

Multidimensional Implementation Evaluation of a Residential Treatment Program for Adolescent Substance Abuse

Leyla Faw

National Clinical Assessment Authority

Aaron Hogue

Columbia University

Howard A. Liddle

University of Miami

Abstract: The authors applied contemporary methods from the evaluation literature to measure implementation in a residential treatment program for adolescent substance abuse. A logic model containing two main components was measured. Program structure (adherence to the intended framework of service delivery) was measured using data from daily activity logs completed by program staff. Treatment process, conceptualized as therapeutic milieu, was measured using an adapted version of a scale used to measure implementation in therapeutic communities. In addition, variability in implementation was measured using statistical process control (SPC) procedures. Adolescents completed, on average, 50% of the weekly prescribed services. The milieu of the program was rated by the adolescents as highly therapeutic. Moreover, preliminary psychometrics suggest therapeutic milieu can be measured reliably in adolescents. These two main variables were implemented with consistency across adolescents. These findings are discussed along with implications for evaluation work in similar fields.

Keywords: *implementation evaluation; residential treatment; adolescent substance abuse; therapeutic milieu; treatment process*

Leyla Faw, Evaluation Research and Development Officer, National Clinical Assessment Authority, 1st floor, Market Towers, 1 Nine Elms Lane, London SW8 5NQ, England UK; Phone: +44 (0) 207-084-3808; e-mail: lfaw@ncaa.nhs.uk.

Authors' Note: This research was supported in part by National Institute on Drug Abuse (NIDA) Grants P50 DA11328 and DA0165930 and is based on the doctoral dissertation of the first author. The authors thank Linda Albergia for her assistance collecting and coding the data; Gayle Dakof and Cynthia Rowe for their research contributions to the study; and Kristi Ferraro for her assistance coding the data.

American Journal of Evaluation, Vol. 26 No. 1, March 2005 77-93

DOI: 10.1177/1098214004273183

© 2005 American Evaluation Association

Why Evaluate Implementation?

Recent calls for accountability and better understanding of intervention processes (Dane & Schneider, 1998; Gresham, Gansle, Noell, Cohen, & Rosenblum, 1993; Moncher & Prinz, 1991; Waltz, Addis, Koerner, & Jacobson, 1993) have been answered by a growing emphasis on implementation evaluation (Dusenbury, Brannigan, Falco, & Hansen, 2003; Hogue et al., 1998; Orwin, 2000; Schlosser, 2002). This has included much theoretical literature promoting increasingly sophisticated methods for this implementation of evaluation and progress toward routine inclusion of implementation and process evaluation procedures in reports on treatment and services outcomes (e.g., Dewa, Horgan, Russell, & Keates, 2001; Kam, Greenberg, & Walls, 2003). Evaluation or research that neglects the issue of implementation has been described as no longer acceptable (Mowbray, Holter, Teague, & Bybee, 2003).

At the least, implementation evaluation is critical for documenting program integrity, defined as the degree to which a service is delivered as intended by the program theory (Summerfelt, 2003). Beyond this, two of the most compelling reasons for evaluators to measure implementation are (a) to monitor program activities in order to identify problems in program implementation and (b) to measure variability in program delivery for later use in statistical analysis of program impact (Scheirer, 1994). The current study aimed to apply newly emerging methods from the evaluation literature to measure implementation of a residential treatment program for adolescents with substance abuse problems. The study highlights some of the lessons learned from applying up-to-date methodology and resulting implications for future implementation studies, including those that go on to evaluate outcomes in relation to implementation. The study also advances the framework of implementation evaluation methodology for residential treatment in particular.

How Implementation Is Typically Evaluated in Residential Treatment

Although much theoretical literature exists to guide implementation evaluation (e.g., Mowbray et al., 2003; Scheirer, 1994; Weston, 2004; Yeaton & Sechrest, 1981), there is little evidence that these ideas have been put into practice in the field of residential substance abuse treatment. Evaluations of residential treatment typically measure dose, defined as length of time in treatment or number of sessions attended (for review, see Grella, Hser, Joshi, & Anglin, 1999; Orwin, 2000). Dose has been positively related to outcome for both adolescents (Hser et al., 2001) and adults (Jainchill, Hauke, De Leon, & Yagelka, 2000). As yet, however, no research has explicitly linked individual components of residential treatment theory to client outcomes. The next step in evaluating residential treatment for substance abuse should involve taking a theory-based approach.

Evaluating program implementation by measuring dose alone neglects the concept of treatment process or the theorized mechanism of change. As such, it cannot be determined whether the program theory has received a "fair test" (Durlak, 1998; Scott & Sechrest, 1989). If the process elements of a treatment model are not delivered with integrity, negative or null findings cannot be unequivocally attributed to flaws in the program theory. These findings may instead be attributable to a failure to implement the program theory as it was intended. This has been called Type III error or the failure to identify significant effects due to lack of control over the independent variable (Durlak, 1998). Theory-based evaluation addresses this point by requiring clarification, operationalization, and measurement of the processes proposed to induce desired changes.

Crew and Anderson's (2003) evaluation of charter schools in Florida illustrates the benefits of theory-based evaluation. First, the theory behind charter schools was clearly laid out, and the logic model was then operationalized and measured. The study found that charter schools did not have the positive impact on the overall school system that had been expected, and this was mainly attributable to a failure to implement one critical element of the charter school program theory. Specifically, charter schools were not able, through superior performance, to create new forms of accountability for public schools with whom they were in direct competition. This represented a breakdown in the logic model. The evaluation provided clarity around the next steps that would need to be taken for charter schools to attempt to improve their outcomes—they must find ways to create new forms of accountability for public schools in order for an improved education system to be achieved. If accountability were established, and the outcomes for the education system remained negative or unchanged, this would imply the charter school program theory itself is flawed.

How Implementation Evaluation in Residential Treatment Should Be Improved

Implementation evaluation in residential treatment should begin to take a theory-based approach so that the program theory can undergo a “fair test.” This will include developing a logic model that includes treatment processes that can be operationalized and measured. Furthermore, multiple methods and multiple sources should be used to match the complex nature of residential treatment theory (Scheirer, 1994). Finally, variability in implementation must be measured because programs are likely to experience changes over time. Any substantial variance in the independent variable must be addressed to provide statistical power to later outcome analyses (Lipsey, 1998; Summerfelt, 2003).

Mowbray et al. (2003) offer a conceptual model for sophisticated implementation evaluation such as that described above. Specifically, they suggest measuring elements of both structure (the framework for service delivery) and process (the way in which services are delivered). This reflects the multimethod approach that is now considered essential to evaluation of programs with multiple components (Forsetlund, Talseth, Bradley, Nordheim, & Bjorndal, 2003; Orwin, 2000; Scheirer, 1994) and offers a practical way to organize the main components of any theory-based program. Mowbray's model maps exactly on to the goals of the current study that aimed to measure both the framework of service delivery in residential treatment, that is, what was delivered and for how long, and the content of the service that was provided. Mowbray et al. (2003) discussed the fact that implementation evaluations do not often include both structure and process elements.

