

# Overcoming Challenges in Outcome Evaluations of School Mental Health Programs

Laura A. Nabors, Mark D. Weist, Matthew W. Reynolds

---

**ABSTRACT:** Significant growth and improvement of school mental health programs has occurred in recent years. However, evaluation of outcomes for children receiving these services is needed to provide accountability data and ensure the sustainability of these programs. When designing studies, evaluators must overcome several challenges that may threaten the validity of their conclusions. In this paper, threats or challenges to the internal and external validity of results from evaluation studies are reviewed. Suggestions are provided for overcoming these challenges, in order to encourage future evaluation activities in this developing field and to document the impact of services for youth and their families. (*J Sch Health*. 2000;70(5):206-209)

---

Significant growth and improvement of mental health programs has occurred in US schools in the last 20 years.<sup>1</sup> School mental health programs have reached a level of maturity and complexity to suggest that attention must turn to documenting whether school mental health services are successful in producing positive outcomes for youth. Establishing the merit of such programs is central to their continued growth and presence in schools.<sup>2</sup>

Outcome evaluation examines the relationship between program involvement and subsequent benefits for youth. Desired outcomes include positive changes in behavioral and emotional functioning, family relationships, academic achievement, or community functioning. Clearly, this represents a high stakes issue for programs and one sometimes characterized as an area where “angels fear to tread.”<sup>3</sup> Numerous questions need to be resolved: Who will conduct the evaluation? How will overburdened clinicians be persuaded to participate? What is the best way to involve key stakeholders such as school staff, youth, and families? Research design issues, finding appropriate comparison groups, and controlling for attrition are ongoing challenges in outcome evaluations. This paper offers suggestions for overcoming some of these challenges and improving program accountability.

## BACKGROUND

In 1996 the Center for School Mental Health Assistance (CSMHA) convened a panel of national experts to establish the rationale for outcome evaluations of school mental health programs and to explore issues related to this process.<sup>4</sup> Among many reasons identified for conducting outcome evaluations, two important ones emerged:

- data can inform staff and program leaders of strengths and areas needing improvement; and
- findings can help secure funding through grants, contracts, and fee-for-service activities, such as through relationships with health care plans.

The panel noted that outcome evaluations are challenging because they require involvement by diverse stakeholders including students, families, school staff, clinicians, and community leaders in all phases of the evaluation process.

---

*Laura A. Nabors, PhD; Mark D. Weist, PhD; and Matthew W. Reynolds, Center for School Mental Health Assistance, University of Maryland School of Medicine, 680 W. Lexington St., 10th floor, Baltimore, MD 21201. Supported by project # RO3 HSO9847-01 from the Agency for Healthcare Research and Quality. Also supported by project #MCJ24SH01-01-0 from the Office of Adolescent Health, Maternal and Child Health Bureau (Title V, Social Security Act), Health Resources and Services Administration, US Dept. of Health and Human Services.*

This in turn, requires choosing appropriate methods to measure variables of interest to important stakeholders. This is critical because outcome indicators of interest to mental health professionals, such as decreased depression and improved self-esteem, might not be as important to school administrators or teachers who might place more value on indicators related to behavior and academic performance.

Other factors for consideration include determining an evaluation design appropriate for the size of the program or that can be used across school types and levels. Small programs might use a case study design and focus on outcomes at the individual student level. For moderate size programs and for those beyond the initial development stage, improvements in individual academic performance and school attendance or fewer disciplinary referrals can serve as measures of improved functioning for individuals receiving services. Ratings of satisfaction with clinical services from students, parents, school administrators, and teachers<sup>5</sup> are another appropriate measure for midsize programs. Other measures gaining popularity include the effect of services on the development of assets or protective factors for youth.<sup>6,9</sup> Larger programs might design evaluation studies with treatment and comparison groups.

