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

A Report to the Juvenile Justice Services Division, Administrative Office of the Courts, State of Arizona

Prepared by

Mark W. Lipsey, Ph.D.
Center for Evaluation Research
Vanderbilt Institute for Public Policy Studies
Vanderbilt University, Nashville TN

January, 2008

Executive Summary

The Standardized Program Evaluation Protocol (SPEP) is an evidence-based rating scheme for assessing the effectiveness of programs for reducing the recidivism of juvenile offenders. The staff of the Juvenile Justices Service Division (JJSD) of the Arizona Administrative Office of the Courts began implementing the SPEP rating scheme for their contract service providers in five pilot counties in the fall of 2006 (Maricopa, Yavapai, Yuma, Pima, and Pinal). Follow-up activities in the latter half of 2007 were then aimed at prompting providers to plan program improvements that would raise their SPEP scores.

This report summarizes an analysis of (a) the initial SPEP ratings derived from the service records of juvenile probationers completing services paid by JJSD and occurring between February, 2005, and February, 2006, in the five pilot counties and (b) the 6-month and 12-month recidivism (new complaints for delinquency or status offenses) of those juveniles for whom these data were available. The main purpose of this analysis was a preliminary investigation of whether the SPEP ratings of the service programs were related to the recidivism outcomes for the juveniles they served. Programs with higher SPEP ratings were expected to show better outcomes. Early evidence of this relationship would provide support for the validity of the SPEP and the value of the JJSD initiative to implement it as a tool for program evaluation and improvement.

Recidivism rates are a function of the initial risk level of the juveniles served as well as the effectiveness of the program serving them. A low recidivism rate, therefore, is not necessarily indicative of an effective program—it may only reflect a low risk clientele. To better assess program effectiveness in relation to the SPEP ratings, recidivism rates were risk adjusted. Expected levels of recidivism for the juveniles in the study sample were predicted from their prior risk factors and demographic characteristics. The predicted recidivism was then compared with the actual recidivism for the juveniles served by each provider and the difference was used as an index of program effectiveness. The main question for the analysis reported here, then, was whether service programs with high SPEP ratings also had favorable recidivism outcomes as indicated by recidivism rates lower than predicted on the basis of the risk levels of the juveniles served.

The major findings of this investigation of SPEP ratings and recidivism were as follows:

- Despite limitations in the data available and inherent limitations in the ability to adjust recidivism for initial risk, the results of the analysis of the relationship between SPEP ratings and recidivism were encouraging. The SPEP scores showed statistically significant and relatively strong relationships with the risk-adjusted recidivism outcomes for the juveniles served by the respective service providers. ***Juvenile offenders served by providers with higher SPEP scores had lower than predicted recidivism; juveniles served by providers with lower SPEP scores recidivated at a rate closer to what was predicted. The main conclusion of this report, therefore, is that the SPEP scores are working as expected and do show promising empirical validity as guides to effective programming for juvenile offenders.***

- In light of the encouraging findings about the validity of the SPEP ratings for assessing program effectiveness, it is relevant to consider how well the Arizona programs scored on the SPEP. Of a maximum possible score of 85 (15 points remain for service quality, but this component is not yet scored in the SPEP), more than 70% of the programs that could be rated scored below 50, with most between 30 and 49. The SPEP sets a high standard and these ratings do not indicate that the Arizona programs are especially weak, but there is clear room for improvement. The SPEP component scores show that these programs generally use effective types of services and serve many juveniles of sufficiently high risk to warrant service. The greatest shortfall appears for the amount of service—the duration of service and, especially, the number of service contact hours. There is room for better targeting of high risk cases as well. ***The second major conclusion of this report, therefore, is that there is ample room for improvement in the effectiveness of the Arizona programs as indexed by the SPEP ratings. Increases in the amount of service provided, along with more focus on high risk cases, while maintaining an emphasis on the more effective types of service is likely to yield the largest effects on recidivism reduction.***

