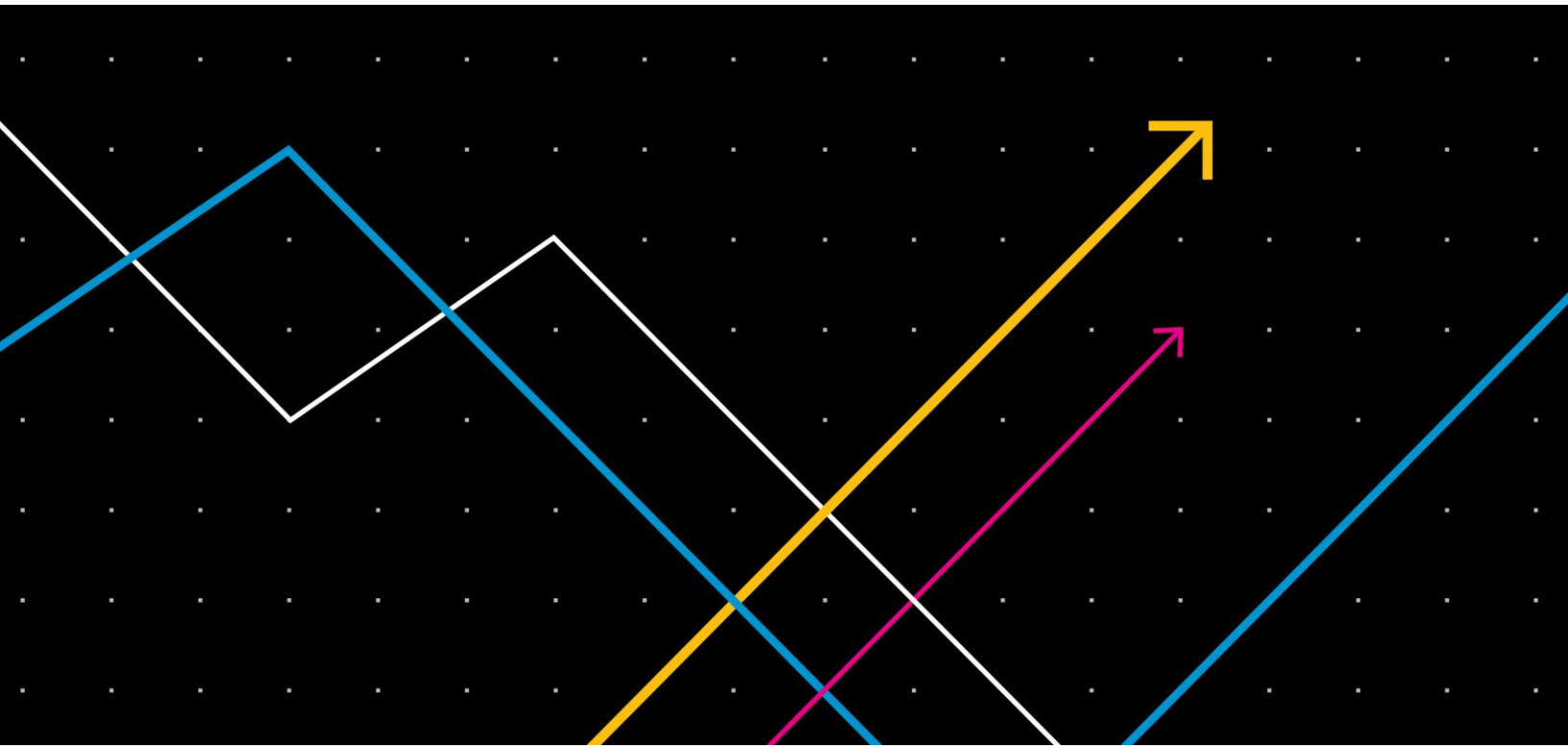


RESEARCH REPORT

Local Validation of SPEP™ Ratings of Juvenile Justice Program Effectiveness

A Case Study

Akiva Liberman *Jeanette Hussemann*

April 2017

URBAN INSTITUTE · ELEVATE · THE · DEBATE

ABOUT THE URBAN INSTITUTE

The nonprofit Urban Institute is dedicated to elevating the debate on social and economic policy. For nearly five decades, Urban scholars have conducted research and offered evidence-based solutions that improve lives and strengthen communities across a rapidly urbanizing world. Their objective research helps expand opportunities for all, reduce hardship among the most vulnerable, and strengthen the effectiveness of the public sector.

Copyright © April 2017. Urban Institute. Permission is granted for reproduction of this file, with attribution to the Urban Institute. Cover image by Tim Meko.

Contents

Acknowledgments iv

1. Introduction 1

Background on JJRRI 2

Risk Assessment 2

JJRRI Sites 3

2. The SPEP™ and Its Requirements 5

Program-Level Elements 5

Client-Level Elements 6

Data Requirements 7

SPEP™ Ratings in JJRRI Sites 8

Interim SPEP™ Ratings 8

3. A Case Study of SPEP™ Ratings in Iowa 10

Types of Services 11

Ratings on SPEP™ Elements 12

SPEP™ Ratings Excluding Risk 13

Correlations among SPEP™ Elements 14

Risk Data in Iowa 14

4. Attempting Local Validation of the SPEP™ 16

Validation Considerations 16

Risk Data in SPEP™ Validation 16

Multiple Concurrent Services 17

Comparison of Arizona and Iowa Samples 18

5. Conclusion 20

Appendix A. Sample SPEP™ Scoring Sheet 21

Appendix B. Iowa Ratings on SPEP™ Elements 22

Appendix C. Comparing Iowa and Arizona Samples 24

Notes 26

References 27

About the Authors 28

Statement of Independence 29

Acknowledgments

This report was funded by Grant No. 2012-PF-F2-K001 awarded by the Office of Juvenile Justice and Delinquency Prevention, Office of Justice Programs, US Department of Justice. Points of view expressed in this document are those of the authors and do not necessarily represent the official position or policies of OJJDP or the US Department of Justice. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission.

The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at www.urban.org/support.

1. Introduction

At the end of 2012, the Office of Juvenile Justice and Delinquency Prevention launched the Juvenile Justice Reform and Reinvestment Initiative (JJRRI) at three demonstration sites in Delaware, Iowa, and Milwaukee County, Wisconsin. The goal of JJRRI was to bring evidence and best practices to bear on juvenile justice operations using three types of tools:

- Dispositional matrices provide evidence-based recommendations concerning dispositional options.
- The Standardized Program Evaluation Protocol (SPEP™) rating system brings evidence concerning program effectiveness to bear on juvenile justice services and guides improvements to those services.
- Validated risk assessments are necessary for both dispositional matrices and SPEP™ ratings. Together, these tools are intended to increase effectiveness and efficiency in the use of juvenile justice resources.

Concurrent with JJRRI implementation, the Urban Institute evaluated the initiative to understand whether it improved the quality and effectiveness of juvenile justice programming.

One goal of the evaluation was to understand the implementation of the SPEP™ rating system, and this was the subject of our first evaluation report (Liberman and Hussemann 2016). The rating system is quite simple in conception, and is described briefly in chapter 2 of this report. Despite its simplicity, however, conducting a first round of SPEP™ ratings is usually an intensive effort that can take two to three years to complete.