Applying Up-to-Date Methods for Implementation Evaluation

The current study attempted to apply Mowbray's model to an implementation evaluation of a residential treatment program for adolescent substance abuse. Structure and process were conceptualized as the two main categories within which elements of the program theory would fit. The first step was to clarify the theory of the residential treatment program. The Adolescent Treatment Program, or ATP, is a modern-day residential program located in an inner-city area for adolescents with substance abuse problems. ATP is based on a social learning approach that emphasizes positive reinforcement for appropriate coping behavior and social skills. ATP shares many of the core assumptions of other residential treatments. For example, the ATP program theory assumes it is essential to separate the adolescent from both the home environment

and the school environment for a period of time. The program also makes the assumption that the adolescent must be provided with a warm, accepting, empathetic context that features positive models for parental and authority figures. ATP follows a schedule of service delivery, whereby adolescents are to complete a regimen of therapeutic activities each week. These services include individual and group therapy, vocational training, educational activities, and recreational therapy. Two main aspects of the program are theorized to lead to individual change: the weekly prescribed schedule of therapeutic activities and the maintenance of a therapeutic milieu in the program. These variables represent structure and process at ATP and as such were the key elements of the program theory that were measured for the study.

Step 2 involved operationalizing these two main components of the program theory. Structure was conceptualized based on the work of Holland (1986), who piloted a model to measure adherence to the service delivery framework of residential substance abuse treatment programs that follow the therapeutic community model (De Leon, 2000). Her model separates treatment services into five distinct categories: treatment activities (e.g., individual and group therapy), functional activities (e.g., resident and staff meetings), productive activities (e.g., educational programs outside the community), reentry activities (e.g., vocational training), and interpersonal activities (e.g., recreational therapy). Holland recommends calculating the percentage of a client's total time spent per week involved in activities falling into these five categories to provide a simple but useful summary of program activity. Holland's model represents a significant advance over traditional evaluation of residential treatment implementation in two ways. First, it allows for comparison of actual service provision to intended service provision, instead of restricting the analysis to the total number of services provided or the dose of treatment received in terms of days completed. Second, the model is specific to residential substance abuse treatment and generalizable to any study focusing on this type of treatment.

Process was operationalized as therapeutic milieu at ATP. Residential treatment settings often incorporate a therapeutic milieu, which serves as an overall environmental factor that facilitates change in the residential clients. There are presumably different aspects of therapeutic milieu that might drive its impact, but these have yet to be researched. Although therapeutic milieu was included in evaluations of residential treatment in the 1970s and 1980s, it has not been the focus of more recent evaluations. Early evaluations found positive correlations between patients' perceptions of therapeutic milieu and outcome. Study samples included substance-abusing adolescents receiving outpatient treatment (Friedman, Glickman, & Kovach, 1986), adults in state hospital wards (Klass, Growe, & Strizich, 1977), and incarcerated men in an enclosed psychiatric facility (Alden, 1978). Patient ratings of therapeutic milieu have also been correlated with retention in residential treatment for substance-abusing adults (Verinis & Flaherty, 1978). These studies were correlational, and potential confounds should be acknowledged; for example, clients with fewer symptoms or a more positive orientation toward treatment may perceive the milieu as more therapeutic and experience more positive outcomes. However, given that both client and staff ratings of therapeutic milieu have predicted outcomes (Friedman et al., 1986; Main, McBride, & Austin, 1991), and given the emphasis of therapeutic milieu in the program theory, we felt this was the clear choice for measuring treatment process in the current study.

The third step in developing the implementation evaluation model was to establish a method to measure variability in implementation. Recent methods for measuring variability at the program level include the use of probability charts to compare means and standard deviations within a sample or program (Weersing & Weisz, 2002). Statistical process control (SPC) methods provide an efficient way to perform this type of analysis. Implementation scores for participants are plotted on a chart and then inspected for patterns that would be unlikely to occur by

chance. Although SPC was originally developed primarily in industrial settings as a way to monitor product output, the method has recently been used in mental health settings to measure consistency of service delivery at outpatient mental health centers (Green, 1999), to analyze variations in staff performance in residential treatment programs for adults (Dey, Sluyter, & Keating, 1994), and to analyze baseline variability in applied behavior analysis (Pfadt, Cohen, Sudhalter, Romanczyk, & Wheeler, 1992).

The main benefit of using SPC rather than more traditional measures of dispersion is SPC's ability to provide information on the nature of variance. In program evaluation, certain patterns of variance in implementation are not always detrimental. For example, when implementation scores are significantly above the mean for clients who entered the program around a certain date, these cases can be isolated by SPC, and the evaluator can then analyze what was happening at the program during that time that may have contributed to better implementation. Causes of low implementation can be probed in the same way when scores fall significantly below the mean.

Summary of Aims

The principal aim of the study was to use a theory-based approach to measure implementation of a residential treatment program for adolescent substance abuse. The measurement model included two main variables (structure and process) as well as analysis of variability in implementation. The potential benefits of this model include (a) its ability to offer information about implementation that is more specific than what has been obtained in past implementation evaluations in residential treatment and (b) its generalizability to any residential treatment program for substance abuse that follows a therapeutic community model. Moreover, the overall framework of the evaluation (structure, process, variability) is applicable to any implementation evaluation.

Method

Participants

The study used data from 43 consecutive admissions to the ATP as part of a federally funded study of treatment options for adolescents with substance abuse problems in Dade County, Florida. ATP is an 8- to 10-month program, but not all participants completed the full course of treatment. Inclusion criteria were as follows: (a) between the ages of 13 and 17; (b) dually diagnosed meeting *DSM-IV* criteria for substance abuse or dependence and major depressive disorder, bipolar disorder, conduct disorder, or oppositional defiant disorder; (c) currently living with at least one parent or parental figure (e.g., aunt, uncle, grandparent, etc.); and (d) provided informed consent from a formal guardian and assent from the adolescent. Adolescents were excluded from participating in the study if they demonstrated suicidal intent or had schizophrenia, mental retardation, eating disorder, or pervasive developmental disorders. Data were used from all residents who received at least 1 week of treatment. The final sample consisted of 43 adolescents (average age = 15.4 years) who enrolled between May 1999 and October 2002. Thirty-two were male (74%), and 11 were female (26%). Adolescents identified themselves as Hispanic ($n = 31$, 72%), African American ($n = 7$, 16%), Caucasian ($n = 3$, 7%), American Indian ($n = 1$, 2%), and Haitian ($n = 1$, 2%).

Measures

ATP daily activity logs. ATP daily log data included one SPSS file for each adolescent enrolled in the ATP program. ATP daily logs were completed by all ATP staff members who provided services to the adolescent during a routine program day, including basic living services (e.g., meals, school, hygiene), therapeutic services (e.g., therapy sessions, milieu groups, and psychological and psychiatric consultations), and recreational services. Daily logs are routinely completed at ATP, that is, this was not introduced as a feature of the randomized clinical trial. The amount of time spent in each contact, the general goal of the contact, the identity of the staff member involved, and any notes or clinical observations were logged. Data from paper-based logs as described above were entered into SPSS in cells divided into columns that represented (a) the date, (b) the service type (e.g., group therapy, school, family session, etc.), (c) the times the service began and ended on that day, and (d) the duration of the service in minutes.

