outcomes might include the assessment of physical, emotional, cognitive, and behavioral signs related to a disorder. The reduction of symptoms of depression (evidenced by a change in score on a depression inventory, for example) would be one example of a clinical outcome. The value of clinical outcomes is in their individuality; however, they lack generalizability.

Functional outcomes reflect the effect of services on skills needed for an individual to succeed in his or her own community and lead a productive life (the end results). Examples of functional outcomes could include employment, school attendance, longevity of friendships, one's living situation, and so forth. Often, outcome measurement systems focus primarily on process or clinical, and not on functional, outcomes. In the development of SumOne for Kids, functional outcomes were emphasized over clinical because they are more universally measurable, more comparable across agencies, and provide data that can be readily understood by the public.

Formation of Development Groups

A consensus-building approach using small development groups of agencies to critique and test draft instruments was selected. Two functional development groups were created: (1) the identification and testing of a set of data elements to describe child and/or family clients and the services delivered to them (child descriptor database group) and (2) the development and testing of the outcome-gathering instruments (outcome protocol group). From the pool of 31 participating PCCS member agencies, 8 volunteered to participate in each area as a development group. PRS served as a pilot agency in both areas, and thus 17 agencies were involved in the development process. The remaining 14 PCCS member agencies were kept informed of developmental activities through a newsletter and by attending annual membership meetings.

Child Descriptor Database

The purpose of this development group was to identify a set of data elements to serve as descriptors of clients at point of entry, during treatment, and at discharge. These descriptive elements would be the foundation of the agency database from which a common multiagency database (central database) could be developed.

The first draft instrument was formed by the data elements in the existing pen-and-paper data-gathering forms used by PCCS and PRS, as well as other data elements from the outcome research literature. Building on existing data-gathering practices was useful because it both increased the speed of development and helped create consensus. From this first draft, a computerized prototype was created for demonstration purposes. Administrators and staff could actually see this early product in operation, which helped them remain committed to completing the laborious, detailed steps required to develop a final, reliable product.

After several iterations of agency feedback and revisions, a draft instrument emerged for formal testing. The instrument collected such information as the child's demographic characteristics, family characteristics, history of the child's residential and educational placements and their

restrictiveness levels, psychiatric diagnoses and medications, insurance information, behavioral/physical problems, type of program serving the child, specific services provided, and permanency goals.

Three rounds of reliability testing were conducted involving staff from three different agencies per round. Staff participating in the test included intake workers, direct care/therapeutic staff, and middle managers. The testing involved two staff from each agency independently coding the same five cases using the draft instrument. Project staff reviewed their work to determine the nature and amount of coding differences. Analysis of each test led to a revised draft instrument to be used in the next test round. In every test round, the average time required to code the last case was less than half the average time required to code the first case. The pair-agreement rate on the 45 test cases ranged from a low of 66% in the first round to a high of 93% in the third round; overall agreement for the third round was 84%.

There were two results from this testing process: (1) a comprehensive child descriptor data-gathering instrument was produced that would form the basis of an electronic case file and (2) a standard was established that at least one staff member in each participating agency must meet. The standard is to achieve at least 85% agreement with project staff on the task of translating narrative case record information into code using a set of test cases.

Outcome Protocol

The purpose of this development group was to create an outcome measurement system. The first step in completing this task was to determine what outcomes the outcome-monitoring system (SumOne for Kids) would measure. This step was the most sensitive part of the development of the system. Because the ultimate aim was to collect data on common outcome measures, it was extremely important to minimize controversy in the selection of measures.

To ensure that the selection would be politically defensible, project staff decided that the measures should not simply be culled from the literature in order to avoid criticism on what literature was consulted, and should not be determined primarily by the opinions of "experts" (agency administrators, project staff) in order to avoid the criticism that the selection was skewed to make the programs appear successful. Instead, project staff decided that the outcomes selected for measurement would be first determined by use of social validation techniques,[12] a methodologically conservative approach for targeting such subjective issues as outcome definition. Such an approach allows issues of social importance to be judged by society, since these issues are based on subjective values.[13]

A social validation process polls members of society who are potentially most affected by the service.[14] These key members of society, or stakeholders, represent diverse societal groups (e.g., direct consumers of services; community members and leaders; community program personnel; advocacy groups; and local-, state-, and federal-level policy and political leaders).[15] Using social validation techniques, stakeholders can be asked what outcomes they expect of the services or program. The values of stakeholders make up the foundation on which the project is based.