Our previous report detailed the implementation requirements of the SPEP™: strong data systems and reliable and timely risk assessment. The SPEP™ often reveals deficiencies in these systems and can then help drive improvements. These improvements, in turn, require support from a range of juvenile justice stakeholders and considerable technical assistance. After reviewing SPEP™ implementation requirements, that report described the challenges the JJRRI sites encountered and how they were addressed. We concluded that a first round of ratings tends to uncover deficiencies that then motivate reforms with considerable potential to improve the effectiveness of the juvenile justice system.

The evaluation also aimed to locally validate the relationship between SPEP™ ratings and reduced recidivism, which is the subject of this report. A local validation would replicate previous work done in

Arizona by Lipsey (2008) and Redpath and Brandner (2010). Of the three JJRRI sites, only Iowa seemed a suitable site for local validation. Iowa's Division of Criminal and Juvenile Justice Planning provided the Urban Institute with data for this purpose that are analyzed in the current report. As a prelude to attempting to locally validate the SPEP™, chapter 3 reports on Iowa's first round of SPEP™ ratings, collected between 2012 and 2015. However, although results in Iowa were promising in some respects, they ultimately did not provide an opportunity for local validation.

The rest of this chapter briefly introduces JJRRI. Chapter 2 then briefly reviews the SPEP™ rating system and its data requirements. Chapter 3 examines the data collected in Iowa. Chapter 4 explores the possibility of using that data to locally validate the relationship between SPEP™ ratings and reduced recidivism. Chapter 5 concludes with a summary and lessons learned.

Background on JJRRI

The JJRRI demonstration program aimed to use evidence to increase effectiveness and efficiency in the use of juvenile justice resources to improve services for youth in the justice system and reduce recidivism. Implementing the SPEP™ rating system was a centerpiece of JJRRI.

In addition to the SPEP™, JJRRI also worked to develop dispositional matrices to guide juvenile justice decisionmaking. Dispositional matrices typically combine information about risk of recidivism (as determined by a validated risk assessment tool) with current case information to generate recommended dispositional options. For example, someone with violent felony charges and a medium risk of recidivism might be recommended for out-of-home placement followed by intensive supervision, but someone with a misdemeanor charge and low risk level might be recommended for court diversion.

Risk Assessment

JJRRI focused considerable effort on ensuring that risk assessment data, used in both the SPEP™ system and in dispositional matrices, were valid at the local level. Note that "risk" here refers strictly to the actuarial risk of recidivism.

The SPEP™ system and the dispositional matrices both use categorical levels of risk (e.g., low, medium, or high risk), which are generally produced from continuous risk scores. The cut-points for different risk levels can vary locally. Thus, one important aspect of local validation is setting appropriate

local cut-points for levels of risk that take into account the distribution of recidivism risk among the local juvenile justice population. This work can be time consuming and data intensive, and JJRRI provided considerable technical assistance (TA) to help the demonstration sites work through risk assessment data issues.

JJRRI Sites

IOWA

JJRRI was implemented in the first, third, and sixth Iowa judicial districts with strong support from the chief juvenile officers in those districts. The chief juvenile officer in each district oversees juvenile program and service contracts and community supervision of youth formally charged with a delinquent act. Implementation was coordinated by the Iowa Department of Human Rights' Division of Criminal and Juvenile Justice Planning, which serves as the Iowa Statistical Analysis Center.

The three districts implemented SPEP™ ratings for both local, community-based services and statewide residential programs. Risk information was obtained from the Iowa Court Information System, a data system that stores juvenile and criminal justice case information, including court services, case processing, financial reporting, and appellate record review. The system is regularly updated by juvenile court staff in all jurisdictions.

DELAWARE

JJRRI was coordinated by the Division of Youth Rehabilitative Services, a branch of the Department of Services for Children, Youth and their Families that provides supervision to preadjudicated and adjudicated youth.

Delaware focused on community-based services for high-risk youth and decided not to rate services only provided to low-risk youth. In its JJRRI grant application, Delaware also proposed focusing its SPEP™ ratings on newly contracted programs that the state had not yet evaluated or audited.

By policy, at the time of the first round of SPEP™ ratings, risk assessments were administered after adjudication and only to youth who received community supervision. Because risk assessment data are central to the SPEP™ rating system, services provided to youth in residential settings could not be rated. Department of Youth and Rehabilitative Services providers used the Family and Child Tracking System to track juvenile cases and services, which allowed agency staff members to access juvenile justice information and youth assessment and service information in other divisions, including the

Division of Family Services, Division of Prevention and Behavioral Health Services, and the Division of Management Support Services. The Family and Child Tracking System did not include risk data, which were stored in a separate database.

MILWAUKEE COUNTY, WISCONSIN

JJRRI was coordinated by the Milwaukee County Delinquency and Court Services Division (DCSD), which administers nonjudicial operational services, intake and probation services, and provision of purchased services. The agency also operates a 120-bed detention facility. DCSD serves youth from referral though the end of their court-imposed supervision.

Programs rated with the SPEP™ included those funded directly through DCSD or through Wraparound Milwaukee, a managed care program operated by the Milwaukee County Behavioral Health Division that serves clients with a DSM-IV diagnosis. Only some Wraparound Milwaukee clients are juvenile justice youth, and most Wraparound Milwaukee services are funded through DCSD, child welfare agencies, mental health agencies, or Medicaid capitation. Residential and community-based programs from each agency were included in the SPEP™.

Juvenile justice information was retrieved from multiple data systems. Case processing information was retrieved from the Juvenile Information Management System, which tracks youth from court referral through the end of their court-imposed supervision. Case management and service information for youth receiving services through Wraparound Milwaukee was retrieved from that program's database, Synthesis, and risk assessment data were retrieved from a separate database.

2. The SPEP™ and Its Requirements

The SPEP™ gives local jurisdictions an evidence-based tool to rate juvenile justice services based on their potential to reduce recidivism. The SPEP™ is based on a quantitative synthesis of evidence from hundreds of controlled evaluations of interventions intended to reduce recidivism among justice-involved youth, which was conducted by Mark Lipsey of Vanderbilt University's Peabody Research Institute. Lipsey's published meta-analysis of the characteristics of effective programs involved 548 independent study samples (Lipsey 2009; Lipsey et al. 2010). He identified four basic characteristics of effective juvenile justice services—the type of service, the quality of service delivery, the service dosage, and the risk level of youth receiving the service—and used them to develop the SPEP™ tool. The tool rates services on each of these four elements on a total scale from 0 to 100. A sample SPEP™ rating tool is included in appendix A.

The SPEP™ is generally used as part of a continuous quality improvement cycle of ratings and improvement: a round of SPEP™ ratings guide program improvement, which should lead to improved ratings in the next round. The SPEP™ is designed to promote improvement in individual programs as well as the larger juvenile justice system. At the program level, ratings provide feedback on areas of possible improvement. If most programs being used by a jurisdiction are rated on the SPEP™, these ratings can be used to assess the current array of programs. An assessment may show that the array of programs could be improved through additional program options, that the wrong youth are being referred to programs, or that programs are generally not being contracted for effective service dosages. Such changes at the system level are sometimes referred to as improving "system alignment."