How outcome evaluations will be used also influences the design. By the time a program has matured to the point where a full-scale outcome evaluation is realistic and feasible, it is probably at the stage where accountability data are critical for continued funding. Program administrators want to be able to provide clear quantitative evidence that program services relate to positive changes in academic performance and behavior within the school or district. These various needs are sometimes at odds, requiring program evaluators to work with program administrators, clinicians, students, school staff, and families to educate them about the various types of evaluation, the professional standards that must be met, and the iterative nature of the process.<sup>4</sup>

## Threats to Validity

While program evaluations generally have different purposes than research, for evaluations to be useful, consumers and funders must have confidence in the integrity of the process and the conclusions. Several principles that guide research are equally important in designing quality evaluations; validity foremost among them. Internal validity relates to the ability of investigators to conclude with certainty, based on the findings, that improvements in

students' functioning were the direct result of the treatment intervention and not some other factor. External validity refers to the conclusion that improvements resulting from a particular treatment intervention can be similarly expected with different youth, in different schools, at different times.<sup>10</sup> To achieve both internal and external validity is the "gold standard" in research and requires an "experimental" design, that is, random assignment of individuals to either a treatment or comparison group. Applied research or evaluation in schools seldom achieves this level. Nonetheless, if school mental health programs are to survive and prosper, program evaluators must be aware of and try to control for several common problems that threaten validity. The following section identifies several of these factors and suggests strategies to minimize their effect.

**Selection Bias.** Selection bias occurs when substantive differences exist between the characteristics of students participating in an evaluation and those who do not, or between treatment and comparison groups.<sup>10</sup> Evaluators must take steps to ensure that all eligible youth have an equal opportunity to be selected for inclusion in the evaluation. However, because parental or student consent to participate is generally required, this can result in selection bias. Often the youth whose parents consent for them to participate in an evaluation differ in important ways from youth whose parents do not return consent forms. In addition, students with relatively high levels of functioning from intact families may be more likely to turn in consent forms than students with more serious problems. One way to minimize these problems is to group students according to risk status at initial assessment (eg, minimal, moderate, and severe risk).<sup>11</sup> The consent rate for each risk status group can then be monitored. By using specialized recruitment strategies for each group, a balanced rate of consent and participation across groups can be achieved. Incentives such as small monetary awards or gift certificates also can increase participation rates.

Once students receiving services are selected for inclusion, a comparison group must be selected. Matching those randomly selected students receiving services with comparison students the same age who are not receiving services allows reasonable conclusions about the treatment group relative to age level peers. Another option is available if the mental health program operates out of a school-based health center; comparison students can be recruited from among students who receive other center services.<sup>12</sup> Similarly, comparisons can be made between students of similar backgrounds and presenting problems who receive services in an expanded school mental health (ESMH) program compared to those receiving traditional school counseling services in a demographically similar school. This will provide information about whether students who receive ESMH services demonstrate greater improvement than students receiving more traditional services.<sup>11</sup> In a recent study, Bruns et al<sup>13</sup> found that staff in elementary schools with ESMH programs felt their schools had a better climate for addressing students' emotional and behavioral issues than staff at schools that did not have such programs.

**Practice Effects.** Practice effects occur when the instruments or method used to obtain information from students influences responses during subsequent assessments. For example, students might report similar answers on self-report measures at subsequent follow-ups because they

recall the responses they provided when completing the measures the first time. A related problem is the Hawthorne Effect, which occurs when participants, whether clinician or student, realize they are participating in an evaluation. This is a real problem in studies that involve repeated clinician ratings of youth. When clinicians know they are part of an evaluation, a strong bias exists toward indicating that youth are improving, regardless of the true status. Obtaining ratings on the same instrument from different informants including parents, teachers, and students themselves helps address this problem.

**Regression to the Mean.** If the same instrument is used repeatedly to collect data, investigators must be sensitive to the tendency of extreme scores to move toward the mean score over time. Regression to the mean could account for self-rating scores of extreme depression at pretest appearing to improve over time. In this case, the treatment intervention cannot be credited with the improvement. Design and analysis methods for addressing this bias include a randomly assigned wait-list control group and controlling for baseline scores on measures when conducting analyses to examine changes in scores over time, providing a mechanism for "teasing out" artificial improvement for extremely low scores.