ATP daily logs often included systematic duplication errors as well as data entry errors such as incorrect service duration estimates and inclusion of services that did not occur. A process for cleaning the data was developed for the current evaluation. By comparing the raw data (from paper logs completed daily by staff at ATP) with the data that were entered into SPSS for each participant, inaccuracies such as those listed above were identified. These inaccuracies were then corrected in the SPSS files to ensure that the SPSS data matched exactly with the raw data. This somewhat painstaking process resulted in SPSS files that were more accurate with respect to the adolescents' daily activities during their time in treatment. The entire log was used for each adolescent, regardless of the amount of time spent in treatment. Although it may have been possible to establish an estimate of implementation for a given point in the program's history using a smaller amount of data, it was important in the current study to include as much data as possible given the nature of the SPC methods used and given that one goal of the study was to illustrate how implementation can be measured as an ongoing process throughout the life of a program.

ATP Environment Scale. The ATP Environment Scale consists of 21 items that measure therapeutic milieu. The scale was completed by the adolescent clients once a week while they were enrolled in the program. In addition to the priority that multiple raters of implementation variables (e.g., staff versus clients) be used, the decision to have adolescents rate the milieu was supported by studies revealing a correlation between therapeutic milieu and outcome based on client ratings of therapeutic milieu (Alden, 1978; Klass et al., 1977; Friedman et al., 1986).

Nineteen items were taken directly from the Community as Therapeutic Agent (CTA) subscale of the Survey of Essential Elements Questionnaire (SEEQ; Melnick & De Leon, 1999). The SEEQ measures adherence of drug treatment programs to therapeutic community principles. The SEEQ consists of 139 Likert-type items and can be administered to both program directors and clients. The SEEQ was drawn on because its CTA subscale measures a concept similar to therapeutic milieu. High interrater reliability across 59 program directors has been found for the CTA subscale ($r = .94$; Melnick & De Leon, 1999).

The items on the ATP Environment Scale represent various aspects of the therapeutic milieu construct such as supportive feedback by peers; client involvement in the residential community; supervision of contact with individuals outside the residential treatment program; and incorporation of behavioral norms, written rules, disciplinary actions, and use of privileges and sanctions. Items taken from the CTA subscale were altered in three ways. First, the reading level was lowered for adolescents. Second, 10 items that were not directly relevant to ATP were deleted. For example, the item, "Clients confront the negative behavior and attitudes of each other and the community," was not included because confrontation is not a feature of group

Table 1
Holland's (1986) Treatment Subcategories Applied to ATP Prescribed Services

Services Provided at ATP	Prescribed Number of Hours per Week for ATP Services	Holland's Treatment Categories	Total Prescribed Weekly Hours of Holland's Categories
Open-ended groups	4	Treatment activities	14.5
Family counseling	.5		
Individual counseling	1		
Chemical education	2		
Substance abuse groups	2		
12-step meetings	5		
Community meeting	2	Functional activities	2
Educational services	30	Productive activities	30
Specialty groups, for example, anger management skills training, problem solving, assertiveness training, relapse prevention	8 total	Reentry activities	8
Recreational therapy	20	Interpersonal activities	20

Note: ATP = Adolescent Treatment Program.

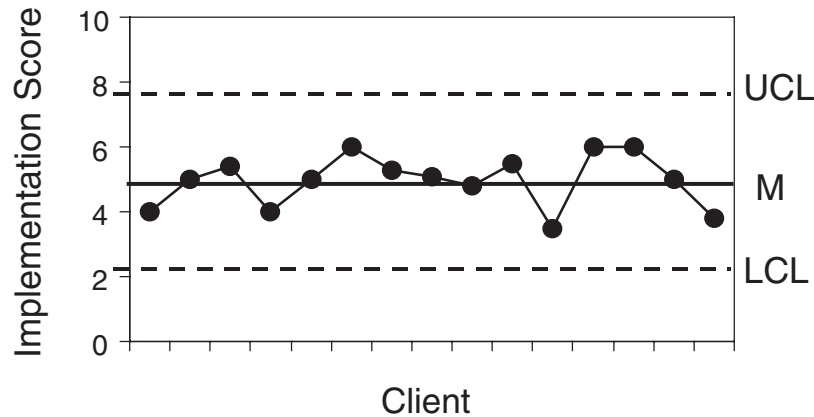
activities at ATP. Third, two items were specifically added to measure implementation of ATP's behavioral reward system, a featured component of the treatment model. ATP Environment Scale items are rated from 0 to 5 with respect to the degree to which the client believes each item describes the program.

Measuring Structure: Holland's (1986) Model

Structure, or the framework of service delivery, was measured using Holland's (1986) model. First, services provided at ATP were placed into the five main treatment categories (treatment, functional, productive, reentry, and interpersonal). Next, using a list of the prescribed number of weekly hours for each service provided at ATP, the number of hours fitting within each of Holland's categories was determined as demonstrated in Table 1. Assignment of ATP services to Holland's categories was completed by the first two authors based on consensus. There were no disagreements during the first round of assignment.

The model was applied to daily log data for each individual by implementing a five-step process: (a) recoding the original daily log data into Holland's treatment categories, (b) computing the total number of hours spent in each service category per week for each week the individual was in treatment, (c) computing the percentage of prescribed activities completed within each category for each week, (d) averaging those percentages to result in one score for percentage of prescribed services completed (across all five categories for each week), and (e) averaging all weekly scores across the individual's time in treatment to obtain one summary score for that person. The final score thus represented the average percentage of prescribed services actually provided on a weekly basis for the adolescent, across all five categories of treatment services, and during the course of the adolescent's stay in the program. If the resulting score for an individual were .62, this would indicate that he or she completed, on average, 62% of the weekly services prescribed at ATP. One score was computed for each adolescent, and these scores were then averaged to compute an overall program score for ATP.

Figure 1
Sample SPC Chart



Note: SPC = statistical process control; UCL = upper control limit; LCL = lower control limit.

Measuring Process: Therapeutic Milieu

For each adolescent, all weekly scores on the ATP Environment Scale were averaged to compute one score reflecting how therapeutic that person found the milieu at ATP during their time there. These individual adolescent scores were then used for analysis of the overall program. The sample size was reduced to 27 because the scale was introduced into the study after 16 adolescents had already exited the program.

Measuring Variability in Implementation

Variability in implementation across adolescents was measured using SPC procedures available through the program Minitab (Minitab, 2000). In SPC, samples are taken from the process under investigation (e.g., therapeutic milieu) and plotted on a chart such as the one shown in Figure 1. The data are then examined for patterns that suggest systematic variance within the sample. Points are inspected in relation to the sample mean as well as upper and lower control limits (labeled as UCL and LCL in Figure 1). These control limits represent three standard deviations above and below the sample mean. There are preset criteria for patterns that would indicate systematic variance, such as one point falling outside the control limits. These criteria have developed throughout the history of SPC as used in industrial settings and reflect patterns of data that are unlikely to occur by chance (Hoyer & Ellis, 1996). SPC charts were generated for both implementation variables (structure and process). The points on the charts represent individual adolescents' scores on these variables.