Program-Level Elements

Two SPEP™ elements—type of service and quality of service delivery—are assessed at the program level. Evidence indicates that some types of services are more effective than others in reducing juvenile recidivism. Programs with a therapeutic orientation (e.g., counseling and skill building programs) are much more effective than programs with a control orientation (e.g., boot camps and "scared straight" programs), and some therapeutic programs are more effective than others. Rating this element involves categorizing services according to the research literature. Note that when one program delivers multiple services, such as individual counseling and social-skills training, to the same youth, each service must be rated separately because each has different potential to reduce recidivism.

The research on which the SPEP™ is based has found that service effectiveness does not change between community and residential settings (Lipsey 2009). Therefore, services in both settings can be rated on the same criteria. Because youth in a residential setting often receive "bundles" of services, these services must be "unbundled" before they can be rated.

The SPEP™ also rates a program's ability to deliver its service consistently and reliably, referred to as "quality of service delivery." This element considers several factors, such as whether programs have manuals, initial and continuous staff training, program monitoring, and protocols for remedial action to correct program gaps or drift. This is the least standardized element of the rating system, and SPEP™ developers have worked with sites to customize this element to accommodate local circumstances (Lipsey et al. 2010).

Client-Level Elements

The remaining SPEP™ elements—service dosage and the risk level of youth served—require data about individual clients that are then aggregated to the service. Service dosage data includes detailed information on how much of a service each youth received. The research underlying the SPEP™ is based on evaluations of recidivism reduction for services delivered with some dosage (e.g., 20 hours of service). When the same service is delivered at a lower dosage, previous evidence of effectiveness does not generalize. Dosage targets for the SPEP™ are based on the number of contact hours for each youth (e.g., 3 hours per week) and the duration of service delivery (e.g., 12 weeks). The SPEP™ rates a program based on the percentage of its juvenile justice clients who meet dosage targets.

When youth receive multiple services from one program, dosage data must be established separately for each service. For example, if Alex received 15 hours of social-skills training with a dosage target of 24 contact hours and 15 hours of family counseling with a dosage target of 30 contact hours, then neither service met its target despite Alex receiving 30 total hours of services. Records of client contacts must be specific to each service being rated.

The final element of the SPEP™ looks at which youth receive a service. Research consistently finds that programs are more effective in reducing recidivism when delivered to higher-risk youth (Lipsey 2009). This is often referred to as the "risk principle" (Andrews, Bonta, and Hoge 1990). Basically, recidivism cannot be reduced much further in youth already at low risk of recidivism, and the research does not support mandating services—especially intensive services—for low-risk youth.[1] Therefore, to

be rated on the SPEP™, a system must assess a youth's risk of reoffending with a validated risk tool *before* services are delivered.

Note that the SPEP™ does not rate how well youth are matched to services based on their specific needs. Although it is generally accepted that not all youth need the same services, and treatment matching is believed to benefit program effectiveness, Lipsey's meta-analyses of program evaluations have not shown that better matching to needs is associated with more effective recidivism reduction (Howell and Lipsey 2012). As a result, the SPEP™ does not rate needs matching, nor does it have separate ratings for services for populations with specialized needs, such as sex offenders or youth with mental illness.[2]

Data Requirements

Assessing each element rated on the SPEP™ requires data, and collecting that data and completing a first round of ratings requires considerable time and effort. For program-level elements, which require only one datum for each service, rating may take time, effort, and TA, but does not impose a large burden on data or data systems. In contrast, the data elements that must be assembled at the client level have substantial data requirements that may require modification of existing systems.

SPEP™ ratings are based on the percentage of clients who exceed thresholds of risk and required dosage. These two elements require well-developed management information systems (or other data systems to store and update this data systematically) as well as systems to generate those data. For risk data, a risk assessment system, including a validated risk tool, is needed, as are policies and practices concerning who conducts risk assessment, when, and how often.

Client-level elements are based on a cohort of youth identified by their entry into each service during a well-defined period of time (e.g., a 12-month period starting on January 1, 2013). Of course, these measures more accurately describe a service when they are based on a larger number of clients, and SPEP™ developers require a cohort of at least 10 to rate a service. It may take smaller services a considerable amount of time to accrue enough clients for a reliable rating.

Service dosage data may be produced by existing billing or other systems, but in some cases, new systems and procedures for recording client contact must be established. Whether client dosage data are systematically recorded often depends, in part, on the nature of the service contract and billing system, who pays for services, and if there are intermediary organizations.

Client-level data must be relatively complete for the period during which a program is rated. If client data are too incomplete, they cannot be taken as representative of the clientele as a whole. To that end, SPEP™ developers require individual risk data for at least 80 percent of youth in a service. And because risk is dynamic, risk data must be based on assessments conducted no more than 90 days (preferably 60 or fewer) before the start of service delivery to produce a full SPEP™ rating of that service.

To summarize, there are three requirements for valid risk data for the SPEP™: (1) a risk assessment must be done on all clients with at least 80 percent data availability); (2) risk assessments must be timely and recent, which typically means that jurisdiction must conduct repeated risk assessments; and (3) the service cohort for each service must consist of at least 10 clients who have received and terminated from the service.

If the first two conditions are met, cohorts can be made up of youth who have recently terminated and data can be assembled *retrospectively*.

Risk data are often found to be incomplete, outdated, or otherwise suboptimal. This leads to a reexamination of the risk assessment system, who conducts risk assessments and when, and the data system that captures risk data. Risk data in JJRRI sites were incomplete in various ways, as discussed in the next chapter.

SPEP™ Ratings in JJRRI Sites

Interim SPEP™ Ratings

Jurisdictions often fail to meet at least one of the first two data requirements: either risk assessments are conducted inconsistently and risk data is incomplete or assessments are not being repeated and risk data are outdated. Risk data on a service cohort can only be assembled once these issues are addressed, meaning that data for SPEP™ ratings must be assembled *prospectively* on a new cohort of youth. Accruing that prospective data can take some time, especially for smaller services.

Meanwhile, programs involved in SPEP™ implementation are often awaiting feedback after providing data on the other SPEP™ elements. For these circumstances, where data on other elements, such as dosage data, are otherwise complete, SPEP™ developers have created several types of interim

ratings that can be used to provide preliminary feedback to programs.[3] However, none of these interim ratings were conducive to a local validation of the SPEP™.

Table 1 shows the services that were rated during the first round of ratings at the JJRRI demonstration sites, along with what timely risk data was available. Both community-based and residential services were rated. In Iowa, 71 services were rated (19 community based and 52 residential).[4] In Milwaukee County and Delaware, 11 and 14 services were rated, respectively.

The JJRRI evaluation focused on the Iowa site because it alone seemed to have rated enough services to warrant examination. However, we note that only nine of the services—all community based—obtained a "full" SPEP™ rating.

