Effective Use of the Large Body of Research on the Effectiveness of Programs for Juvenile Offenders and the Failure of the Model Programs Approach

Mark W. Lipsey
Vanderbilt University

Michael Baglivio, Kevin Wolff, Katherine Jackowski, Gabrielle Chapman, Mark Greenwald, and Katherine Gomez (2018, this issue) report the relationship of ratings of the quality of service delivery for 56 residential programs for juvenile offenders in Florida to the subsequent recidivism of the youth who participated in those programs. Quality of service delivery is one of the four components of the Standardized Program Evaluation Protocol (SPEP™), a scheme for assessing the expected effects of therapeutic juvenile justice programs on recidivism. The other SPEP™ components include the generic program type (cognitive-behavioral therapy, family therapy, individual counseling, etc.), the amount of service provided (duration and contact hours), and the recidivism risk of the youth served. The focus on these factors, how each is defined, and the associated rating and weighting scheme are based on the program features found to be associated most strongly with recidivism reductions in a meta-analysis of 548 controlled studies of interventions with juvenile offenders (Howell and Lipsey, 2012; Lipsey, 2009).

The SPEP™ allows for a certain amount of flexibility and local tailoring of the data sources that provide input for the ratings and the data elements incorporated into those ratings. The recidivism risk rating, for instance, is based on scores from a validated risk assessment instrument, but no particular instrument is specified and this rating is adapted to whatever eligible assessment is in use by a juvenile justice system. Similarly, even though the SPEP™ specifications for the quality of service delivery rating address the treatment...
protocol, provider training, monitoring of adherence to the protocol, and corrective action when adherence drifts, SPEP™ users are encouraged to include further differentiation that reflects their system’s quality assurance perspective and practice. In that spirit, the Florida Department of Juvenile Justice (FDJJ) has added a staff turnover element to its rating scheme, differentiated the training element to ask whether actual skill in delivering the treatment is part of personnel evaluation, and enhanced the monitoring element to include direct observation of treatment sessions by FDJJ personnel as well as review of the provider’s monitoring procedures.

The results of the FDJJ analysis of the relationship of this quality of service delivery rating to recidivism are remarkable in several regards. Although that rating is only one of the four SPEP™ components designed to work in combination to predict recidivism reductions, it shows a surprisingly strong predictive relationship to recidivism all by itself. The summary ratings are related to lower recidivism in all analyses, and they are statistically significant for most. Indeed, all the relationships between the three kinds of recidivism examined and the average rating across programs are statistically significant, indicating that the average rating of the quality of service delivery is the better summary index. Moreover, the relevant unit of analysis is rather global, a residential facility, and the 56 facilities represented do not constitute a very large sample for attaining statistical significance, making such findings all the more remarkable. And, even though there is sufficient variation around the overall mean ratings to support analysis, those means are modest with few facilities presenting notably high scores. At the same time, the sample is nearly the statewide population of facilities, so these are results at scale for representative performance in a state with an exceptionally progressive juvenile justice system.

Especially important is the way the Baglivio et al. (2018) study complements the two SPEP™ validation studies that have been conducted to date, both in Arizona. In those studies, the relationship between overall SPEP™ scores for a sample of juvenile justice programs and the recidivism of the participating youth was examined, controlling for their baseline risk and related background factors. The first of these studies was done with a mixed sample of 66 community and residential programs (Lipsey, 2008). The second used later data with expanded coverage of 90 programs (Redpath and Brandner, 2010). However, neither of these studies included the quality of service delivery component in the summary SPEP™ scores because it was still under development at the time. Even without that component, both Arizona studies found relationships in the expected direction between the partially incomplete summary SPEP™ scores and 6- and 12-month rearrest recidivism, with recidivism differentials between higher and lower scoring programs in the range of 7 to 12 percentage points. The Baglivio et al. (2018) Florida results showing the independent predictive strength of the quality of service delivery component that was omitted in the Arizona studies thus fills that gap and provides further evidence of the validity of the SPEP™ for rating the expected effects of programs for juvenile offenders on their subsequent recidivism.
Different Perspectives on What Constitutes an Evidence-Based Program

The conventional conception is that an evidence-based program is a brand-name manualized program model supported by research demonstrating favorable effects on an outcome of interest. These are typically identified by one or another more or less authoritative organization that has reviewed the pertinent research and reported its rating of the program in its registry of evidence-based programs. Familiar registries of this sort in juvenile justice include the Blueprints for Violence Prevention (now expanded as Blueprints for Healthy Youth Development), the Office of Juvenile Justice and Delinquency Prevention (OJJDP) Model Programs Guide, and CrimeSolutions.gov. Examples of familiar programs targeting recidivism outcomes for juvenile offenders that appear in these registries include Multi-systemic Therapy (MST), Functional Family Therapy (FFT), and Aggression Replacement Training (ART). Although it does not always receive much emphasis in the respective registry presentations, it is understood that model evidence-based programs must be implemented with fidelity to the program specifications before the effects demonstrated in the supporting research can be expected to be produced in practice.