- The SPEP assessment is derived from research studies; for some of the Arizona programs there is not sufficient research to support a SPEP. Few of the general types of programs used in Arizona fall outside the scope of the SPEP but one category of those, the behavior specific programs, was notable for the large number of juveniles served—more than a third of all the cases in the research sample. Youth in these programs participated in one or two classroom sessions, as might be appropriate for low risk offenders assigned to relatively perfunctory services. The risk scores for these juveniles, however, showed only about one-fourth rated as low risk and nearly half rated as high risk. Also, the analysis of their recidivism suggested that the behavior specific programs were not especially effective. ***The final major conclusion of this report, therefore, is that the brief behavior specific programs that seem to be designed for low risk youth are most likely not very effective in reducing the recidivism of the large number of high risk juveniles referred to them. The better SPEP-rated programs showed stronger indications of effectiveness with such youth.***

Report

Background

The Standardized Program Evaluation Protocol (SPEP) is a rating instrument for assessing programs for juvenile offenders with regard to their expected effectiveness for reducing recidivism. In the fall of 2006, a version of the SPEP was adapted for Arizona programs serving probationers under a contract between the Administrative Office of the Courts, Juvenile Justice Services Division (JJSD), and the Center for Evaluation Research and Methodology at Vanderbilt University.

The SPEP is based on research studies of programs for juvenile offenders drawn from an archive of nearly 600 controlled studies of effects on recidivism assembled by Dr. Mark Lipsey, Director of the Vanderbilt Center for Evaluation Research and Methodology. Using a technique known as meta-analysis, the characteristics of the programs with the largest effects on recidivism have been identified from that research and translated into guidelines for effective interventions. Based on those guidelines, the SPEP is designed to rate programs according to how closely their characteristics resemble the characteristics shown by research to be most strongly associated with recidivism reductions.

The characteristics rated by the SPEP include the primary type of service provided, any supplemental services, duration and frequency of service, quality of service, and the risk level of the juveniles served. Figure 1 shows the rating categories and general form of the SPEP instrument. Individual programs receive a SPEP score derived from records kept by the Juvenile Justice Services Division that provide data on the type and amount of service and the risk assessment scores for the juveniles served. Program quality is not currently rated, but JJSD personnel are working on a scheme to add that soon. SPEP scores can range from 15 to 100 but, because quality of service is not currently rated and accounts for 15 points, 85 is the maximum possible score for the SPEP ratings examined in this report.

Because the SPEP rating assigns points for the respective program characteristics according to research findings about their relationship to recidivism, each program's overall SPEP score constitutes an evaluation of its expected effectiveness for reducing recidivism. The ratings for the different program characteristics, in turn, identify the areas in which a program has the greatest potential for improvement. The purpose of the Arizona SPEP, therefore, is to both evaluate the effectiveness of the programs for juvenile probationers with which JJSD contracts and guide improvements that will make those programs more effective.

The purpose of this report is to summarize the data available to date from the five pilot counties about the SPEP scores for the providers funded by JJSD for services to Arizona juvenile justice probationers, the recidivism of those juveniles after receiving services, and the relationship of the SPEP scores to recidivism. The first two components are mainly descriptive—they simply paint the picture of what the SPEP scores look like for the providers and what the recidivism rates look like for the juveniles. The third component, however, addresses the validity of the SPEP scores. If the SPEP scores are useful guides to effective programming, we should find that juveniles of

the same risk level have lower recidivism rates when served by programs with higher SPEP scores than when served by programs with lower SPEP scores.