TABLE 1

Services Rated on the SPEP™

		Timely Risk Data Available				
		≥ 80%	**60–79%**	**< 60%**	**< 10 youth**	**All services**
Iowa	Community	9	7	3	0	19
	Residential	0	0	52	0	52
Milwaukee County	Community	3	3	0	0	6
	Residential	1	4	0	0	5
Delaware	Community	13	0	0	1	14
	Residential	0	0	0	0	0
Total		**26**	**14**	**55**	**1**	**96**

Note: Full SPEP™ ratings require cohorts of more than 10 youth and timely risk data for at least 80 percent of clients.

3. A Case Study of SPEP™ Ratings in Iowa

Seventy-one services were rated using data on 778 youth who received juvenile justice services from 2010 to 2012. Of these youth, 381 participated in at least 1 of 19 services delivered in community settings and an additional 397 participated in at least 1 of 52 services in residential settings. Table 2 describes the sample. Demographically, the sample was 83 percent male and 32 percent black. Most participants (64 percent) were between 15 and 17 years old at the time, but 7 percent were younger than 15 and 29 percent were older than 17 by service end.[5] It is worth noting that 252 of the 397 youth who received services in residential settings were from different judicial districts than the youth in community settings in the participating judicial districts. A residential sample limited to youth from these three districts would have been just 145.

TABLE 2

Characteristics of Juveniles in SPEP™-Rated Programs

	Clients of Community-Based Services		Clients of Residential Services		Full Sample	
	N	%	N	%	N	%
Sample	381	100%	397	100%	778	100%
Judicial district						
First	178	47%	50	13%	228	29%
Third	27	7%	43	11%	70	9%
Sixth	176	46%	52	13%	228	29%
Other	0	0%	252	63%	252	32%
Gender						
Male	298	78%	347	87%	645	83%
Female	83	22%	50	13%	133	17%
Race						
Black	146	38%	100	25%	246	32%
White	201	53%	234	59%	435	56%
Hispanic/Latino	23	6%	40	10%	63	8%
American Indian	4	1%	7	2%	11	1%
Asian American	3	1%	3	1%	6	1%
Other	4	1%	13	3%	17	2%
Age at service end						
12–14	51	13%	7	2%	58	7%
15–17	256	67%	240	60%	496	64%
18–19	74	19%	150	38%	224	29%

Types of Services

Different types of services can receive different maximum SPEP™ ratings, and only a service type rated at 35 can receive the maximum total SPEP™ rating of 100. Services can also have a supplemental service, specific to each service type, that affords additional points for service type. The maximum service score for a primary service is 30, which can be raised to 35 through supplemental services.

Table 3 shows how many services of each type were rated and the number of service episodes (i.e., how often those services were used). Services are ordered by the maximum possible rating for their service type, which determines the maximum possible SPEP™ rating.[6] The 778 youth in the sample generated 2,904 service participation records, meaning that many received multiple services.[7] Group counseling was by far the most common service delivered in residential settings, and cognitive behavioral therapy was the most common in community-based settings.

TABLE 3

Types of Services Rated

	Maximum rating	Number of Services			Service Episodes		
		Community	Residential	All	Community	Residential	All
Cognitive behavioral therapy	35	8	3	11	288	71	359
Group counseling	30	0	10	10	0	384	384
Mentoring	30	0	5	5	0	223	223
Family therapy	20	2	0	2	147	0	147
Social-skills training	20	7	21	28	168	1,032	1,200
Restitution/ community service	15	1	6	7	15	233	248
Remedial academic services	15	0	3	3	0	235	235
Individual counseling	10	1	1	2	13	33	46
Job-related training	10	0	3	3	0	62	62
Total		**19**	**52**	**71**	**631**	**2,273**	**2,904**

Table 4 looks at services at the client level and shows the top-rated service for each youth. Because youth could receive multiple services, we expect that the recidivism reduction potential of the set of services to be at least as high as that of the single most effective service in the set.

TABLE 4

Types of Services Received

	Maximum rating	Top-Rated Service for Each Youth		
		Community	Residential	All
Cognitive behavioral therapy	35	186	34	220
Group counseling	30	0	104	104
Mentoring	30	0	0	0
Family therapy	20	109	0	109
Social-skills training	20	73	119	192
Restitution/community service	15	0	0	0
Remedial academic services	15	0	140	140
Individual counseling	10	13	0	13
Job-related training	10	0	0	0
Total		**381**	**397**	**778**

In addition, many youth received the same "bundle" of services, especially in residential settings. At one female residential facility, 30 of 36 girls received the same set of six services (alternative decision making, Girls Circle, individual counseling, individual skill building, levels group, and social skills), and 23 received Aggression Replacement Training.

Ratings on SPEP™ Elements

Table 5 shows the average ratings given for service type, quality, and dosage at both the service level (N = 71) and client level (N = 778). If a youth received multiple services, the client-level rating reflects the service with the highest SPEP™ rating. As a result, these ratings are somewhat more favorable. The table also expresses the average ratings as a percentage of the maximum possible rating on each element. For example, the average quality rating for community-based services is 13.7 of a possible 20 points, 68 percent of the maximum possible score. Full distributions of ratings are shown in appendix B.

TABLE 5

Average Ratings on SPEP™ Elements

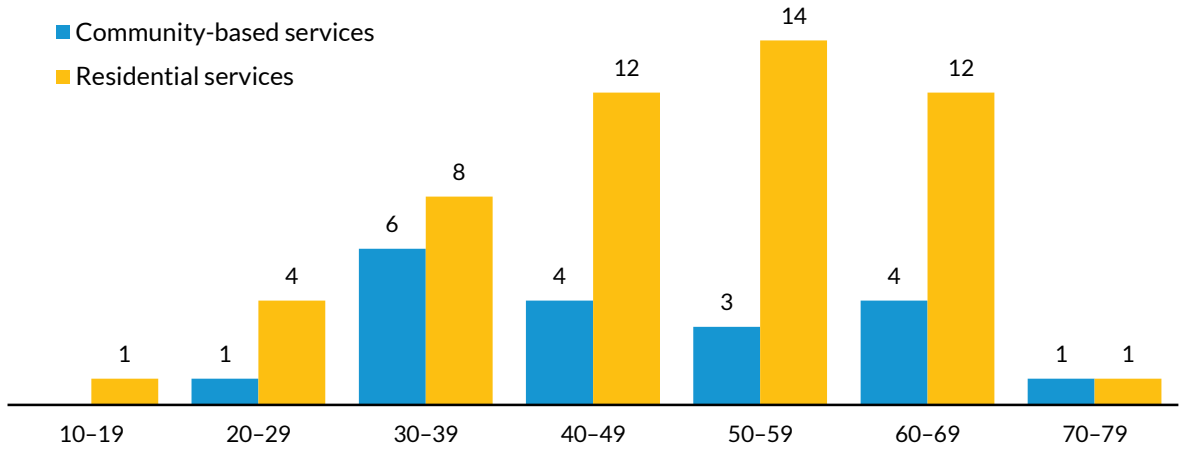
	Services			Youth		
	Community	Residential	All	Community	Residential	All
N	19	52	71	381	397	778
Service type (max 35)	25.5	22.1	23.0	27.0	22.1	24.5
Service quality (max 20)	13.7	15.0	14.6	17.3	16.3	16.8
Dosage (max 20)	13.7	15.0	14.6	17.3	16.3	16.8
As percentage of maximum rating						
Service type	73%	63%	66%	77%	63%	70%
Service quality	68%	75%	73%	87%	81%	84%
Dosage	68%	75%	73%	87%	81%	84%

SPEP™ Ratings Excluding Risk

When we put these three elements of the SPEP™—service type, service quality, and dosage—together, we get a maximum possible subscore of 75.[8] Because the risk level of clients referred to a program—the other component of a full SPEP rating—is largely out of the program's control, this subscore may be a reasonable proxy for how programs are doing within their sphere of influence. Figure 1.A and 1.B show the distribution of these ratings across services and across clients.