The SPEP™ is based on a broader conception of evidence-based programs. Rather than orienting to specific brand-name programs, its foundation is research on generic program types, such as cognitive-behavioral therapy, group counseling, family therapy, mentoring, and the like. Many brand-name programs fall into one or another of these generic categories, but so do many homegrown and one-off programs. And there are far more evaluation studies of the generic versions of these programs than for any of the brand-name versions and, thus, much more evidence about their effectiveness. That evidence shows favorable average effects on recidivism for a wide range of programs, although with systematic differences for different program types. Moreover, for the generic program categories that include the familiar brand-name, evidence-based programs, the recidivism effects of those programs are not systematically better than those of the larger number of no-name programs in those categories (Lipsey, 2012; Lipsey, Howell, Kelly, Chapman, and Carver, 2010).

As with brand-name programs, the way in which programs of a given generic type are implemented has considerable influence on the direction and magnitude of their effects. The main thrust of the meta-analysis of evaluation studies on which the SPEP™ is based has been to identify the program characteristics that are most strongly and robustly associated with favorable effects on recidivism (Lipsey, 2009). Those have been incorporated into the SPEP™ assessment to provide a systematic basis for differentiating the program types, service delivery characteristics, and youth characteristics expected to yield larger or smaller effects on subsequent offending. This assessment allows juvenile justice systems to determine which of their programs, brand-name or otherwise, are broadly supported by evidence of effectiveness and whether they are implemented in a manner consistent with what the evidence indicates is associated with positive outcomes of practical significance.

This generic approach has some intrinsic advantages. It builds on the program array already in place in a juvenile justice system rather than promoting conversion of existing
local programs to brand-name versions. In addition, it extends an evidence-based program framework to many common types of programs for which there are few or no brand-name evidence-based versions, for example, individual and group counseling, vocational training, social skills training, and victim-offender mediation. Via the SPEP™, this approach is then supported by a broadly applicable, data-driven assessment that can be used to routinely and systematically track the expected effectiveness of a large proportion of the programs a system uses. Perhaps most importantly, the SPEP™ assessment provides specific feedback about areas of program weakness that can be addressed within system-wide continuous improvement and quality assurance procedures. The Baglivio et al. (2018) study adds to the evidence that this approach provides valid guidance to juvenile justice systems about how to effectively use the results of the large body of research on interventions with juvenile offenders to improve the outcomes for the youth they serve.

Failure of the Model Program Approach for Improving Juvenile Justice Practice

Registries of evidence-based model programs appropriate for use in juvenile justice systems have been available for many years. Blueprints was launched in 1996, and the OJJDP Model Programs Guide began in 2000. During that period, the potential of evidence-based programs to improve juvenile justice outcomes has been increasingly recognized and many systems have made concerted efforts to implement model programs identified by these sources. The history of these initiatives and implementation efforts has identified serious problems with the model program approach, and it is now clear that it has failed to have much impact on the effectiveness of juvenile justice practice at a scale that would be consequential for youth outcomes. The major problems are twofold: (a) limited evidence that the model programs central to this approach are capable of having broad impact, and (b) obstacles to uptake of the model programs identified by what evidence is available.

The problems with the evidence base are manifold. First, few manualized programs suitable for dissemination have been evaluated with sufficiently rigorous research designs to meet the model program criteria of the registries that review the respective research. Blueprints, which rightly prides itself as having the highest standards for the quality of the supporting research, has identified only three model programs over its 20-year history that target system-involved juvenile offenders and show effects on recidivism, and two of those are variations on family therapy (MST and FFT), whereas one is specialized foster care—not a very broad repertoire for juvenile justice practice. Other registries identify more programs, but only by applying less stringent standards for the quality of the supporting research. Moreover, the research they rely on to deem programs for juvenile offenders “evidence based” typically consists of few studies, most often only one or two, and those studies are usually conducted by the program developers as small-scale demonstration projects. Not only is the generalization of the results of these studies to routine juvenile justice practice questionable, but developer involvement in the implementation of the program is
a circumstance that tends to produce overestimates of the effects that will be found with independent replication (Petrosino and Soydan, 2005).