Figure 1: General Form of the SPEP Rating Instrument

Standardized Program Evaluation Protocol (SPEP) for Services to Probation Youth		
	Possible Points	Received Points
Primary Service: High average effect service (35 points) Moderate average effect service (25 points) Low average effect service (15 points)	35	
Supplemental Service: Qualifying supplemental service used (5 points)	5	
Treatment Amount: Duration: % of youth that received target number of weeks of service or more: 0% (0 points) 60% (6 points) 20% (2 points) 80% (8 points) 40% (4 points) 100% (10 points)	10	
Contact Hours: % of youth that received target hours of service or more: 0% (0 points) 60% (9 points) 20% (3 points) 80% (12 points) 40% (6 points) 100% (15 points)	15	
Treatment Quality: Rated quality of services delivered: (Currently unscored)	[15]	
Youth Risk Level: % of youth with a risk score over .50 (medium): < 85% (0 points) 85-98% (5 points) 99+% (10 points) % of youth with a risk score over .70 (high): < 70% (0 points) 70-89% (5 points) 90+% (10 points)	20	
Provider's Total SPEP Score:	100	[INSERT SCORE]

Data Available for the Analysis

The source data for this report were (a) service records for juvenile probationers who completed service in the pilot counties (Maricopa, Yavapai, Yuma, Pima, and Pinal) between February, 2005, and February, 2006, and contributed to the initial SPEP scores for the providers of those services, combined with (b) JOLTS offense records for those juveniles through August, 2006. Some juveniles had multiple service episodes recorded. Prior services create ambiguity about the extent to which recidivism was influenced by the most recent service, which is the one under study in this analysis. Records were therefore dropped from the analysis for juveniles with more than three service episodes; for the remainder, the most recent service episode was used. Records were also dropped for juveniles who did not have sufficient time past the end of service for either 6-month or 12-month recidivism data to accumulate, and those whose age at the conclusion of service did not leave the respective 6- or 12-month interval before they turned 18.

SPEP Scores for Service Providers

The juveniles with 6-month recidivism data available for analysis were served by 66 SPEP rated programs and 6 programs that could not be rated by the SPEP. The latter provide types of service for which there is insufficient research to provide the evidence base to construct a SPEP rating. Table 1 shows the service categories into which the programs were classified by JJSD staff, the number of programs of each type, and the number of juveniles with recidivism data served by programs of that type. Overall, 59% of the juveniles in the analysis sample were in SPEP rated programs; 41% were in 'behavior specific' programs that could not be SPEP rated.

Table 1: Service Categories for Programs that Served Juveniles with at least 6-Month Recidivism Data

Service Category	Number of programs	Number of juveniles
SPEP rated programs		
Individual counseling	9	265
Group counseling, community	5	92
Group counseling, residential	5	77
Family counseling/therapy	9	334
Life skills training	2	63
Mentoring	1	49
Restitution	1	7
Cognitive behavioral, community	1	37
Cognitive behavioral, residential	2	39
Substance abuse, community	17	339
Substance abuse, residential	3	70
Sex offender, community	8	90
Sex offender, residential	3	28
Programs not SPEP rated		
Behavior specific	6	1024

The largest number of juveniles by far was served by the behavior specific programs—short-term educational programs on topics related to juvenile behavior problems for which no SPEP ratings could be made. Most of the juveniles participating in these programs attended no more than one or two classroom sessions. With such brief service, this would appear to be a program category for relatively low risk juveniles, but the risk scores from the risk assessment instrument completed by the probation officers does not bear that out. Only 24% were classified as low risk (0-.50 risk score). Of the remainder, 28% were medium risk (.51-.70 risk score) and 48% were high risk (score >.70).

For the SPEP rated programs, higher SPEP scores identify programs expected to be more effective on the basis of available research on similar programs. The total SPEP scores, as currently computed, are composed of three main component scores, each further divided into two subparts, as follows:

- Type of service (primary service and qualifying supplementary services);
- Amount of service (service duration in weeks and total number of contact hours);
- Risk level of juveniles served (proportions reaching moderate and high risk thresholds).

In addition, there is a fourth component for quality of service, worth up to 15 points, which is not currently being rated. The maximum possible SPEP score for the programs in the present analysis, therefore, is 85.