Following specific criteria, the 31 agencies nominated potential survey responders in each of nine stakeholder groups: child clients (age 12 or older), parents of child clients, agency board members, school board members, community agency workers, poverty-level representatives, juvenile court judges, local legislators, and informal community leaders. This nomination process ensured that the stakeholders would represent the various regions of the state, different funding streams, and different program types. The final survey population included 700 children's service stakeholders across Pennsylvania.

The survey instrument was designed to gather ratings on 49 children's issues in terms of two dimensions: importance and satisfaction. The stakeholders used two 5-point scales to rate the importance of each issue and their satisfaction with the local services that addressed each issue. The issues were organized by life domain areas of educational, vocational, safety, living arrangements, family life, emotional/psychological, medical/psychiatric, social/recreational, cultural, spiritual, and legal.

Stakeholders were also given the opportunity to add issues not included on the list and to rate those issues in terms of importance and satisfaction. A 90% response rate was achieved.[16]

Issues of high importance were identified when 90% or more of the stakeholders rated the issue a "5." Likewise, issues of low satisfaction were identified when 10% or more rated the issue a "1." Seven issues met the criteria for high importance; three of those seven also met the criteria for low satisfaction.

To ensure the buy-in of the project agencies, after informing the agency administrators of the results of the stakeholder survey, administrators were asked to rank order the same set of 49 issues. The top 10 from the administrators' ranking included six of the seven issues of high importance from the larger survey. Safe communities for raising children, the one high-importance criteria-level issue from the stakeholder survey, was not included among the administrators' top 10 because it was considered outside the realm of this set of agencies' services.

Six issues both met criteria for high importance in the stakeholder survey and were ranked among the top 10 of the agency administrators:

1. Children being taught the values of right and wrong, good and bad
2. Child neglect and abuse
3. Use of drugs and alcohol by children and teenagers
4. High school graduation
5. Parents being involved with their children in positive ways
6. Children attending school regularly.

The following four issues, although not meeting the criteria level in the stakeholder survey, were still of considerable importance, and were added by the administrators. The stakeholder survey rank ordering of these four issues is shown in parentheses following each issue:

1. Children having stable, long-term places to live (9th)
2. Children learning not to be aggressive (14th)
3. Children being protected from aggression or harm (8th)
4. Youth being taught skills for independent living (13th).

Project staff researched appropriate measures for each of the administrators' top 10 issues and developed a protocol for gathering the related outcome information. The protocol included (a) six draft interview instruments (three for use with children in age groups 3-6, 7-12, 13-18; the other three for use with the respective caregivers), (b) a monthly collection schedule for up to one year postdischarge, (c) a policy of confidentiality to child clients (except in instances of risk of harm to the child) in exchange for honest answers, (d) a policy on staff eligibility to conduct the outcome interviews, (e) a method for comparing an agency's or system's outcomes to existing norms, and (f) a component to determine the severity of a child's entering problem(s).

These materials were reviewed to determined whether implementation was possible using the existing agency staff. Since a primary goal of the project was to develop a low-cost, easy-to-use system, if agencies reported that they appeared to need additional staff for implementation, the protocol would be reduced in scope. As a result, the protocol was determined to be too large, complex, and burdensome. It was felt that 10 issues were too many to measure, the structured interviews should be more generic to eliminate the need for different instruments for children of different ages, and the frequency of collection should be reduced to fit an agency's existing routine. In response, the six instruments were reduced to two (one for the child and one for the caregiver), and the proposed interview frequency was reduced from monthly to quarterly throughout the duration of services and up to one year after discharge or completion of services.

Most significant, outcomes were consolidated into five major areas:

Productivity: school attendance, graduation, and/or employment;
Antisocial activity: drug and/or alcohol use;
Living environment: stability and restrictiveness of the child's living environment;

Protection from harm: frequency of injury or abuse by peers or adults and frequency of threats of harm from peers or adults;
Client satisfaction: satisfaction with living arrangements, school or work, and with life in general.

For the most part, the first four outcome areas above abstract the larger themes in the administrators' top 10 list while retaining most of the criteria-level issues of importance of the stakeholder survey. The fifth outcome area, client satisfaction, while not a functional outcome, was added to supplement information on program effectiveness.