FIGURE 1.A

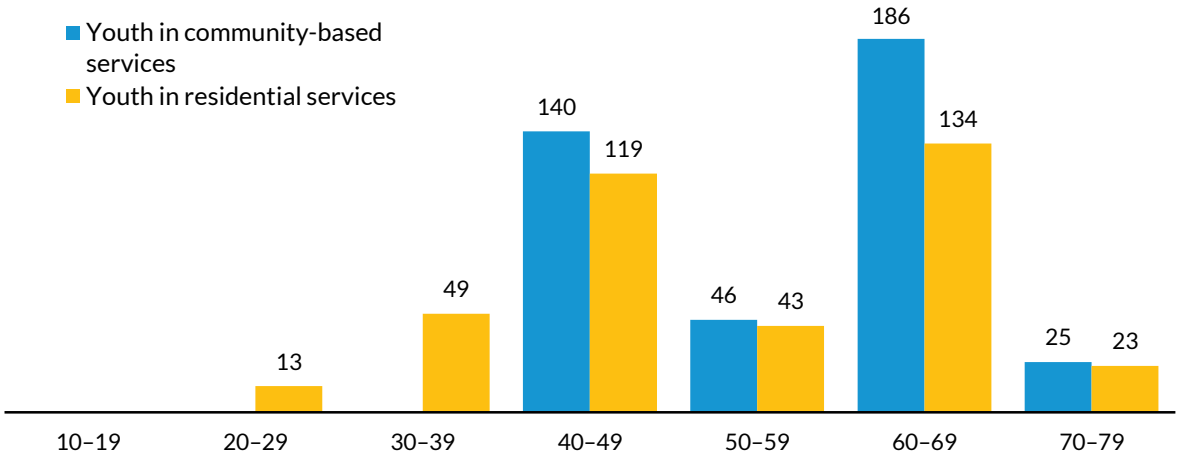
SPEP™ Ratings Excluding Risk, Service Level



Note: The total sample size was 71, divided into community-based (19) and residential services (52).

FIGURE 1.B

SPEP™ Ratings Excluding Risk, Client Level



Notes: The total sample size was 778, divided into community-based services (381) and residential services (397). SPEP™ ratings reflect the highest-rated service that each youth received.

Correlations among SPEP™ Elements

Were services that rated highly on one SPEP™ element likely to rate highly on other elements as well? Or were service elements relatively independent features? Table 6 shows the correlation among SPEP™ elements. In general, elements were positively correlated, and services of a potentially more effective type were also more likely to be of better quality and delivered at an effective dosage. Some of this reflects differences between residential and community-based services. When examined separately, residential services show larger correlations than community-based services (although the number of data points is small). The notable difference between the two is that more effective types of services were more likely to be delivered at effective dosages in residential settings (r = .68), a relationship not seen in community-based services.

TABLE 6

Correlations among SPEP™ Elements

	Service type	Service quality	Dosage
All services			
Service type	1		
Service quality	0.28**	1	
Dosage	0.12	0.15	1
Residential services			
Service type	1		
Service quality	0.28	1	
Dosage	0.68**	0.31	1
Community-based services			
Service type	1		
Service quality	0.29**	1	
Dosage	0.00	0.12	1

Notes: The total sample size was 71, divided into community-based (19) and residential services (52).
** = $p < .05$.

Risk Data in Iowa

Recall that "risk" in the SPEP™ refers strictly to the actuarial risk of recidivism, and services have more potential to reduce recidivism when they serve higher-risk youth. Thus, programs that serve more high-risk clients receive higher SPEP™ ratings.

In 2007, Iowa began using the Iowa Delinquency Assessment (IDA) to assess the risk and needs of juvenile justice youth. All youth complete a short-form IDA at intake to assess their risk level and eligibility for diversion. A more qualitative long-form IDA is given to youth for whom an adjudication is being pursued and is used in case planning. Risk information is stored in the Iowa Court Information System, a statewide data system that stores juvenile and criminal justice case information and is regularly updated by juvenile court staff.

A key issue was that quantitative risk assessment data from the short-form IDA were often too old to meet SPEP™ inclusion criteria. Some qualitative data were available from the long-form IDA, but this data could not easily be transformed into the quantitative risk data needed. TA providers worked with the Iowa site to devise proxy risk measures from the long-form data and available criminal history information to use in the preliminary round of SPEP™ ratings. Considerable effort was also made to establish consistent risk assessment practices so that data would be more readily available in future cycles of SPEP™ ratings.

As discussed in chapter 2, services cannot receive a full SPEP™ rating without timely, consistent, and complete risk data, so SPEP™ developers created interim ratings to provide preliminary feedback. Table 1 in chapter 2 displays the timeliness of risk data that were available for the first round of ratings in the three demonstration sites. Unfortunately, only about half of the community-based services in Iowa (9 of 19) and none of the residential services had enough timely risk data available to support a full SPEP™ rating.

Table 7 shows the distribution of risk among youth in the sample using the best risk data available. Among youth served by community-based programs, risk is fairly equally distributed (although risk data is missing for 10 percent of clients). Among youth in residential settings, there are relatively more high-risk youth and fewer low-risk youth as well as more missing risk data (34 percent of clients). Risk profiles seem to differ, as one would hope, between community and residential settings, but the amount of missing risk data makes that finding somewhat tentative.

TABLE 7

Assessed Risk Levels among Youth

	Clients in Community-Based Services		Clients in Residential Services		Full Sample	
	N	%	N	%	N	%
Low	113	29.66	34	8.56	147	18.89
Moderate	127	33.33	76	19.14	203	26.09
High	102	26.77	152	38.29	254	32.65
Risk data missing	39	10.24	135	34.01	174	22.37
Total	**381**	**100.00**	**397**	**100.00**	**778**	**100.00**

4. Attempting Local Validation of the SPEP™

One of the goals of the JJRRI evaluation was to conduct a local validation of the presumed relationship between the SPEP™ rating received by a service and recidivism among its clients. Reduced recidivism is the basic finding from Lipsey's meta-analysis that motivated the creation of the SPEP™. However, those studies occurred in different regions and time periods. Local validation seeks to examine the local relationship between services' SPEP™ ratings and their ability to reduce recidivism. (Here, "local" may mean a state or a smaller jurisdiction.)