This thin and rather narrow research support for programs identified as evidence-based in the influential registries raises serious questions about the likelihood that they will be effective when implemented at scale on a routine basis by personnel other than the original program developers. The program evaluation literature is replete with examples of programs that performed well in demonstration studies but for which the results were disappointing when they were implemented and evaluated under more realistic practical conditions. One recent example comes from the Education Innovation and Research program, a tiered-evidence initiative at the U.S. Department of Education that supported evaluation of programs with promising initial evidence of effectiveness when implemented under conditions more typical of routine practice. The evaluation results released early in 2017 revealed that only 13 of the 44 programs generated overall positive impact findings with 7 others showing mixed results; that is, more than half did not yield significant positive effects when implemented under realistic conditions (Lester, 2017). A similar pattern appeared in the tiered-evidence results for teen pregnancy prevention programs (Farb and Margolis, 2016). Only 12 of the 41 studies found overall positive effects on the key behavioral outcomes. For the Tier 1 programs, those with the strongest initial evidence of effectiveness, only 4 of 19 demonstrated overall positive effects.

No similar systematic assessment of the effects in routine practice of a selection of programs for juvenile offenders from evidence-based program registries is available. Nevertheless, there is little reason to suppose those programs are any more likely to retain their effectiveness in routine practice than programs in other areas (Rhoades, Bumbarger, and Moore, 2012; Walker, Bumbarger, and Phillippi, 2015). FFT, one of the most widely used and well-documented programs for juvenile offenders, has failed to perform as expected in some studies of its effects when implemented in practice (Barnoski, 2002; Humayun et al., 2017; Sexton and Turner, 2010). And, a recent meta-analysis of MST (van der Stouwe, Asscher, Stams, Deković, and van der Laan, 2014), another widely used and well-documented program, found a modest mean effect size for the 11 studies that reported delinquency outcomes with juvenile offenders (0.16) and, overall, a dramatically smaller mean effect size for the 12 studies conducted by independent researchers than for the 8 studies conducted by researchers associated with the program developers (0.08 vs. 0.42). Program developers typically cite weak implementation as the source of such poor performance, and there is evidence to support that claim, but it begs the question of how realistic it is to expect the same level of fidelity to the program protocol in widespread practice as the developers are able to achieve in the small-scale demonstration projects they organize themselves.

There are many potential differences between the conditions that allow high-fidelity implementation of model programs in research demonstration projects and those present in juvenile justice systems that attempt to adopt and maintain those programs for everyday use.
For example, the service infrastructure for delivering the program is likely to be weaker in routine practice than that organized by program developer when conducting their evaluation research. Such resources as trained service providers and funds for personnel and capital expenditures may not be sufficient to fully meet the requirements of a model program. Also, it may not be possible to restrict the program to the same population represented in the research. In real-world settings, programs may of necessity be required to serve a more heterogeneous population than was used in the research. As a result of circumstances such as these, the desirable program effects found in the supporting research studies are often attenuated when those programs are adopted for general application (Welsh, Sullivan, and Olds, 2010).

Even if, despite good reasons to be skeptical, we suppose that the research-practice effectiveness gap for model programs could be bridged, there are many obstacles to widespread uptake of those programs by juvenile justice systems. Despite years of emphasis on evidence-based model programs, they still constitute a very small proportion of the repertoire of programs in use by juvenile justice systems. There is no national census of juvenile justice programs, but our team has made an inventory of the programs in the systems with which we have worked to implement the SPEP™ assessment. This sample includes juvenile justice systems in Arizona, Connecticut, Delaware, Florida, Georgia, Iowa, North Carolina, Pennsylvania, Tennessee, and Wisconsin. These systems are likely more progressive than average, and many have made notable efforts to implement brand-name model programs. Their uptake of evidence-based model programs is thus likely to be an overestimate of what has occurred nationally.

Only therapeutically oriented programs were counted in this inventory—those directed at bringing about positive psychological or behavioral change expected to endure well past the end of treatment—and each site with a separate treatment implementation was counted as a distinct program even if administered by the same multisite provider. Across these 10 systems, a total of 1,087 programs were identified. Of those, only 79 (7.3%) were programs listed in the Blueprints, OJJDP Model Programs Guide, or CrimeSolutions.gov evidence-based program registries. Although no client counts are available for these programs, we can assume that a correspondingly small proportion of the juvenile offenders served by these systems participated in one of these evidence-based model programs. That assumption is consistent with the Henggeler and Schoenwald (2011) estimate that only about 5% of eligible high-risk offenders are treated with evidence-based model programs. By comparison, 786 (80%) of these programs fell into one of the categories of generic program types with meta-analytic evidence of effectiveness to which the SPEP™ applies.