Another approach to control for regression to the mean is to augment statistical analysis on groups of students with separate analyses of individual students with extreme scores at any measurement point. This strategy has benefits beyond controlling for bias. This approach might reveal a unique pattern within a subgroup, suggesting more intensive qualitative analyses or clinical intervention. For example, when a subgroup is identified who continue to present high levels of depression while participating in a particular intervention program, evaluators can assist program staff by suggesting these youth receive alternative care. Alternatively, it might suggest some other variable is operating such as association with acting out peers, pointing to the need for further study and tailoring the intervention for the subgroup.

**Timing of Measurements.** If data collection intervals are too close, practice effects can effect the validity of findings. Conversely, if measurement intervals are too far apart, participants — especially high-risk students — can be lost to attrition. Because behavioral change and adaptation among children is not linear or continuous over time, observations must be spaced to detect possible changes toward the desired outcome without producing practice effects. For example, if the desired outcome is increased involvement in activities considered assets or protective factors, such as reading for pleasure or getting involved in community activities, these changes can be detected after a relatively short period of time, perhaps after one month. However, for variables such as improved school performance or behavioral functioning, realistic changes may not occur for months or more. Data collection intervals need to be determined based on length of treatment and the interval needed to detect changes in different variables. In addition, observations need to be timed to minimize any burden for students or interference with their treatment. Experience over five years at CSMHA indicates that once the anticipated length of a treatment has been estimated, data should be collected at baseline, at the midpoint of this interval, at the conclusion of treatment, and at subsequent follow-up dates.

**Attrition.** Unequal dropout rates for different subgroups in a sample threatens both internal and external validity and can severely compromise interpretation of evaluation findings. This can be the most critical threat to the validity of program evaluations in schools, because youth with severe problems are more likely to drop out of treatment and the evaluation.<sup>14</sup> If participation in the evaluation is limited to low-risk students, external validity is seriously threatened, and findings are not applicable for the high-risk students often treated in schools.<sup>15</sup>

Without a plan for sustaining program and evaluation involvement, sample sizes can be significantly diminished. Evaluators must devise alternative strategies to maintain contact with program participants such as mailing surveys and questionnaires to absent students, conducting phone interviews, holding make-up data collection sessions for absent students, and finding students who transfer to different schools.<sup>16-18</sup> Successes and failures related to these efforts must be documented and analyzed (eg, are there differences in face-to-face, mail, and telephone interview data?) Providing students with incentives such as coupons for fast food restaurants, candy, and raffle tickets for prizes is another strategy. Tracking attrition and learning about reasons for ending treatment can provide useful information to clinicians about the types of students who might terminate treatment prematurely. Documenting outcomes for students who sporadically participate in treatment yields helpful information about outcomes for infrequent users. In addition, this group can be used as a quasi-comparison group for youth who participate more intensively in the program. Alternatively, attrition can be viewed positively as an indicator that better methods are needed to retain children in treatment or in evaluation studies.

**Examiner Bias.** Many students need help reading or interpreting questionnaire items. They sometimes need prompting to review all items and complete ones they have missed. But, if service providers administer the instruments they can (knowingly or unknowingly) influence students to respond in ways that suggest positive treatment effects. Alternatively, if clinicians administer and view the results, it might alter their treatment plan, thus influencing outcomes. For this reason, it is usually a better strategy to have trained assistants who are blind to the purposes of the study and its interventions, collect evaluation data from youth. Another strategy is to hire community members or paraprofessionals who have no knowledge of clinical practice or research to gather data. If clinicians are needed to collect data, students should complete surveys and questionnaires in a separate room, place their materials in a sealed envelope, and return them to a secretary or nurse to maintain confidentiality.