Two items from the review packet were tabled for future consideration: (1) norms for comparing results of agencies' outcomes and (2) severity rating of each child's disabilities or problems. Comparisons of agencies' outcomes with existing state or national norms on the above indicators would likely lead to erroneous conclusions because such norms are based on different child populations than those being served by the agencies. Because existing norms provide information on a representative sample of all children, the outcomes of agencies serving seriously troubled children would likely compare unfavorably. Educating policy makers on how to use norm comparisons would be too time consuming to consider at this stage in the development of SumOne for Kids. In addition, the ideal standard for comparisons should be based on the children served by participating agencies. However, such comparisons would be available only after the creation of a central database from which benchmarks or standards could be generated.

Severity of disturbance was seen as an important variable to include in the outcome evaluation system to ensure fair comparisons of outcomes across similar groups of children and, potentially, to use any change in severity as another measure of effectiveness. However, at the time of the review, existing instruments were considered too costly (required skill level and administration time) for large-scale use, and the methodology for developing severity levels based on the project's core data elements was not likely to be achieved quickly. Given the need for rapid development of useful products for agency managers, project staff decided to table the development of severity scales until the basic measurement systems and the software were produced.

Confidentiality of the children's answers to the outcome questionnaire produced the most disagreement. The policy was originally proposed to ensure honest answers to outcome questions. But some agency executives objected to granting children confidential interviews while in the custody of a treating agency. Their disapproval was primarily due to their desire to use outcome information for clinical purposes. Plus, the staffing patterns demanded by a policy of confidentiality created implementation problems. That is, confidentiality demanded case-neutral or nontherapeutically involved individuals to collect the outcome data if the child were being interviewed. Without confidentiality, caseworkers could collect the outcome information for their own clients. With confidentiality, the caseworker and all others directly involved in the case would not be permitted access to the outcome information except in aggregate form. Case-neutral persons could be caseworkers or other workers involved with children, but not with the children for whom they are collecting outcome information. Such sharing of responsibility between involved and noninvolved workers is common in the field of peer review. As is discussed below, the policy of confidentiality was ultimately adopted when, after four rounds of testing, it was demonstrated that confidentiality was essential in order to obtain honest answers from the children.

After revising the instruments, two levels of further testing were conducted. The first level was carried out in three rounds involving staff from three different agencies in each round. After a thorough review of the child and caregiver structured interview instruments and an opportunity to use each instrument in a role play interview, staff completed a questionnaire about each instrument. They answered questions such as (a) How clear are the instructions you will be expected to follow when you use these instruments?, (b) How clear is each question in the interview?, (c) How easy will it be for your clients (children and caregivers) to answer these questions?, (d) Will your clients be able to give real answers to these questions or will they say "don't know"?, and (e) How likely are your clients to give honest answers or will concerns about negative consequences affect their

responses? Changes were made to the child and caregiver interviews following each test round for use in the next round. The final versions of the instruments took between seven and nine minutes to complete.

After completing the first level of testing with agency staff, participating members were sent revised instruments with introductory scripts for their use in a second level of testing, which was gathering feedback from a sample of child clients and respective caregivers. These respondents had been selected during the earlier rounds by use of a table of random numbers against numbered caseload lists. The questions agency staff asked were similar to those the children and their caregivers had answered earlier (i.e., clarity of the questions, ease of answering, likelihood of honest answers). Those responding included 37 children and 38 caregivers, which represented 97% and 95% of each sample, respectively.

This dual-level structured testing process allowed for the identification and correction of poorly worded instructions, questions, and other troublesome concepts with the instruments. For example, project staff members' earlier concern that confidentiality would be essential to obtain honest answers was confirmed by this process. Significant proportions of agency staff, child clients, and caregivers all indicated that some questions were not likely to be answered honestly unless the children were granted freedom from consequences. When the issue was revisited, administrators' concerns were eased in two ways. First, agency staff indicated that they, in essence, already knew the answers to the outcome questions, at least while children were in their care, so administrators were assured that important information of clinical significance would not be missed by granting confidentiality in exchange for a better chance at honest answers. Second, new instructions included stronger language to alert the interviewer and the child that any response that suggested the child may be at risk of harm could not be held in confidence. This dual-level testing process also revealed that the outcome-gathering instruments were best used with children aged seven or older, since participating staff consistently expressed the view that younger children would not be able to answer the questions.