There are many readily identifiable reasons why model programs have not been widely adopted in juvenile justice systems. Local providers perceive confidence in their own programs and are reluctant to replace them with model programs developed elsewhere that have not been proven to be more effective than what they are already doing. Those providers are also part of an established social and political infrastructure that juvenile
justice administrators may not wish to challenge and, indeed, may find it beneficial to support. Furthermore, there are real costs associated with adoption of model programs—not only licensing fees but also training and maintenance expenses that must be assumed. Moreover, the repertoire of model programs for juvenile offenders with evidence of effects on recidivism is limited and not obviously sufficient to address all the diverse needs of youth who become involved with the juvenile justice system. Perhaps the biggest obstacle, however, is the “by-the-book” fidelity required by model programs that providers often view as a straitjacket on their professional judgment about what is best for their clients and local circumstances.

**Alternative Focus for Evidence-Based Practice**

Manualized intervention programs are generally understood to be recipes for services that can be expected to produce the same effects on similar consumers wherever and whenever applied so long as the recipe is followed with high fidelity. This perspective justifies promotion of the use of such a program recipe once it has been shown effective in a demonstration study or two, with the caveat that there be no deviation from the original. This is a flawed conception. Local ingredients vary widely as do consumers, and local cooks have their own interpretation of the recipes and how they should be adapted to their purposes. In practice, there will inevitably be significant variation across implementations of any program taken to scale with correspondingly significant variation from the model version no matter how fully it is specified in the program manual. In practical application, settings and context vary, the skill and personal style of providers vary, the nature and background of the individuals served vary, and the resources and infrastructure for supporting program implementation vary. It is unrealistic to believe that all implementations of a model program can be constrained to mirror the particulars of the original demonstration research.

The program recipe conception might still be viable for at least some model programs if there were sufficient evidence to establish that the effects from good faith attempts to implement that model were robust to the range of variation expected to occur at scale in routine practice. But, as described earlier, there is no such evidence for model juvenile justice programs, and the results of such evidence in other areas where programs that showed promise in initial efficacy studies were taken to scale is not encouraging. This circumstance is familiar—the relative effects of fidelity versus adaptation when implementing intervention programs has been discussed for decades without resolution (e.g., Bopp, Saunders, and Lattimore, 2013; Castro and Yasui, 2017; Hansen, 2013; Kemp, 2016).

An alternative perspective is to shift the focus from the recipe to the fidelity specifications. That is, it may not be evidence-based recipes that are most critical for effective practice but evidence-based standards for implementation that identify the program features most crucial for attaining the desired outcomes. If such standards are measureable and can be assessed at scale in routine practice, they can then provide ongoing practical guidance for keeping implementation within a range where positive effects can be reasonably expected.
Many model programs have accompanying fidelity measures, but these are typically created on the basis of theory by the developer and mainly assess the extent of compliance with what the developer deems to be the most important instructions in the program manual. By contrast, evidence-based implementation standards would identify a profile of program characteristics shown empirically to be related to positive program outcomes across a large number and diversity of implementations in routine practice.

Empirically establishing such evidence-based standards would define the essential characteristics of an effective program and provide validated guidance for implementation that would retain that effectiveness. At the same time, they would allow providers the flexibility to make local adaptations to the program protocol in areas that were not constrained by those standards, thus addressing the perennial fidelity versus adaptation issue. Moreover, for programs already successfully implemented, such evidence-based standards would provide a basis for quality assessment tools that could be used routinely to measure expected program effectiveness and provide feedback for program improvement. From this perspective, the primary challenge for researchers is not to develop and test recipes for programs that will be broadly effective, although that is important, but to tackle the measurement problem of developing empirically valid predictors of program effectiveness across a range of real-world implementations.

The SPEP™ assessment is such a measurement scheme, albeit one based on rather simple and crude program features. It focuses on a small set of generic program characteristics found in a meta-analysis of hundreds of evaluation studies to be empirically predictive of favorable effects on recidivism. If future evaluation research produces more differentiated data on the characteristics of the programs evaluated, further meta-analysis should yield an even more refined set of empirically based factors with which to assess the expected effects of different juvenile justice programs on recidivism. The Baglivio et al. (2018) study of the predictive validity of the FDJJ rating scheme for quality of service delivery adds a key piece of evidence about the validity and utility of this approach when used at scale in routine juvenile justice practice across a wide range of programs. And, in so doing, it makes an important incremental contribution to a better understanding of the elements of an evidence-based juvenile justice system (Howell, Lipsey, and Wilson, 2014).

References


Mark W. Lipsey is a research professor in the Peabody Research Institute and the Department of Human and Organizational Development at Vanderbilt University. His research specialties are evaluation research and meta-analysis investigating the effects of social interventions with children, youth, and families. Recent research topics include risk and intervention for juvenile delinquency and substance use, early childhood education programs, methodological quality in program evaluation, and ways to help practitioners and policy makers make better use of research to improve the outcomes of programs for children and youth. Prof. Lipsey’s published works include textbooks on program evaluation, meta-analysis, and statistical power as well as articles on applied methods and the effectiveness of school, community, and residential programs for youth.