Lipsey (2008) conducted such a validation on services in Arizona, which was repeated with a larger cohort by Redpath and Brandner (2010). Essentially, they found that services with higher SPEP™ ratings had lower rates of recidivism among their clients than would have been expected based on their criminal history and risk scores.

Validation Considerations

Risk Data in SPEP™ Validation

The SPEP™ attempts to rate the ability of services to *mitigate* recidivism; as such, risk data serves two purposes that are somewhat at odds. Services have more potential to reduce recidivism among higher-risk youth, and services with more high-risk clients receive higher SPEP™ ratings despite their clients being more likely to recidivate. If program A serves more high-risk youth than the otherwise equivalent program B, program A would receive a higher SPEP™ rating even if the recidivism rate among its clients is higher. As a result, simple comparisons of recidivism rates that do not control for baseline risk cannot be used to assess the utility and validity of the SPEP™.

Because controlling for the risk of recidivism is so important, prior validation attempts have also brought whatever criminal history data were available to bear on predicting recidivism. In the Iowa sample, criminal history data were not complete enough to remedy the deficiencies in the risk data per se.

Multiple Concurrent Services

How should we handle youth who participated in more than one service? They seem to muddy the waters because we are not sure how to characterize their service experience, which poses a difficult problem for local validation. Which service's SPEP™ rating should we assign? Each youth has only one outcome regardless of how many services they received.[9] As a result, "bundles" of services must essentially be reduced to one SPEP™ rating.

Lipsey (2008) and Redpath and Brandner (2010) simply excluded all youth with more than three service participation records from their validation studies regardless of whether those service records were concurrent (i.e., all during the rated period) or from throughout their criminal history.

It may make sense to handle current and prior service episodes differently. Excluding all youth with more than three prior service participation records would remove more than half of the sample and limit the validation sample to just 104 youth in community settings and 166 youth in residential settings. However, it might be defensible to ignore prior services, as subsequent system involvement generally signals the failure of those services to prevent recidivism. These services might be thought of simply as part of criminal history and risk.

Excluding all youth who received more than three services concurrently from the Iowa study would eliminate all youth in residential settings because they all received "bundles" of services. Alternatively, one might assign to each youth the highest SPEP™ rating among the concurrent services they received, as seen in the descriptive data presented above. The reasoning is that an effective service should not lose its effectiveness when supplemented by additional services.[10] This approach seems conservative when one considers that receiving multiple effective services might, collectively, be more effective than any individual service.

In view of the considerations regarding risk data and multiple concurrent services, the Iowa data do not seem capable of supporting a local validation of the expected relationship between SPEP™ ratings and recidivism. When a locality can support local validation, the unit of analysis will be another important consideration, as discussed in box 1.

BOX 1

Unit of Analysis: Service or Client?

Validation might be attempted at either the service level or at the individual client level. Data are needed at both levels, and validation requires both enough individuals and enough services.

At the service level, validation compares the client recidivism *rates* of services with high versus low SPEP™ ratings. Validation is supported if services with higher SPEP™ ratings achieve better outcomes for their clients. Lipsey (2008) and Redpath and Brandner (2010) took the following approach:

- Services were split into those with high and low SPEP™ ratings.
- Individual risk of recidivism was aggregated to each service as its predicted rate of recidivism.
- Observed recidivism rates were compared to predicted rates, for each service.
- The difference between actual and predicted recidivism rates was compared between low and high SPEP™ services.

Validation at the client level involves a correlational study of the effects of services on individual clients. Each service is considered a different treatment condition in a large effectiveness study with different clients in each condition. Individual-level regressions control for baseline risk using assessed risk and available criminal history. Validation would be supported by an association between higher SPEP™ ratings and less recidivism.

At the client level, there are more data points and thus more statistical power, but services with more clients carry more weight in the analysis. In the service-level analysis, each service's recidivism rate is equally weighted.

Comparison of Arizona and Iowa Samples

How do the available data in Iowa compare to data from prior validation reports in Arizona? Table 8 compares the Iowa sample to the Arizona samples used by Lipsey (2008) and Redpath and Brandner (2010). More extensive information is provided in appendix B.

- In Arizona, the total sample sizes were 1,490 (Lipsey) and 3,571 (Redpath and Brandner) compared to 778 in Iowa. Most of the Arizona youth received community-based services, unlike the Iowa sample, of which roughly half received residential services.
- Although the samples were much larger in Arizona, the number of services rated, 66 and 90 respectively, seems roughly comparable to the 71 services rated in Iowa. However, many services in Iowa, especially in residential settings, were delivered in "bundles."

- The available recidivism data in Iowa included new case filings in either juvenile or adult courts, but Arizona data was limited to juvenile courts. Although these recidivism measures are not comparable, for descriptive purposes the rates are shown in table 8.

In summary, the Iowa data covered considerably fewer youth. And although a similar number of services were rated, for validation purposes there were much fewer distinct services offered in residential settings. Ultimately, following the exclusion criteria used in Arizona would exclude all youth in residential settings and over half of youth in community settings.

TABLE 8

Comparison of Arizona and Iowa Samples

	Arizona		
	Lipsey (2008)	**Redpath and Brandner (2010)**	**Iowa**
Years of service receipt	2005–06	2005–07	2010–12
Geography	Five counties	Statewide	Community-based services for youth in the first, third, and sixth judicial districts. Residential services for youth statewide.
Sample	Juvenile probationers served by community-based or residential programs rated on the SPEP™.		Youth served by programs rated on the SPEP™, including youth who were not (yet) petitioned.[c]
Additional exclusion criteria	Youth with more than three service episodes (prior or current).		None
Number of juveniles[a]			
Community	1,276	3,215	381
Residential	214	356	397
Total	*1,490*	*3,571*	*778*
Number of services			
Community	53	77	19
Residential	13	13	52
Total	*66*	*90*	*71*
Measure of recidivism	New complaints for a delinquency or status offense within 6 or 12 months after end of service.		Any misdemeanor or felony offense filed in the juvenile or adult system after end of service.
Recidivism rates[b]			
6 months	27%	23%	38%
12 months	44%	38%	53%

[a] Lipsey (2008, 6); table 1 in Redpath and Brandner (2010, 4).

[b] Lipsey (2008, 14); Redpath and Brandner (2010, 9). Because of their different definitions, recidivism rates are not comparable across states.

[c] Of the SPEP™ cohort in community-based services, 72 of 381 (19 percent) had no current or prior petitions.

5. Conclusion

Iowa's SPEP™ data cannot support an attempt to correlate SPEP™ ratings with recidivism after controlling for prior risk of recidivism, which is at the heart of local validation. We reach this conclusion after considering three factors, which serve as rough guidelines for future attempts at local validation.

The first issue is that, given the centrality of risk to the analysis, Iowa's timely risk data are too incomplete. However, we note that the state has been working to improve the timeliness of its data.

The other considerations are more specific to the SPEP™ itself than to Iowa's operations. Iowa delivers residential services in "bundles," which raises questions about how to assign SPEP™ service ratings to people receiving multiple services. Under the guidelines used by Lipsey as well as Redpath and Brandner, we would not be able to include *any* youth receiving residential services in our sample. Even if we assign each youth the highest rating associated with any service they received, as we have done, youth receiving residential services are preserved for analyses but the number of residential services that can be used would be reduced to a set too small for local validation. We are then left with 19 community-based services, another fairly small set for validation.