**Measurement Biases.** If methods selected by the evaluator to document outcomes are not adequate or appropriate, treatment effects might not be detected. Consequently, it is important to use reliable and valid measures used with similar youth who have undergone similar interventions.<sup>3,19</sup> Instruments should have established psychometric qualities, be relatively brief and stable over relatively short time intervals, and ideally, be appropriate for youth, parents, and teachers. Although many measures of child and adolescent psychosocial functioning exist, relatively few measures provide all these qualities, requiring use of several measures.

Regardless of the number of instruments used, measures selected should be proximally related to constructs associated with the intervention. For example, if the desired outcomes and clinical interventions with a group of children exposed to high levels of violence focus on improved coping skills and anxiety reduction, then instruments should measure these factors and not others. However, this tailored approach often contrasts with standard practice wherein a battery of assessments is administered to all youth at baseline and at regular intervals regardless of the presenting problem or the intervention they experience.

**Historical Events.** Unplanned events occurring in the students' lives can affect student outcomes and the evaluation study. When the school shooting and bombing in Littleton, Colo., occurred, several clinicians participating in our evaluation reported they did not have time to complete the study protocol because they were dealing with numerous "copycat" emergencies in the aftermath of this tragedy. In addition, fewer children attended school in the weeks after these events and, not surprisingly, some reported heightened anxiety. Under these circumstances, which are beyond control, serious threats to validity exist that must be acknowledged in interpreting study findings.

**Maturation Effects.** Adolescence is a challenging time for youth. As adolescents mature they learn a variety of coping skills. These developmental changes, alone, can result in improved behavior and emotional functioning, regardless of treatment intervention. This threat to validity is always present in school-based programs. Even with random assignment to treatment and control groups, it is not possible to ignore the impact of historical events and maturation effecting outcomes. One approach to this problem is to use a cross-sequential cohort design, in which the progress of children of the same age levels is examined using the same measures over several years. This reduces the impact of cohort effects and allows for examination of developmental changes, but is obviously a more cumbersome, time-consuming, and expensive approach.<sup>9</sup>

**Context Effects.** Factors such as income level and level of violence in the environment are other variables difficult to control that can negatively affect treatment outcomes.<sup>20</sup> Other contextual factors, such as number and extent of services received must be identified and described for each school as quantity and type of services can effect outcomes. School policies also should be considered when analyzing findings. If school policies allow suspensions and expulsions as a means of discipline for noncompliant behavior or truancy, students with behavioral and emotional problems frequently are disproportionately represented. Changes in program policy, such as new billing procedures or additional paperwork, can negatively impact clinician availability and their desire to complete evaluation protocols. These confounding factors typically cannot be eliminated or controlled, but they should be recognized as covariates or moderator variables in analyses and described as part of the study results.

## CONCLUSION

Most outcome evaluations do not use experimental designs and therefore investigators must be vigilant against threats to validity. The complexities of and differences between mental health programs across service providers, schools, and clientele makes outcome evaluation a complex

process. In response, a tendency exists to avoid outcome evaluations altogether or assume that the conclusions were compromised by too many confounding factors. However, if school mental health programs are to fulfill their promise, program planners must demonstrate that the programs can achieve desired outcomes. Outcome evaluations must become routine. As researchers, program leaders, staff, and stakeholders rise to this challenge, we must recognize the realities of schools, strive to understand all the factors that influence outcomes, and generate ideas for future study.

Outcome evaluations of school mental health programs are in their infancy. Most focus on examining the short-term impact of services on individuals. This focus may need to expand to explore the short-term and long-term effects of these programs in different schools and in various communities. As evidence mounts that ESMH programs have multiple beneficial outcomes for youth, it will become possible to determine with more precision which specific treatments and treatment elements are related to positive changes for children with specific problems, such as phobias or depression. Outcome evaluations are a necessary step to assuring that school mental health programs can continue to provide opportunities for advancing the mental health of children and youth. ■