Results

Preliminary Statistics

Preliminary analyses were run to examine the potential effects of demographic variables and the duration of time spent in treatment on implementation data. Demographic variables were

first considered in relation to the adolescents' scores on adherence to the weekly ATP service delivery framework, computed using Holland's model. An independent samples *t* test showed that the mean score for girls ($n = 11$, $M = .48$, $SD = .12$) was not significantly different from that for boys ($n = 32$, $M = .50$, $SD = .11$; $t = .47$, $p = .64$). Regression analysis revealed there were no age effects ($\beta = .09$, $p = .56$). Next, one-way ANOVA was used to test for differences in adherence scores across ethnicity. Five groups were represented (Hispanic, $n = 31$; African American, $n = 7$; Caucasian, $n = 3$; American Indian, $n = 1$; and Haitian, $n = 1$). Results were nonsignificant ($F = 1.96$, $p = .12$). To compensate for low power in the ANOVA due to small subgroup sizes, a *t* test was also run to test for differences in adherence scores between Hispanics ($n = 31$) versus non-Hispanics ($n = 12$). Results were nonsignificant ($t = .76$, $p = .45$).

The same demographic variables were also considered in relation to ATP Environment Scale ratings. One score, averaged across all weeks in treatment, was used for each person. Using identical methods to those described above, differences were not observed in terms of gender ($t = -.54$, $p = .60$), age ($\beta = .09$, $p = .67$), or ethnicity ($F = .60$, $p = .76$; $t = -.08$, $p = .94$).

A regression analysis was conducted to look for any relationship between total time spent in treatment and implementation scores. The average number of weeks completed by the adolescents was 16.5 ($SD = 10.3$), with a range from 1 to 33 weeks. Total number of weeks was entered as the predictor variable with implementation scores as the criterion variable. Results from the regression analyses were nonsignificant for both adherence (Holland) scores ($\beta = .14$, $p = .39$) and ATP Environment Scale scores ($\beta = .09$, $p = .64$). These results provide evidence that implementation scores were not significantly affected by the amount of time the adolescents spent in treatment.

Program Structure

All individual scores on program structure, computed using Holland's formula as described previously, were averaged. The resulting program score across all participants was .50 ($SD = .11$), suggesting that ATP provided, on average, 50% of the prescribed services across the five categories each week. To provide a more detailed picture of program structure, program scores were also computed for each category. This was done by first computing each person's average score for each category across weeks spent in treatment and then averaging the category scores across all individuals. Results demonstrated that, on average, adolescents completed 61% of the weekly prescribed amount of treatment services, 47% of the prescribed amount of time in functional activities, 63% of the weekly prescribed productive activities, 60% of the prescribed number of reentry activities, and 15% of the prescribed number of hours of interpersonal activities. In general, the overall program score (.50) accurately reflects the amount of prescribed services provided in the first four categories. The fifth category, interpersonal services, consisted solely of recreational therapy, such as field trips, sports, and so on. This may have been difficult to implement due to the large number of prescribed hours (20 per week) in relation to the large number of hours of other services expected to be provided each week.

Program Process: Therapeutic Milieu

Preliminary analyses were run to ensure the reliability of the ATP Environment Scale. Internal consistency reliability was examined by computing Cronbach's coefficient alpha for the scale using the average score across weeks in treatment for each adolescent. The scale demonstrated adequate internal consistency ($\alpha = .70$). The alpha remained high when computed using

Week 1 scores only ($n = 22$, $\alpha = .77$). Test-retest reliability was subsequently examined by computing a correlation between Week 1 and Week 2 scores for all participants for whom such scores existed ($n = 19$). For the total scale, the correlation was significant ($r = .52$, $p < .03$).

Individual ATP Environment Scale scores ($n = 27$) were averaged to compute a program score for therapeutic milieu. Individual scores used for this analysis were previously determined by computing the average score across all weeks in treatment for each adolescent. The average score across the 27 clients was 3.79, based on a scale from 0 to 5, suggesting that, on average, the adolescents rated the environment of the program as moderately to highly therapeutic on a week-to-week basis.

Variability in Implementation

SPC was used to measure variability in both structure and process at ATP. Two types of SPC charts were generated for each variable—*mean* charts and *range* charts. Mean charts reflect a sample's tendency toward a mean value, whereas range charts reflect the amount of dispersion in a sample. In mean charts, each participant's mean score on a given variable is plotted on a graph. The center line, drawn horizontally through the graph, represents the average score of the sample. In range charts, a range of scores, computed as a person's highest score minus their lowest score, is plotted on the graph for each participant. The center line represents the average of these range scores across all participants. In both mean and range charts, UCLs and LCLs signify three standard deviations in either direction of the sample mean.

Although SPC does not yield inferential statistical results, it was well suited for the moderately sized sample, the nature of the data collected, and the purposes of the current study. SPC is particularly useful for implementation evaluation because it allows the evaluator to pinpoint anomalies in service delivery that might indicate a systematic flaw in the process. It is also useful for creating program implementation histories for ongoing and formative evaluation, as recommended by Orwin (2000). Logically, adolescents at ATP should receive an average amount of the prescribed services. Although occasional outliers would be expected, larger patterns of outliers would suggest that the program's service implementation should be further investigated. Cases identified as out of range can further be inspected to diagnose the cause of parametric failure (e.g., program breakdown, data collection problem, staffing changes, etc.).

The Minitab software package specifies several patterns that would indicate significant variability in a given set of data. For the current study, the following criteria were chosen: (a) One point falls more than three standard deviations from the center line, (b) nine consecutive points are on the same side of the center line, (c) six consecutive points are all increasing or decreasing, (d) four out of five consecutive points are more than one standard deviation from the center line (all on the same side), and (e) eight consecutive points fall more than one standard deviation from the center line (on either side). These criteria were the most relevant for the questions of the study. If any of these criteria were met, the data would be subjected to further inspection.

Variability in structure at ATP. SPC charts were generated to examine variability in structure, that is, adherence to service parameters across the 43 adolescents who entered ATP during the study. By simply eyeballing the charts, inferences can be drawn regarding changes in the structure of the residential program during the course of the study. For example, peaks and valleys in the chart would signify times of higher and lower program structure. When a group of points on a chart follows such patterns, two courses of action can be followed. First, characteristics of the program that may have contributed to these patterns can be sought by program evaluators with access to this kind of information, for example, staffing changes, number of adoles-

cents enrolled, changes to the physical environment, and so on. Second, by changing the unit of analysis to weekly scores for one individual, additional SPC charts can be generated to examine the pattern of implementation for each adolescent represented in the group of anomalous data on the program chart. This allows for a more in-depth picture of implementation across time in treatment for these adolescents and may offer clues as to why their scores were inconsistent with the rest of the sample, for example, sharp improvement or decline in scores. These methods can assist in the exploration of the characteristics of both the program and the adolescents that might have contributed to variability in implementation.