Figure 2 reports the distribution of total SPEP scores for the programs that served juveniles for whom we have 6-month recidivism data. The most notable aspect of the data in Figure 2 is the relatively low and restricted range of total SPEP scores for these programs. Of a maximum possible total score of 85, 73% of the providers scored under 50. Only 6% of the providers scored 70 or higher. For many of the programs, therefore, there is ample room for improvement.

Figure 2: Number and Percentage of Programs with Different Total SPEP Scores

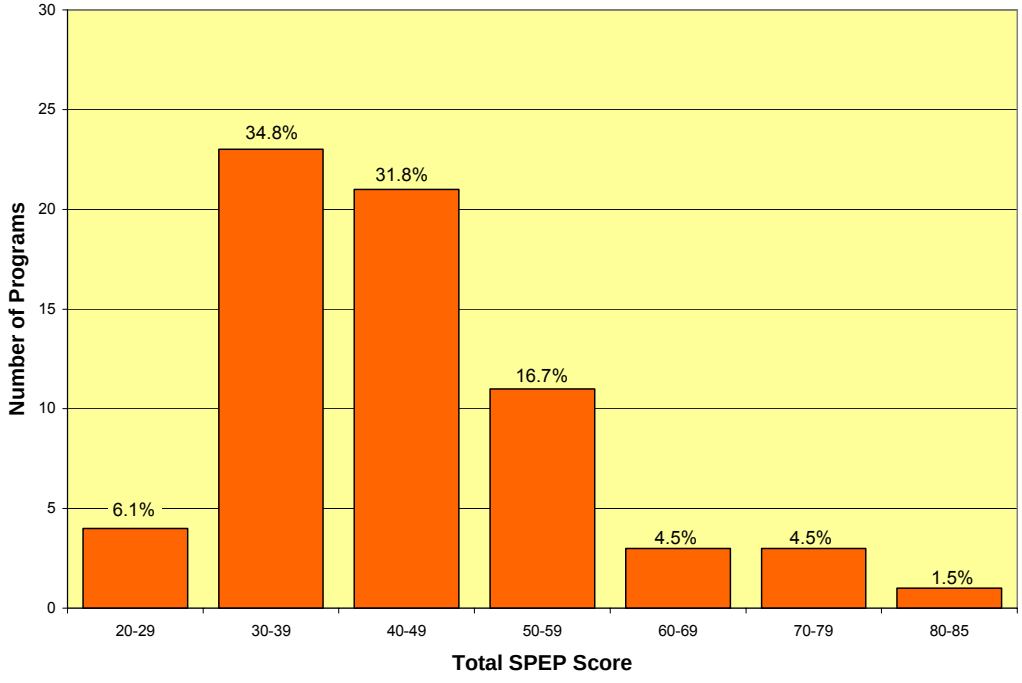


Table 2 shows the number of juveniles served by providers in the different SPEP score ranges. About 75% of the juveniles in a SPEP-rated program were in programs with ratings of less than 50; nearly 48% of the juveniles were in programs with ratings of less than 40. Only 3% were in programs with ratings of 70 or higher.

Table 2: Number of Juveniles Served by Providers with Different SPEP Scores

SPEP Score	Providers N	Providers %	Juveniles N	Juveniles %
20-29	4	6.1	168	11.3
30-39	23	34.8	543	36.4
40-49	21	31.8	400	26.8
50-59	11	16.7	275	18.5
60-69	3	4.5	60	4.0
70-79	3	4.5	22	1.5
80-85	1	1.5	22	1.5
Total	66		1490	

Table 3 reports the distributions of component scores that went into the total SPEP scores for the SPEP rated programs. These breakdowns reveal that the greatest shortfall was for amount of service, where nearly a third of the providers scored zero and more than 80% received fewer than half the 25 points available in that category. The next largest shortfall was for juvenile risk with 76% of the providers receiving half or fewer of the 20 points available. The strongest category was type of service with only 15% of the providers scoring half or fewer of the 40 points available.