#### References

1. Weist MD, Schlitt J. Alliances and school-based health care. *J Sch Health*. 1998;68(10):401-403.
2. Weist MD, Nabors L, Myers P, Armbruster P. Evaluation of expanded school mental health programs. *Community Mental Health J*. In press.
3. Forness SR, Hoagwood K. Where angels fear to tread: issues in sampling, design, and implementation of school-based mental health services research. *Sch Psychol Q*. 1993;8:291-300.
4. *Documenting the Effectiveness of School Mental Health Programs*. Baltimore, Md: Center for School Mental Health Assistance; 1996.
5. Nabors L, Tashman N, Myers P, Weist M. Quality assurance activities in high school mental health centers. Poster presented at Eleventh Annual Conference: System of care for children's mental health: Expanding the research base. Tampa, Fla; 1998.
6. Masten AS, Coatsworth JD. The development of competence in favorable and unfavorable environments: lessons from research on successful children. *Am Psychol*. 1998;53:205-220.
7. Nabors LA, Reynolds MW. Program evaluation activities: outcomes related to treatment for adolescents receiving school-based mental health services. *Child Serv: Soc Policy, Res, & Practice*. In press.
8. Benson PL, Leffert N, Scales PC, Blyth DA. Beyond the "village" rhetoric: creating health communities for children and adolescents. *Appl Dev Sci*. 1998;2:138-159.
9. Leffert N, Benson PL, Scales PC, Sharma AR, Drake DR, Blyth DA. Developmental assets: measurement and prediction of risk behaviors among adolescents. *Appl Dev Sci*. 1998;2:209-230.
10. Cook TD, Campbell DT. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston, Mass: Houghton Mifflin; 1979.
11. Werthamer-Larrson L. Methodological issues in school-based services research. *J Clin Child Psychol*. 1994;23:121-132.
12. Weist MD, Paskewitz DA, Warner BS, Flaherty LT. Treatment outcome of school-based mental health services for urban teenagers. *Community Mental Health J*. 1996;32:149-157.
13. Bruns E, Walrath C, Acosta O, Anderson K, Glass-Siegel M, Weist M. Surveying school climate in schools with and without expanded mental health programs. 1999. unpublished manuscript.
14. Kazdin AE, Mazurick JL, Bass D. Risk for attrition in treatment of anti-social children and families. *J Clin Child Psychol*. 1993;22:2-16.
15. Dryfoos JG, Brindis C, Kaplan DW. Research and evaluation in school-based health care. *Adolesc Med*. 1996;7:207-220.
16. Ellickson PL. Getting and keeping schools and kids for evaluation studies. *J Community Psychol*. 1994;102-116.
17. Mowbray CT, Bybee D, Collins ME, Levine P. Optimizing evaluation quality and utility under resource constraints. *Eval Program Plann*. 1998;21:59-71.
18. Ribisl KM, Walton MA, Mowbray CT, Luke DA, Davidson WS, Bootsmiller BJ. Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: review and recommendations. *Eval Program Plann*. 1996;19:1-25.
19. Reynolds WM. Self-report methodology. In: Ollendick TH, Hersen M, eds. *Handbook of Child and Adolescent Assessment*. Needham Heights, Mass: Allyn & Bacon; 1993.
20. Mercier C, Fournier L, Peladeau N. Program evaluation of services for the homeless: challenges and strategies. *Eval Program Plann*. 1992;15:417-426.

**New  
Member  
Benefit  
from  
ASHA**

**Communicate,  
Collaborate**

## School Health ListServ

An electronic forum for seeking help and sharing insights related to school health.

To become a participant:  
**Send an e-mail to: [listserv@ilstu.edu](mailto:listserv@ilstu.edu)**

In the body of the message, type:  
**SUBSCRIBE schoolhealth-1 [Your name]**  
The "1" after school health is a lowercase "L", not a number 1.  
The listserv recognizes the e-mail address from which you send the message as your e-mail address.

**This is a closed listserv only available to ASHA members.**  
There is no cost other than your cost of ASHA membership.  
For information, contact ASHA at 330/678-1601 or  
<[asha@ashaweb.org](mailto:asha@ashaweb.org)>.