The SPC mean and range charts for implementation of program structure at ATP are shown in Figure 2. The points on each chart flow consecutively across time. The first point represents the mean score of the first adolescent to enter the study, whereas the last point represents that of the last adolescent to enter the study during the course of 42 months. UCLs and LCLs represent three standard deviations above and below the sample mean. The mean chart shows that the chart met the fourth criteria chosen for the study, which involves four out of five points in a row falling beyond one standard deviation of the sample mean. Specifically, the point labeled 4 in Figure 2 is the final point in a set of four points that fall more than one standard deviation below the center line.

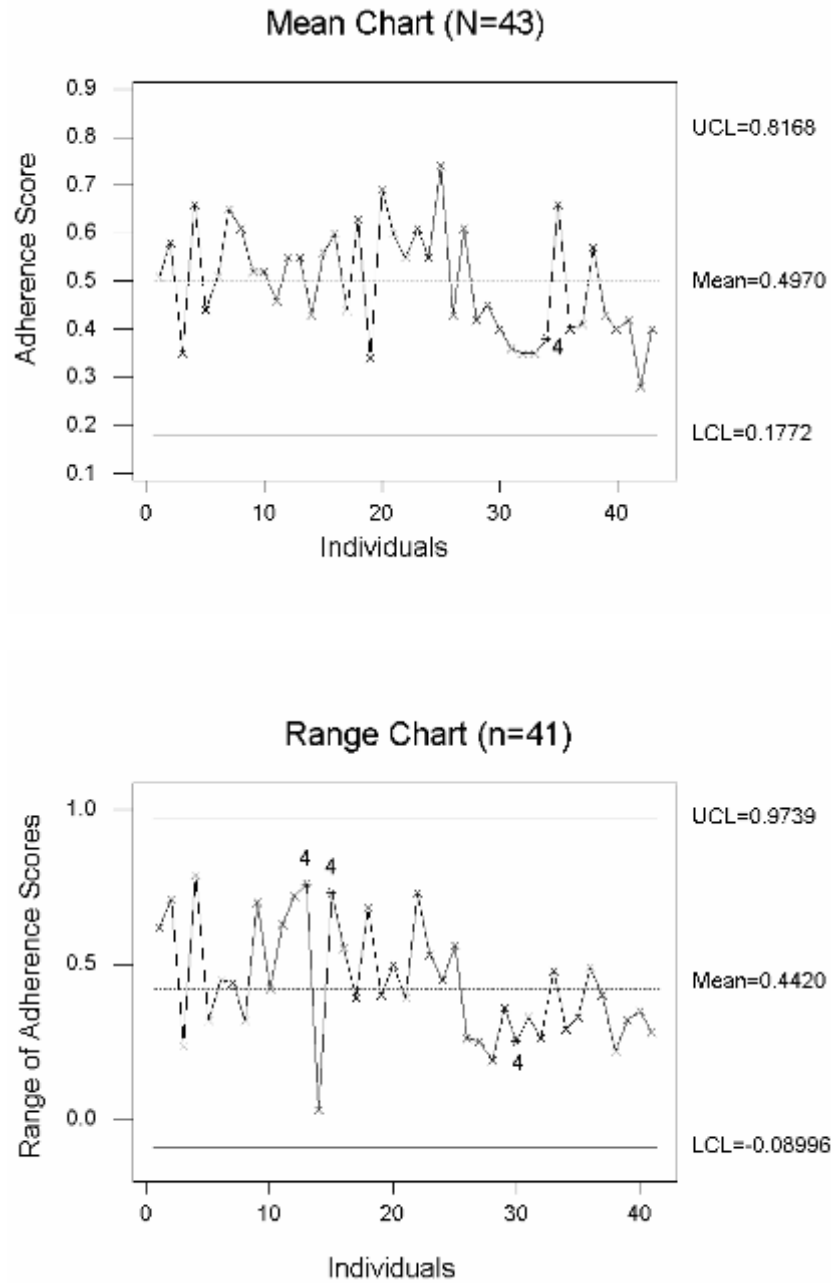
To examine this anomalous subset of scores more closely, mean charts were generated for the individuals represented by these four points. The unit of analysis (i.e., samples plotted) was changed to weekly scores for one person across his or her time in treatment. This prevented range charts being generated for these individuals because each point on the chart represented one weekly score instead of an average of weekly scores as was the case for the program chart. In individual mean charts, the center line represented the average score across all weeks in treatment (M) for the individual. For SPC analyses of individual cases, control limits were set at three standard deviations above and below the person's average weekly score.

Analysis of the SPC charts generated for the four individuals with unusual scores revealed that structure scores for one person were inconsistent across 26 weeks in treatment, with no recognizable pattern emerging. Several criteria were met to suggest significant variability in this person's scores across time. There were no anomalies in the score patterns for the other three cases. A new program mean chart was then constructed with this outlier case removed, and the resulting chart revealed no anomalous pattern of scores. These analyses show that ATP program structure was consistent across adolescents in the sample, with one notable exception.

The range chart (see also Figure 1) identified three clusters of scores that met Criterion d. Individual mean charts were generated for each adolescent represented in the three clusters. Patterns of systematic variability appeared for two of the adolescents in the first cluster. For one of these individuals, the SPC mean chart showed low adherence to the intended program structure in the early weeks, followed by a sharp rise in scores in the late weeks, revealing a pattern of improvement over time. The other outlier revealed the opposite pattern of change over time: high scores early in treatment followed by a sharp decline late in treatment. In summary, in the first cluster, the scores for two adolescents showed no significant variability, whereas the scores for two other adolescents did demonstrate such variability. Upon further inspection, one of these latter adolescents demonstrated a sharp increase in adherence to program structure, whereas the other adolescent demonstrated a sharp decrease in adherence to program structure.

In the second cluster of deviant scores identified in the program range chart, only one adolescent demonstrated significant variability in adherence to program structure across weeks in treatment. This person's scores improved dramatically over time in treatment. Finally, in the third cluster, two of the four out-of-range individuals completed only 2 weeks of treatment, making it impossible to generate an SPC chart for either of them. The short treatment tenure of these two adolescents may explain the significantly narrow range of their weekly scores. For the

Figure 2
SPC Mean and Range Charts: Structure
(Adherence to the Intended Framework of Service Delivery)



Note: Ranges were not computed for two adolescents who completed only 1 week of treatment. Upper control limit (UCL) is 3 standard deviations above the mean, and the lower control limit (LCL) is 3 standard deviations below the mean. 6 = criteria (6) met for data to be considered inconsistent. SPC = statistical process control.

two adolescents in the cluster for whom mean charts could be generated, one person demonstrated inconsistent scores across time in treatment. This was the same adolescent whose overall mean score influenced the program mean scores to appear inconsistent on the mean chart for the ATP program. This case demonstrated no pattern of improvement or decline, appearing instead to have experienced erratic changes in adherence to program structure during weeks spent in treatment.