Integrating and Launching the System

Following the testing of the child descriptors instrument and the outcome protocol, the combined set was converted into a PC-based database software system. Members of both development groups learned how to use the software in laboratory settings. Each agency received the software on computer disks to test at its own pace in its own setting. System modifications were made as user feedback was gathered from the on-site tests. Revised SumOne for Kids software and support manuals were distributed to the 17 agencies (those involved in the two development groups plus PRS) for their own use. This initial software, however, was not as flexible as was needed; the debugging process was time consuming and difficult. Although standard reports could be generated easily, they lacked the visual appeal that would encourage use of the data. At the same time, a change of leadership within PCCS led to lessened interest in the project. Finally, increasing amounts of time and costs were being devoted to the complex technology needed to summarize data and develop the central database. It appeared that SumOne for Kids was to be a prematurely completed project rather than a fully functioning outcome-monitoring system.

PRS realized that software improvement, ongoing enhancements, central database management, and distribution of SumOne for Kids would require specialized, technological expertise and resources. Instead of ending the project because of these barriers, PRS pursued additional foundation support to create a separate corporation, the Corporation for Standards & Outcomes (CS&O), for the continuation of SumOne for Kids. CS&O's sole purpose became the pursuit of technologically advanced methods for gathering, summarizing, and sharing information regarding the outcomes of human services. With its focus on technological improvements, CS&O redesigned the SumOne for Kids software to maximize flexibility within a few months' time; they began a replication of the initial project in Maryland and now have replications in Pennsylvania and other states.

Ensuring Data Integrity

As attention turned to implementation issues, the concern for data accuracy became paramount. Accuracy had been an explicit issue in the design of the system and drove decisions on such matters as the standard of accuracy in coding case information, confidentiality of client responses, and case-neutral staff. To ensure data integrity with new users, agencies will learn use of the system from packaged training materials that include a testing/certification process for staff involved in the task of coding case information. In addition, an audit program will be added to the outcome-monitoring system as a way to ensure data integrity on an ongoing basis.

CS&O will lead the development of the audit, and they have considered an array of alternatives for structuring it. A comprehensive audit system operating at the central database level will ultimately be necessary. Responsibility for the audit of an agency's data will reside with the agency itself. Agencies will be provided with an audit program that separates the process into an internal audit, conducted by agency staff or consultants, and an external audit, conducted by the agency's hired financial auditing firm as part of its annual engagement.

The internal audit will have two components, both of which will be documented in the agency's annual outcome report. The first component will be the agency's quality control of the outcome system (i.e., verification of compliance with the outcome protocol requirements and corrective actions taken as needed). The second component will be monitoring for data accuracy. Audit information will be compiled as part of the annual agency outcome report.

One or more of three methods will be employed to verify data accuracy. In the first method, an interviewer rating of each outcome interview will ensure that data entered into the automated system meet a certain criterion of believability. At the end of each outcome protocol, the interviewer uses a 5-point scale to rate the following factors, each of which relates to believability: (a) how well the child/caregiver understood the questions, (b) how cooperative the child/caregiver was, (c) how freely and comfortably the child/caregiver answered the questions, and (d) whether any of the responses were not fully believable. Interviews with an overall low rating (i.e., those considered "not believable") will be held in a separate file in the database and are not used as the basis for outcome reports. Thus, outcome reports will be based on information that has passed a level of rated believability. The number of nonbelievable interviews will be summarized in the agency's annual outcome report and will be reviewed as part of the agency's annual audit.

In the second method, a paired-agreement technique is used to monitor data accuracy. When the answers to the outcome questions posed to a child and his or her caregiver are the same (pair agreement), the information contained in both interviews will be considered accurate. When they do not agree, both interviews become suspect. The system will take the lower or worst case response as "the truth" when a child's and caregiver's responses differ. This default will serve to bias the reported data conservatively and will prevent the aggregated agency data from being inflated. A summary of the number of pair agreements/pair disagreements will also be part of the annual agency outcome report and will be reviewed as part of the annual audit.

In the third method, a sample of outcome reports is examined during the internal audit to determine how well the outcome reports reflect the actual responses from clients and caregivers. A case-neutral staff member or consultant will contact randomly selected child and caregiver respondents shortly after a regular protocol interview and query them again on selected elements of the original interview. The results of these tests will be summarized in the agency's annual outcome report and will be reviewed in the agency's annual audit.

The annual audit of financial statements in Pennsylvania is typically conducted by a CPA firm hired by the agency. SumOne for Kids will furnish an add-on audit program to be conducted by the agency's firm. The purpose of this audit program will be to review the agency's compliance with prescribed policies and procedures and their corrective actions, and to review the agency's monitoring of data accuracy. These reviews will serve as an evaluation of the agency's actual experience with collecting and reporting outcome information using the SumOne for Kids system.