Recidivism of the Juveniles Served

The following recidivism variables were used in the analysis:

- 6-month recidivism—whether any new complaint was recorded for a delinquency or status offense in the six months after the end of service (yes/no);
- 12-month recidivism—whether any new complaint was recorded for a delinquency or status offense in the twelve months after the end of service (yes/no).

The data in the full analysis sample for the five pilot counties included 1542 juveniles who were at least 12 months short of their 18th birthday and whose offense records extended at least 12 months past the date on which their service ended. The average age of these juveniles at the end of service was 15.6; 72% were male, 45% were white, 41% Hispanic, 9% African-American, and 3% Native American.

The 6-month recidivism rate for these juveniles was .27 and the 12-month recidivism rate was .44. That is, 27% and 44% of the juveniles, respectively, had a new complaint for a delinquency or status offense during the six or twelve months after the end of service. An additional 1361 juveniles were at least 6 months before their 18th birthday and had offense records for 6 months past the end date of service (but not 12 months). Combined with the 1542 juveniles above, this made a total sample of 2903 juveniles for whom 6-month recidivism data were available. Their 6-month recidivism rate was virtually the same as above, .26.

Table 3: SPEP Component Scores for 66 Service Programs for Juveniles with 6-Month Recidivism Data

Points for Service Type (Max 40)	N	%	Points for Primary Service (Max 35)	N	%	Points for Supplementary Service (Max 5)	N	%
15	3	4.5	15	10	15.2	0	46	69.7
20	7	10.6	25	21	31.8	5	20	30.3
25	18	27.3	35	35	53.0			
30	3	4.5						
35	25	37.9						
40	10	15.2						

Points for Amount of Service (Max 25)	N	%	Points for Weeks of Service (Max 10)	N	%	Points for Hours of Service (Max 15)	N	%
0	21	31.8	0	23	34.8	0	43	65.2
2-4	18	27.3	2	15	22.7	3	3	4.5
5-7	7	10.6	4	13	19.7	6	8	12.1
8-10	7	10.6	6	14	21.2	9	6	9.1
11-13	2	3.0	8	1	1.5	12	6	9.1
14-16	8	12.1						
17-20	3	4.5						

Points for Client Risk (Max 20)	N	%	Points for Lower Client Risk (Max 10)	N	%	Points for Upper Client Risk (Max 10)	N	%
0	19	28.8	0	24	36.4	0	26	39.4
5	12	18.2	5	28	42.4	5	26	39.4
10	19	28.8	10	14	21.2	10	14	21.2
15	4	6.1						
20	12	18.2						

It should be noted that these recidivism rates are for the select group of juveniles who appear in the analysis sample and are not representative of the recidivism rates for the whole population of juveniles served by the Arizona juvenile justice system. The juveniles in the analysis sample are only those from the five pilot counties who received services paid by JJSD during the restricted time windows that allowed for 6- and 12-month recidivism data to accumulate.

The Difference between Predicted and Actual Recidivism

Recidivism rates are largely a function of risk factors present before receipt of services, so it can be very misleading to compare programs with regard to the recidivism rates of the juveniles they serve. A program may be relatively ineffective but serve low risk juveniles who, therefore, have the same low recidivism rate they would have had without the program. Another program may serve high risk juveniles, who would otherwise have a high recidivism rate, and may be very

effective in reducing their recidivism. Despite the effectiveness of the program, however, the juveniles could still have a recidivism rate higher than a program that started with low risk juveniles. A relatively high proportion of high risk juveniles will recidivate despite effective services, just not as many as with ineffective services.

Under these circumstances, the only way to assess the effects of program services on recidivism with any confidence is through controlled research in which juveniles are randomly assigned to different services and their subsequent recidivism compared. Absent results from such research, the next best thing is to attempt to statistically adjust recidivism rates for prior risk using data on the juveniles' characteristics before they received service. Such adjustments amount to attempting to predict recidivism on the basis of the juveniles' prior risk-related characteristics. If those adjustments are successful, the difference between predicted recidivism and actual recidivism will then be an indication of the effectiveness of service. More effective programs should show an actual recidivism rate lower than the predicted recidivism.