Overall, then, for the program range chart, there appeared to be five cases that caused the data to appear inconsistent: one whose scores dramatically improved during time in treatment, one whose scores dramatically declined during time in treatment, two who completed only 2 weeks from which to compute a range of scores, and one whose weekly scores showed no identifiable pattern across time. As such, the story for the range scores on ATP structure was similar to the story for the mean scores. With the exception of a few cases, ATP was able to adhere to its intended structure, or framework, for service delivery. When the data appeared inconsistent, specific causes could be assigned (e.g., one adolescent with an erratic sequence of scores). There was no evidence, based on implementation patterns across the whole sample and investigation of outlier cases, that the quality or consistency of the program suffered any breakdown over time or serious lapse across cases.

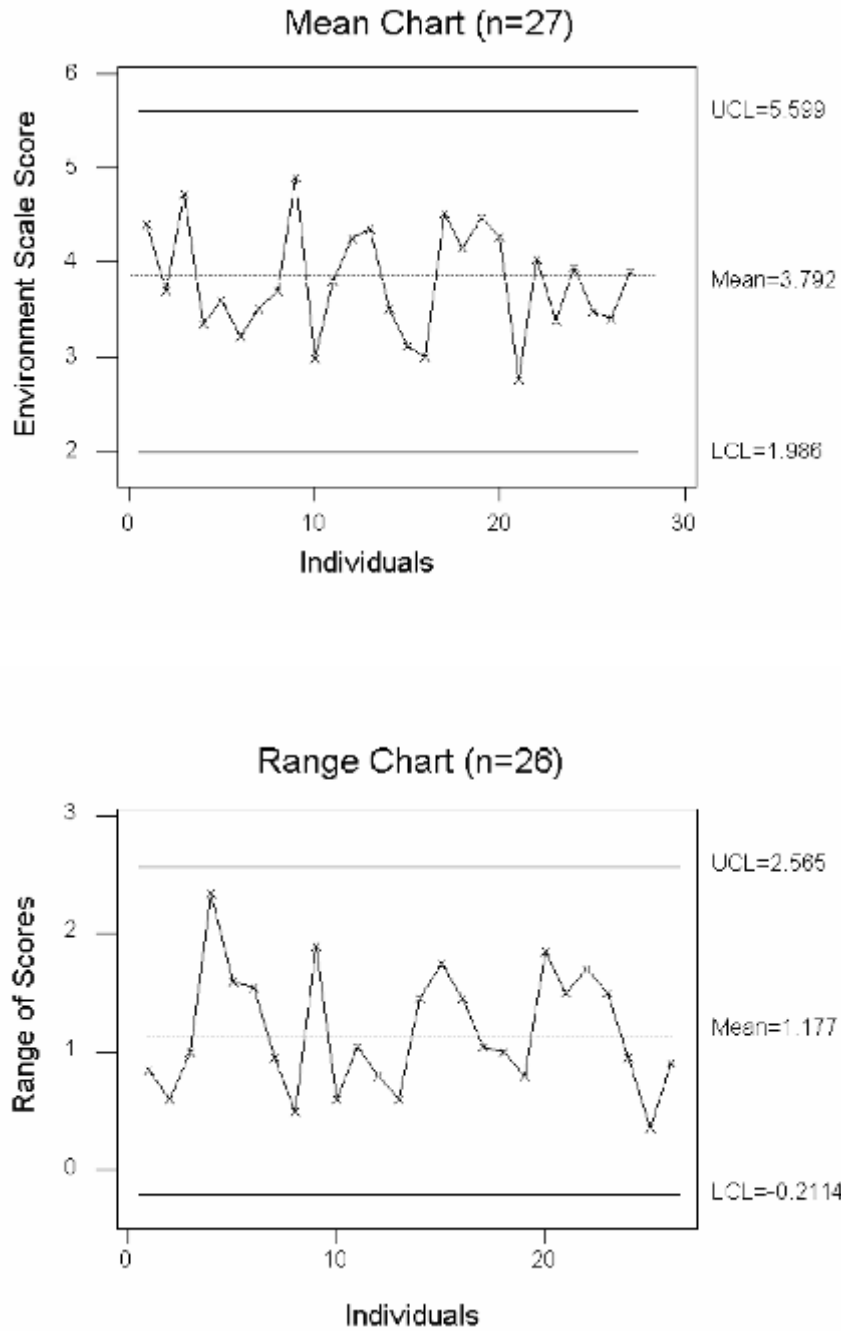
Variability in process at ATP. The program-level SPC mean and range charts for ATP Environment Scale scores are shown in Figure 3. Both charts indicate that implementation of therapeutic milieu was consistent across participants. These findings speak well to the stability of the scale as well as the program's success at maintaining a consistent therapeutic milieu across cases.

Discussion

The study applied a theory-based method to measure implementation of a residential treatment program for adolescent substance abuse. On the whole, the results of the study provide a rich picture of what actually happened during the course of program implementation, in the context of what was intended by the program model. The measurement model incorporated state-of-the-art theory and methods from the program evaluation literature. Following Mowbray et al.'s (2003) model, both structure and process were operationalized and measured. In addition, the consistency of these two variables across adolescents was measured. This approach is generalizable to any residential treatment program for substance abuse that aims to provide prescribed doses of certain activities and that aims to maintain a therapeutic milieu as part of the treatment theory. Moreover, the methods used to measure variability in implementation can be applied in real time for formative assessment as well as retrospectively for summative assessment. For example, SPC can be applied in real time to notify an evaluator when implementation falls below a prespecified level or demonstrates systematic inconsistency. This might signify when intervention is needed. SPC can also be applied retrospectively to detect periods in a program's history when implementation was compromised. If the cause were then identified (e.g., the program was above capacity during this period), the program could use this information to help prevent such compromises from reoccurring.

Structure of the program was measured by applying Holland's (1986) model to the framework of service delivery at the program. Results showed that ATP implemented, on average, 50% of the treatment services prescribed by the program theory on a weekly basis. The meaning of this figure is difficult to interpret. ATP's implementation of interpersonal services was substantially lower than that of the other four categories specified by Holland's model. Operationalizing and measuring these other four categories (treatment services, educational

Figure 3
SPC Mean and Range Charts: *Process* (Maintenance of Therapeutic Milieu)



Note: A range was not computed for one adolescent who completed only 1 week of treatment. Upper control limit (UCL) = 3 standard deviations above the mean. Lower control limit (LCL) = 3 standard deviations below the mean. 6 = criteria (6) met for the data to be considered inconsistent. SPC = statistical process control.

services, program meetings, etc.) is likely easier in a residential program because these services are well-defined clinically and thus specifically documented in program logs. If interpersonal services were a more well-defined, or specific, element of the treatment model, the program estimate might have been based on less diffuse data across categories. The only ATP service fitting into this category was recreational therapy, which is less clearly defined than the services fitting into the other four categories. However, of primary importance, the application of Holland's model yielded programmatically meaningful results, suggesting that any evaluator applying a similar model can make his or her own decisions about the successfulness of implementation, with full knowledge of the degree of fit between the program theory and its operationalization.