Meshing With Concurrent Public Policy Initiatives

Much of the system's statewide implementation will depend on present and future circumstances in the public policy arena. Significant developments at all levels of government will affect the implementation opportunity. One development that has universal application to children's service agencies nationwide is the introduction of the Statewide Automated Child Welfare Information System (SACWIS).[17]

SACWIS is a federal matching grant program that helps states develop and operate comprehensive computer databases that integrate a number of existing state-operated children's service databases. However, SACWIS does not address functional outcomes. The Adoption and Foster Care Analysis and Reporting System (AFCARS) is one database in SACWIS. AFCARS lists its outcome categories as "reason for discharge" (reunification with parent, living with other relative, adoption, emancipation, guardianship, transfer to another agency, runaway, and death) and "date of discharge." Neither is an example of a functional outcome. SACWIS's current requirements do not address impacts or results after discharge, a deficiency addressed by SumOne for Kids. Nevertheless, many of the SACWIS data elements used to describe the population of children and families receiving welfare services will influence subsequent revisions to SumOne for Kids. The database developed through SumOne for Kids will need to be "nested" within the larger databases included in SACWIS to ensure that it is useful to state officials.

Another development at the federal level, in the form of the Government Performance and Results Act of 1993, is the legal requirement that U.S. government agencies must begin reporting on outcomes in their budget submissions. The outcome requirements by the federal government will affect programs at the state and local levels, and eventually all public and private agencies, that receive federal funding. General state legislation is also requiring information on human service outcomes. Furthermore, the Governmental Accounting Standards Board is going to require outcome reporting of all governmental units.[18]

For behavioral health and child welfare programs, which represent the largest and fastest growing costs for states, perhaps the most influential driving force for outcomes assessment is the adoption of managed care. As a means of controlling costs, managed care relies on predetermined standards to limit services to particular populations. Ultimately, standards must be based on the outcomes each service is expected to achieve. Child welfare and behavioral health outcome systems need to respond to funding streams and standard-setting organizations, but in time, these standards will become grounded in actual experience as recorded by outcome-monitoring programs.

Implications for Behavioral Health Services

Administrators of programs serving at-risk children and their families will be turning their attention to outcome measurement as the inducements grow. Whether systems are developed for a single agency, a group of agencies, or for a large system such as a county or state, outcome measurement will be on the frontier of social service evaluation. Ultimately, all efforts are largely idiosyncratic and depend on local circumstances. For outcome system designers at any level, it is far more useful to examine the decisions, assumptions, and principles that underlie the models than to emulate the features of the models. This article attempted to draw attention to some of the salient issues that led to the development of SumOne for Kids. The most important principles and observations to be gleaned from this effort include the following:

1. Involve stakeholders in the development of all aspects of the measurement system, including selection of outcomes.
2. Foster continued commitment to the long development process through any means possible, including group meetings, newsletters, communications regarding mandates for outcomes, and early products.

3. Measure a few things well, keep it simple, and build on what providers are already measuring. The involvement of many stakeholders in the development will naturally result in a system that is somewhat complicated and burdensome initially; realize that through testing and refinement, simplicity will result.
4. Design a system for agency self-evaluation, one that is built into ongoing program operations, and one in which agency staff routinely collect information for midstream correction. Moving toward a culture of outcome accountability requires that providers embrace self-evaluation as part of their program models.
5. Select outcomes based on social validation and ones that are immediately understandable and socially significant for both internal and external audiences. Such an emphasis will produce a greater reliance on functional outcomes.
6. Use rigorous testing procedures in the development of outcome measures, but recognize that naturalistic settings often require great flexibility in research design and methodology. Also, recognize that providers must be involved in product testing.
7. Work toward multiagency adoption of common outcome measures. The power of outcomes is the aggregate of many cases; thus, outcome measurement should increasingly focus on multiagency, multisystem data pools that can provide benchmarks for comparisons and answers to fundamental questions of children's services: which services work better than others, for which children, and at what cost.