None of this undermines the evidence base on which the SPEP™ was designed or its potential to improve the effectiveness of services. But this case study demonstrates that *local* validation of the relationship between SPEP™ ratings and recidivism at a correlational level requires a large sample of rated services delivered independently so that their effects can be distinguished. Sufficiently large cohorts of youth are needed for each service for the individual data to reliably describe the service and its clients. In addition, reliable risk data are needed to produce SPEP™ ratings and assess how risk is mitigated by services.

As part of JJRRI, SPEP™ developers worked with the demonstration sites to devise interim ratings that provide programs with feedback despite small cohorts or suboptimal risk data. These interim ratings were useful for identifying areas of potential improvement, as were the ratings on individual SPEP™ elements. However, they are a poor basis for efforts to locally validate the SPEP™. Validation efforts require adequate risk data to generate a large set of "full" SPEP™ ratings suitable for testing the ability of the SPEP™ to predict the effectiveness of services in mitigating that risk of recidivism.

Appendix A. Sample SPEP™ Scoring Sheet

Standardized Program Evaluation Protocol (SPEP™) for Services to Juvenile Offenders© Recalibrated version, 2013	Points Possible	Points Received
Primary and Supplemental Service Types [Identified based on definitions derived from the research]		
Primary Service Type for Program Being Rated Group 1 services (5 points) Group 4 services (25 points) Group 2 services (10 points) Group 5 services (30 points) Group 3 services (15 points)	30	
Supplemental Service Type Qualifying supplemental service utilized: Yes (5 points) No (0 points)	5	
Quality of Service Delivery [Determined from a systematic assessment of the relevant features of service implementation]		
Rated quality of services delivered: Low (5 points) Medium (10 points) High (20 points)	20	
Amount of Service [Determined from dosage data for the qualifying group of service recipients]		
Duration [Target number of weeks specified for each service type] % of youth who received *at least* the target weeks of service: 0% (0 points) 60% (6 points) 20% (2 points) 80% (8 points) 40% (4 points) 99% (10 points)	10	
Contact Hours [Target number of hours specified for each service type] % of youth who received *at least* the target hours of service: 0% (0 points) 60% (6 points) 20% (2 points) 80% (8 points) 40% (4 points) 99% (10 points)	10	
Risk Level of Youth Served [Determined from risk ratings on a valid instrument for the qualifying group of service recipients]		
% of youth with medium or high risk scores (greater than low): 0% (0 points) 75% (7 points) 30% (2 points) 85% (10 points) 50% (5 points) 95% (12 points) % of youth with high risk scores (greater than medium): 0% (0 points) 25% (8 points) 15% (3 points) 30% (10 points) 20% (5 points) 35% (13 points)	25	
Total SPEP™ Score	100	(Insert Score)

© Copyright held by Mark W. Lipsey, Peabody Research Institute, Vanderbilt University.
Not to be copied or used without explicit permission.

Appendix B. Iowa Ratings on SPEP™ Elements

Tables B.1, B.2, and B.3 show the distribution of ratings on SPEP™ elements other than risk at both the service level (N = 71) and client level (N = 778). Note that client-level data reflect the service with the highest SPEP™ rating among all services a youth received. As a result, SPEP™ ratings at that level are somewhat more favorable.

TABLE B.1

SPEP™ Ratings of Service Type

	Services			Youth		
	Community	**Residential**	**All**	**Community**	**Residential**	**All**
10	1	4	5	13	0	13
15	1	9	10	0	140	140
20	9	21	30	182	119	301
30	0	15	15	0	104	104
35	8	3	11	186	34	220
Total services	**19**	**52**	**71**	**381**	**397**	**778**
Average rating	**25.5**	**22.1**	**23.0**	**27.0**	**22.1**	**24.5**

Note: The highest possible score on this element is 35.

TABLE B.2

SPEP™ Ratings of Quality of Service Delivery

	Services			Youth		
	Community	**Residential**	**All**	**Community**	**Residential**	**All**
5	0	4	4	0	0	0
10	12	20	32	101	148	249
15	0	0	0	0	0	0
20	7	28	35	280	249	529
Total services	**19**	**52**	**71**	**381**	**397**	**778**
Average rating	**13.68**	**15.00**	**14.65**	**17.35**	**16.27**	**16.80**

Note: The highest possible score on this element is 20.

TABLE B.3

SPEP™ Ratings of Dosage

	Services			Youth		
	Community	**Residential**	**All**	**Community**	**Residential**	**All**
0	6	6	12	158	0	158
2	5	1	6	13	12	25
4	2	1	3	0	0	0
6	2	1	3	0	0	0
8	1	1	2	0	0	0
10	7	4	11	53	1	54
12	1	2	3	57	0	57
14	7	1	8	77	33	110
16	2	0	2	0	16	16
18	6	1	7	0	217	217
20	13	1	14	23	118	141
Total services	**52**	**19**	**71**	**381**	**397**	**778**
Average rating	**11.88**	**7.16**	**10.62**	**7.29**	**17.68**	**12.59**

Note: The highest possible score on this element is 20.

Appendix C. Comparing Iowa and Arizona Samples

This appendix presents a more complete description of the Iowa and Arizona samples.

TABLE C.1

Comparison of Arizona and Iowa Samples

	Arizona		
	Lipsey (2008)	Redpath and Brandner (2010)	Iowa
Years of service receipt	2005–06	2005–07	2010–12
Geography	Five counties	Statewide	Community-based services for youth in the first, third, and sixth judicial districts. Residential services for youth statewide.
Sample	Juvenile probationers served by community-based or residential programs rated on the SPEP™.		Youth served by programs rated on the SPEP™, including youth who were not (yet) petitioned.[a]
Exclusion criteria	Youth with more than three service episodes (prior or current).		
Age restriction	Youth who ended service at least six months before their 18th birthday.		No age cutoff; adult recidivism data available.
Number of juveniles[b]			
Community	1,276	3,215	381
Residential	214	356	397
Total	*1,490*	*3,571*	*778*
Juvenile characteristics[c]			
Male	72%	75%	82%
Black	9%	7%	31%
White	45%	48%	55%
Hispanic/Latino	41%	40%	8%
Average age at end of service	16	16	16
Number of services[d]			
Community	53	77	19
Residential	13	13	52
Total	*66*	*90*	*71*
SPEP™ service categories[e]			
Cognitive behavioral therapy	3	4	11
Family counseling	9	12	2
Group counseling	10	14	10
Individual counseling	9	18	2
Job-related training	0	0	3
Social-skills training	2	2	28

	Arizona		
	Lipsey (2008)	Redpath and Brandner (2010)	Iowa
Mentoring	1	2	5
Parent/training or counseling	0	0	0
Restitution	1	0	7
Remedial or education/academic	0	0	3
Sex offender	11	12	0
Substance abuse	20	26	0
Total	*66*	*90*	*71*
Threshold for low vs. high SPEP™ score[f]	50 out of 85	45 out of 85	70 out of 100
Services by SPEP™ rating[g]			
High rating	27%	41%	31%
Low rating	73%	59%	69%
Youth by SPEP™ service[h]			
In high-rated SPEP™ service(s)	25%	44%	54%
In low-rated SPEP™ service(s)	75%	56%	46%
Measure of recidivism	New complaints for a delinquency or status offense within 6 or 12 months after end of service.		Any misdemeanor or felony offense filed in the juvenile or adult system after end of service.
Includes status offenses	Yes	Yes	No
Adult data available	No	No	Yes
Recidivism rates[i]			
6 months	27%	23%	38%
12 months	44%	38%	53%
18 months	NA	NA	61%

[a] Of the SPEP™ cohort in community-based services, 72 of 381 (19 percent) had no current or prior petitions. This was true for 7 of 397 youth in residential services (18 percent).