Treatment process was measured using clients' ratings of how therapeutic they found the milieu of the residential program. Adolescents reported reasonably high levels of therapeutic milieu, indicating ATP was successful in implementing this aspect of the program theory. The fact that treatment process (therapeutic milieu) was implemented to a higher degree than program structure supports the argument that a program with multiple components should be measured using a multidimensional evaluation model (Scheirer, 1994). Divergent implementation findings can help clarify the salience of various program components when they are later analyzed in relation to outcomes. Sophisticated statistical methods, such as structure equation modeling, can help isolate the variance in outcome that is related to the variance in different implementation variables, provided an adequate sample size is available. Multiple regression models can also be of use in isolating the effects of different implementation variables when larger sample sizes are difficult to achieve. This type of analysis can ultimately lead to more refined program models in which more effective components become the featured elements of the model.

The psychometric analyses of the ATP Environment Scale expand the research base on the CTA subscale of the SEEQ (Melnick & De Leon, 1999). Results suggested that therapeutic milieu can be measured with adequate reliability in adolescent clinical samples. The scale showed high internal consistency reliability and high test-retest reliability, albeit with a small sample size. The findings on the therapeutic milieu measure are also important given findings with adults showing a positive correlation between therapeutic milieu and both retention and outcome in residential substance abuse treatment (Bliss, Moos, & Bromet, 1976; Friedman et al., 1986; Verinis & Flaherty, 1978), and findings that positively link retention to outcome for adolescents (Hser et al., 2001; Jainchill et al., 2000). Results from the current study suggest that the CTA subscale might be a promising resource for studying the potential salience of therapeutic milieu for adolescents.

SPC methods were applied to measure variability in implementation. SPC charts provided a picture of the overall pattern of program implementation as well as the patterns of implementation across time in treatment for certain individuals whose data were inconsistent with the rest of the sample. In one such case, further analysis revealed a sharp increase in implementation scores across time in treatment. In this context, variability should not be considered in a negative light. In contrast, the opposite pattern of variability (a sharp decrease in implementation scores) was found in another case. ATP might want to closely examine the course of treatment in this latter case to provide clues as to how to prevent this pattern from occurring in the future.

The study was able to use specific and optimally relevant SPC criteria for determining if significant variability existed in the implementation of important program components. The criteria chosen were based on consideration of the sample size as well as the questions of the study. In future studies that attempt to monitor consistency of implementation scores, the criteria chosen should similarly depend on the goals of the study. If, for example, the goal were to monitor implementation in real time, detecting any possible systematic influences, a conservative

approach might be warranted in which any suspicious pattern of data would be marked for further investigation. When SPC is used to monitor a process in real time, systematic inconsistencies that are identified can be corrected. For a program that simply wished to maintain a certain threshold of implementation for quality assurance purposes, only criteria that include patterns of decreasing scores or outliers falling below the sample mean would be used. In this situation, SPC could be used as a way to identify certain participants or certain periods of time when implementation scores fall below a predetermined threshold so that systematic causes could be sought. This latter example represents a role for SPC that falls more in line with the traditional use of SPC to monitor product output in industrial settings (Hoyer & Ellis, 1996).

The study adds to the program evaluation literature in two ways. First, it provides a practical example of how Mowbray et al.'s general method for comprehensive implementation evaluation can be approached. Moreover, the method used is generalizable to any residential treatment center for substance abuse that follows a therapeutic community model. Second, the study illustrates the utility of measuring variability in implementation. SPC analyses identified patterns of variability and further identified the individuals exhibiting these patterns. This allowed a meaningful examination of variability, in contrast to more traditional methods that only allow binary estimates of dispersion, that is, the results are either reliable or unreliable. Furthermore, the finding that implementation was consistent across participants attests to the validity of the results for the program's mean level of adherence to service parameters and therapeutic milieu. The relative consistency in scores suggests that any adolescent's score could be meaningfully predicted by the program's mean score.

The study had many strengths, above all the focus on a greatly underresearched population (high-risk, adolescent, ethnic minority). Second, the data collected were well suited to the questions of the study. Every minute of each adolescent's activity was logged during his or her stay in the program, and this allowed for strict computations of time spent in specific activities. Furthermore, daily log data were applicable for both across-person analyses and within-person analyses as exemplified most clearly in the SPC analyses. Third, adherence data were collected at the time of program implementation. Previous research has frequently relied on retrospective methods, which can be unreliable. Fourth, implementation data were collected from multiple sources. Various staff members provided daily adherence data, whereas the adolescents themselves provided ratings of therapeutic milieu.

Despite these strengths, the study was limited in important ways. First, therapeutic milieu was rated by adolescents only and not by staff. Past research suggests staff reports may differ from client ratings of residential treatment environments (Friedman et al., 1986; Main et al., 1991). Future studies should improve on the current methods by including multiple raters for each variable. In addition, the lack of prior research on the SEEQ, used to measure therapeutic milieu, must be considered a limitation. The psychometric findings from the current study will be an important contribution to the research base on the scale. Second, client attrition negatively affected the stability of implementation data, of which a significant amount consisted of average scores across time in treatment. At-risk adolescents are notoriously difficult to retain in treatment (Hogue, Johnson-Leckrone, & Liddle, 1999). This is an issue that will require ongoing attention in any future work involving similar samples. Third, the study was not able to examine characteristics of the adolescent participants or the physical environment of the program in relation to implementation findings. Important personal factors such as motivation for treatment may affect an adolescent's ability to engage in the therapeutic milieu. Furthermore, program characteristics, such as staff changes, may play a role in any observed variability in implementation. Future studies should examine the role of such factors in program implementation. Finally, the scope of the study did not allow examination of outcomes in relation to implementation data. This limited the utility of implementation data to affect the residential

program itself and to inform residential treatment program theory. Nevertheless, the study has demonstrated that diverse and thorough implementation data can be collected in a complex treatment environment.

The study carries important implications for future implementation evaluation work. The benefits of performing comprehensive and generalizable implementation evaluation include aiding in dissemination of programs into new environments or populations, as well as increased understanding of how treatment models work when outcome findings are studied in relation to implementation findings. Developing valid and reliable implementation methods represents the first step in these processes. At a more practical level, the study's methodology is applicable in real-world settings.

Although real-world programs may not be as generously resourced as those participating in a randomized clinical trial, any residential program with a quality assurance component that includes electronic logging of clients' activity could feasibly conduct similar analyses with a modicum of data cleaning and a software package such as Minitab that allows for variability checks. The extra analytic staff this would require (i.e., one appropriately trained person) imposes a negligible burden in terms of resource requirements compared to that associated with a randomized clinical trial. In addition, the study's methodology can be applied in real time as a means of formative assessment as well as quality assurance.