Acknowledgments

The success of the SumOne for Kids effort was made possible by the generous funding of the Grable Foundation of Pittsburgh. We wish to also acknowledge the participating agencies for their forward thinking and willingness to participate in the development and testing of its products (pilot agencies noted by asterisks): Auberle Home, Bethany Children's Home*, Children's Aid Society of Somerset County*, Children's Home of Bradford*, Children's Home of Reading, Children's Home of York, Concern*, Craig House, Education Center at D. T. Watson, Friends Association for the Care and Protection of Children, Gannondale, George Junior Republic*, Hoffman Homes for Youth, Holy Family Institute*, Lourdesmont/Good Shepherd Youth and Family Services*, Lutheran Youth and Family Services*, New Life Youth and Family Services*, Northern Tier Youth Services, Perseus House, Pinebrook Services for Children and Youth*, Presbyterian Children's Village*, St. Gabriel's System, St. Michael's School*, Silver Springs-Martin Luther School*, Supportive Child/Adult Network, The Bair Foundation, The Pressley Ridge Schools*, The Wesley Institute*, The Whale's Tale*, Tressler Lutheran Services*, and Women's Association for Women's Alternatives.

References

1. United States House of Representatives Select Committee on Children, Youth, and Families: *No Place to Call Home: Discarded Children in America*. Washington, DC: U.S. Govt. Printing Office, 1989.
2. Nadel MV: *Residential Care*. Report No. GAO HEHS 94 56. United States Government Accounting Office, Health, Education, and Human Services Division, Washington, DC, 1994.
3. Kaplan LW: *Making Quality Count: A New United Way of Allegheny County Program Review and Allocation System*. Pittsburgh, PA: Allegheny County United Way, 1993.
4. Cole RF, Poe SL: *Partnerships for Care: Systems of Care for Children With Serious Emotional Disturbances and Their Families*. Interim Report of the Mental Health Services Program for Youth. Washington, DC: Washington Business Group on Health, 1993.
5. Governor's Select Commission: *Memorandum on the Care of Dependent Children in Pennsylvania*. Commission report (suppl.), 1923.
6. *SumOne for Kids: An Outcomes-Based Measurement Software System*. Pittsburgh, PA: Corporation for Standards & Outcomes, 1994.
7. VanDenBerg J, Beck S, Pierce, J: The Pennsylvania outcome project for children's services. In: Kutash K, Liberton CJ, Algarin A, et al. (Eds.): *A System of Care for Children's Mental Health: Expanding the Research Base*. Tampa: Mental Health Institute, University of South Florida, 1992, pp. 233-238.
8. Fabry BD, Hawkins RP, Luster WC: Monitoring outcomes of services to children and youths with severe emotional disorders: An economical follow-up procedure for mental health and child care agencies. *The Journal of Mental Health Administration* 1994; 21(3):271-282.

9. Usher L: Balancing Stakeholder Interests in Evaluations of Innovative Programs to Serve Families and Children. Paper presented at the annual meeting of the Association for Policy Analysis and Management, Washington, DC, October 30, 1993.
10. Mathison S: What do we know about internal evaluation? *Evaluation and Program Planning* 1991; 14:159-165.
11. Hawkins RP, Fremouw WJ, Reitz AL: A model useful in designing or describing evaluation of planned interventions in mental health. In: McSweeny AJ, Fremouw WJ, Hawkins RP (Eds.): *Practical Program Evaluation in Youth Treatment*. Springfield, IL: Charles C Thomas, 1982, pp. 24-48.
12. Rosenblatt A, Atkisson CC: Assessing outcomes for sufferers of severe mental disorder: A conceptual framework and review. *Evaluation and Program Planning* 1993; 16:347-363.
13. Wolf MM: Social validity: The case for subjective measurement or How applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis* 1978; 11:203-214.
14. Barth RP: *Social and Cognitive Treatment of Children and Adolescents*. San Francisco: Jossey-Bass, 1986.
15. Gold N: Stakeholders and program evaluation: Characterizations and reflections. In: Bryk AS (Ed.): *Stakeholder-Based Evaluation*. San Francisco: Jossey-Bass, 1983, pp. 63-72.
16. VanDenBerg J, Beck S, Howarth D, et al.: *What Pennsylvanians Want From Children's Services: Summary Report on the Social Validation Study*. Pittsburgh, PA: The Pressley Ridge Center for Research and Public Policy, 1992.
17. Data collection for foster care and adoption; statewide automated child welfare information systems; final rule and interim final rule. *Federal Register* 1993; 58(244).
18. *Service Efforts and Accomplishments Reporting: Concepts Statement No. 2*. Governmental Accounting Standards Board No. 109-A. Norwalk, CT: Government Accounting Standards Board, April 1994.