[b] Table 1 in Lipsey (2008, 6); table 1 in Redpath and Brandner (2010, 4).

[c] Lipsey (2008, 8); Redpath and Brandner (2010, 6).

[d] Table 1 in Lipsey (2008, 6); table 1 in Redpath and Brandner (2010, 4).

[e] Table 1 in Lipsey (2008, 6); table 1 in Redpath and Brandner (2010, 4).

[f] Lipsey (2008, 14); Redpath and Brandner (2010, 4). Neither study rated quality, which resulted in a maximum SPEP™ score of 85.

[g] Lipsey (2008, 8); Redpath and Brandner (2010, 5).

[h] Lipsey (2008, 8); Redpath and Brandner (2010, 5).

[i] Lipsey (2008, 14); Redpath and Brandner (2010, 9).

Notes

1. The literature has not found a level of risk so high that it renders programs ineffective.
2. The SPEP™ considers risk of recidivism but not risk of mental health issues, educational failure, or any other risk. In discussing risk and needs assessment, risk is often distinguished from needs, with risk referring to actuarial risk of recidivism and needs referring to issues that could be addressed by services, such as poor family functioning, mental health issues, susceptibility to deviant peers, educational deficiencies, and so on. Different risk and needs assessment systems rate domains somewhat differently (see Hoge, Vincent, and Guy 2012), and there is considerable controversy among researchers over whether and how assessed needs should be combined with actuarial risk of recidivism (see Baird et al. 2013 and commentaries).
3. If the cohort was smaller than 10 youth, scores were considered "advisory." "Preliminary" ratings involved timely risk data for at least 60 percent of the cohort. "Provisional" ratings had timely risk data from less than 60 percent, provided that prior offense data confirmed that the youth with and without timely risk data did not differ systematically.
4. We consider services that were rated in a small pilot round as part of the first round of ratings.
5. The extended age of juvenile court jurisdiction in Iowa is 20 for youth who commit delinquency offenses before 18 (http://www.jjgps.org/iowa).
6. SPEP™ scoring has changed over time. At the time of the Arizona study conducted by Lipsey, the maximum rating for service type was 40 rather than 35 (Redpath and Brandner 2010).
7. If each record contains distinct start and end dates, a youth may have more than one participation record per service.
8. Points were distributed somewhat differently in the Arizona study, with 20 points allocated for risk rather than the current 25.
9. This problem remains whether the data are analyzed at the client level or service level.
10. This seems reasonable when limited to the therapeutic services rated by the SPEP™, although one can imagine other "services" that might undermine each other.

References

Andrews, D. A., James Bonta, and Robert D. Hoge. 1990. "Classification for Effective Rehabilitation: Rediscovering Psychology." *Criminal Justice and Behavior* 17 (1): 19–52.

Baird, Chris, Theresa Healy, Kristen Johnson, Andrea Bogie, Erin Wicke Dankert, and Chris Scharenbroch. 2013. *A Comparison of Risk Assessment Instruments in Juvenile Justice*. Washington, DC: National Council on Crime and Delinquency.

Hoge, Robert D., Gina M. Vincent, and Laura S. Guy. 2012. "Prediction and Risk/Needs Assessments." In *From Juvenile Delinquency to Adult Crime*, edited by Rolf Loeber and David P. Farrington, 150–83. New York: Oxford University Press.

Howell, James C., and Mark W. Lipsey. 2012. "Research-Based Guidelines for Juvenile Justice Programs." *Justice Research and Policy* 14 (1): 17–34.

Liberman, Akiva, and Jeanette Hussemann. 2016. *Implementing the SPEP™: Lessons from Demonstration Sites in OJJDP's Juvenile Justice Reform and Reinvestment Initiative*. Washington, DC: Urban Institute.

Lipsey, Mark W. 2008. *The Arizona Standardized Program Evaluation Protocol (SPEP) for Assessing the Effectiveness of Programs for Juvenile Probationers: SPEP Ratings and Relative Recidivism Reduction for the Initial SPEP Sample*. Phoenix: Arizona Supreme Court, Administrative Office of the Courts, Juvenile Justice Services Division.

———. 2009. "The Primary Factors that Characterize Effective Interventions with Juvenile Offenders: A Meta-Analytic Overview." *Victims and Offenders* 4 (2): 124–47.

Lipsey, Mark W., James C. Howell, Marion R. Kelly, Gabrielle Chapman, and Darin Carver. 2010. *Improving the Effectiveness of Juvenile Justice Programs: A New Perspective on Evidence-Based Practice*. Washington, DC: Georgetown University, Center for Juvenile Justice Reform.

Redpath, David P., and Jeanne K. Brandner. 2010. *The Arizona Standardized Evaluation Protocol (SPEP) for Assessing the Effectiveness of Programs for Juvenile Probationers: SPEP Rating and Relative Recidivism Reduction.* Phoenix: Arizona Supreme Court, Administrative Office of the Courts, Juvenile Justice Services Division.

About the Authors

Akiva Liberman is a senior fellow in the Justice Policy Center at the Urban Institute, where he researches and evaluates crime and justice policy, with a focus on juvenile delinquency and juvenile justice. Recent projects include evaluations of juvenile reentry programs supported through the Juvenile Second Chance Act, Restorative Justice in Rhode Island Schools, Pay for Success initiatives, and Early Access to Medicaid as a Reentry Strategy.

Jeanette Hussemann is a research associate in the Justice Policy Center. She is committed to using rigorous social science methods to inform and improve justice systems. Her research portfolio focuses on indigent defense, victimization, and juvenile justice. Hussemann directs projects that employ both quantitative and qualitative research methodologies, and she has extensive experience conducting fieldwork in diverse settings.

Statement of Independence

The Urban Institute strives to meet the highest standards of integrity and quality in its research and analyses and in the evidence-based policy recommendations offered by its researchers and experts. We believe that operating consistent with the values of independence, rigor, and transparency is essential to maintaining those standards. As an organization, the Urban Institute does not take positions on issues, but it does empower and support its experts in sharing their own evidence-based views and policy recommendations that have been shaped by scholarship. Funders do not determine our research findings or the insights and recommendations of our experts. Urban scholars and experts are expected to be objective and follow the evidence wherever it may lead.

2100 M Street NW
Washington, DC 20037

www.urban.org