By examining both structure and process elements of program implementation in terms of their overall level as well as their variability across participants, the study made the critical first step in subjecting the program theory of residential treatment for adolescent substance abuse to a "fair test" (Durlak, 1998). Continuing to apply and improve methods for implementation evaluation in this way can ultimately enhance the process of ongoing program development. Most important, understanding the effectiveness of treatment for adolescents hinges on the continued development of methods to measure treatment implementation and analyzing these findings in relation to outcomes.

References

- Alden, L. (1978). Treatment environment and patient improvement. *Journal of Nervous and Mental Disease*, 166, 327-334.
- Bliss, F. H., Moos, R. H., & Bromet, E. J. (1976). Monitoring change in community-oriented treatment programs. *Journal of Community Psychology*, 4, 315-326.
- Crew, R. E., & Anderson, R. M. (2003). Accountability and performance in charter schools in Florida: A theory-based evaluation. *American Journal of Evaluation*, 24, 189-212.
- Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, 18, 23-45.
- De Leon, G. (2000). *The therapeutic community: Theory, model, and method*. New York: Springer.
- Dewa, C. S., Horgan, S., Russell, M., & Keates, J. (2001). What? Another form? The process of measuring and comparing service utilization in a community mental health program model. *Evaluation and Program Planning*, 24, 239-247.
- Dey, M. L., Sluyter, G. V., & Keating, J. E. (1994). Statistical process control and direct care staff performance. *Journal of Mental Health Administration*, 21, 201-209.
- Durlak, J. A. (1998). Why program implementation is important. *Journal of Prevention and Intervention in the Community*, 17, 5-18.
- Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research*, 18, 237-256.
- Forsyth, L., Talseth, O. K., Bradley, P., Nordheim, L., & Bjorndal, A. (2003). Many a slip between cup and lip: Process evaluation of a program to promote and support evidence-based public health practice. *Evaluation Review*, 27, 179-209.

- Friedman, A. S., Glickman, N. W., & Kovach, J. A. (1986). Comparisons of perceptions of the environments of adolescent drug treatment residential and outpatient programs by staff versus clients and by sex of staff and clients. *American Journal of Drug and Alcohol Abuse, 12*, 31-52.
- Green, R. S. (1999). The application of statistical process control to manage global client outcomes in behavioral healthcare. *Evaluation and Program Planning, 22*, 199-210.
- Grella, C. E., Hser, Y., Joshi, V., & Anglin, D. (1999). Patient histories, retention, and outcome models for younger and older adults in DATOS. *Drug and Alcohol Dependence, 57*, 151-166.
- Gresham, F. M., Gansle, K. A., Noell, G. H., Cohen, S., & Rosenblum, S. (1993). Treatment integrity of school-based behavioral intervention studies: 1980-1990. *School Psychology Review, 22*, 254-272.
- Hogue, A., Johnson-Leckrone, J., & Liddle, H. A. (1999). Recruiting high-risk families into family-based prevention and prevention research. *Journal of Mental Health Counseling, 21*, 337-351.
- Hogue, A., Liddle, H. A., Rowe, C., Turner, R. M., Dakof, G. A., & LaPann, K. (1998). Treatment adherence and differentiation in individual versus family therapy for adolescent substance abuse. *Journal of Counseling Psychology, 45*, 104-114.
- Holland, S. (1986). Measuring process in drug abuse treatment research. In G. DeLeon & J. T. Ziegenfuss (Eds.), *Therapeutic communities for addictions: Readings in theory, research, and practice* (pp. 169-181). Springfield, IL: Charles C Thomas.
- Hoyer, R. W., & Ellis, W. C. (1996). A graphical exploration of SPC: Part 1—SPC's definitions and procedures. *Quality Progress, 29*, 65-72.
- Hser, Y., Grella, C. E., Hubbard, R. L., Hsieh, S., Fletcher, B. W., Brown, B. S., et al. (2001). An evaluation of drug treatments for adolescents in 4 U.S. cities. *Archives of General Psychiatry, 58*, 689-695.
- Jainchill, N., Hawke, J., De Leon, G., & Yagelka, J. (2000). Adolescents in therapeutic communities: One-year posttreatment outcomes. *Journal of Psychoactive Drugs, 32*, 81-94.
- Kam, C., Greenberg, M. T., & Walls, C. T. (2003). Examining the role of implementation quality in school-based prevention using the PATHS curriculum. *Prevention Science, 4*, 55-63.
- Klass, D. B., Growe, G. A., & Strizich, M. (1977). Ward treatment milieu and posthospital functioning. *Archives of General Psychiatry, 34*, 1047-1052.
- Lipsey, M. W. (1998). Design sensitivity: Statistical power for applied experimental research. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 39-68). Thousand Oaks, CA: Sage.
- Main, S., McBride, A. B., & Austin, J. K. (1991). Patient and staff perceptions of a psychiatric ward environment. *Issues in Mental Health Nursing, 12*, 149-157.
- Melnick, G., & De Leon, G. (1999). Clarifying the nature of therapeutic community treatment: The Survey of Essential Elements Questionnaire (SEEQ). *Journal of Substance Abuse Treatment, 16*, 307-313.
- Minitab, Inc. (2000). *Minitab reference manual release 9*. Lebanon, PA: Sowers Printing.
- Moncher, F. J., & Prinz, R. J. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review, 11*, 247-266.
- Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24*, 315-340.
- Orwin, R. G. (2000). Assessing program fidelity in substance abuse health services research. *Addiction, 95*(Suppl. 3), S309-S327.
- Pfadt, A., Cohen, I. L., Sudhalter, V., Romanczyk, R. G., & Wheeler, D. J. (1992). Applying statistical process control to clinical data: An illustration. *Journal of Applied Behavior Analysis, 25*, 551-560.
- Scheirer, M. A. (1994). Designing and using process evaluation. In J. Wholey, H. Hatry, & K. Newcomer (Eds.), *Handbook of practical program evaluation*. San Francisco: Jossey-Bass.
- Schlosser, R. W. (2002). On the importance of being earnest about treatment integrity. *Augmentative and Alternative Communication, 18*, 36-44.
- Scott, A. G., & Sechrest, L. (1989). Strength of theory and theory of strength. *Evaluation and Program Planning, 12*, 329-336.
- Summerfelt, W. T. (2003). Program strength and fidelity in evaluation. *Applied Developmental Science, 7*, 55-61.
- Verinis, J. S., & Flaherty, J. A. (1978). Using the Ward Atmosphere Scale to help change the treatment environment. *Hospital and Community Psychiatry, 29*, 238-240.
- Waltz, J., Addis, M.E., Koerner, K., & Jacobson, N.S. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology, 61*, 620-630.
- Weersing, V. R., & Weisz, J. R. (2002). Community clinic treatment of depressed youth: Benchmarking usual care against CBT clinical trials. *Journal of Consulting and Clinical Psychology, 70*, 299-310.
- Weston, T. (2004). Formative evaluation for implementation: Evaluating educational technology applications and lessons. *American Journal of Evaluation, 25*, 51-64.
- Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology, 49*, 156-167.