The first step in the analysis of the recidivism rates for the various programs, therefore, was to identify variables relating to juveniles' prior status that were capable of statistically predicting recidivism. These variables were then used to generate a predicted probability of recidivism for each juvenile so that, for juveniles receiving different services, their actual recidivism could be compared with their predicted recidivism. The variables selected for this analysis, based on their significant contributions to the prediction of recidivism, included prior service events, county, age, sex, race, risk rating, and the number and nature of prior complaints recorded. Further details about the variables and the statistical prediction procedure are presented in Appendix A.

The predicted probability of recidivism values that resulted from this analysis for 6-month and 12-month recidivism had correlations ranging from .24 to .29 with actual recidivism. These correlations represent a predictive accuracy of about 65%; that is, the predicted recidivism value was correct for about 65% of the cases. These are modest correlations, however, though in the range typically found in validation studies of recidivism risk rating instruments. They are not such high correlations that we can be certain that all the relevant predictive factors are represented. To the extent that important variables are left out of the prediction equation, the resulting predictions may be too high or too low for some juveniles. This ambiguity should be kept in mind when interpreting the comparisons between actual and predicted recidivism for different service programs. If a program serves juveniles with unmeasured characteristics that make them lower risk than the recidivism prediction indicates, the actual recidivism will be lower because of that prediction error and not necessarily because of the effectiveness of the program. Conversely, unmeasured characteristics that make risk higher than the predicted recidivism indicates will make an effective program look less effective.

Another factor that should be kept in mind is that virtually all the juveniles in the analysis received at least some service. Thus the predicted recidivism estimate is the expected recidivism *with the average service*—the analysis does not include a "no program" control group that allows estimation of recidivism absent service. The value of these predicted rates is that we expect juveniles receiving a service that is more effective than average to show an actual recidivism lower than predicted. On the other hand, juveniles receiving a service less effective than average should show an actual recidivism higher than predicted.

Predicted and Actual Recidivism for the Primary Program Service Categories

Table 4 below reports the 6- and 12-month recidivism for juveniles served in each of the primary service categories with at least 10 cases. Alongside the actual recidivism rates in Table 4 are the predicted rates generated from the statistical procedure described above and in Appendix A. Also shown are the differences between the actual and predicted recidivism rates-- negative differences indicate that the actual rates are lower than predicted; positive differences indicate that the actual rates are higher than predicted.

For some primary service categories, the 6- and 12-month recidivism results are consistent in showing better, worse, or the same recidivism as predicted. For other services the 6- and 12-month results do not agree about how actual recidivism compares with predicted recidivism. These results indicate which types of program, *as they are currently delivered to the juvenile probationers represented in the analysis sample*, appear to be more and less effective than average. Services with similar results in this regard can be grouped as follows.

Program services that appear to be much more effective than average (actual minus predicted recidivism differences of -.10 or more for both 6 and 12-month recidivism):

- community-based cognitive behavioral services
- mentoring
- residential substance abuse services.

Program services that appear to be more effective than average (actual minus predicted recidivism differences of at least -.05 for both 6 and 12-month recidivism):

- residential cognitive behavioral services
- community-based substance abuse services
- residential sex offender services.

Program services that appear to be average in effectiveness (recidivism about the same as predicted):

- individual counseling
- family counseling/therapy

Program services that appear to be less effective than average (actual minus predicted recidivism differences between +.06 and +.09 for both 6 and 12-month recidivism):

- behavior specific services.

Program services that appear to be much less effective than average (actual minus predicted recidivism differences of +.10 or more for both 6 and 12-month recidivism):

- community-based group counseling
- life skills training.

Program services with inconsistent results (actual minus predicted recidivism differences for 6 and 12-month recidivism do not agree):

- residential group counseling
- community-based sex offender services

Table 4: Actual and Predicted Recidivism by Primary Service Category

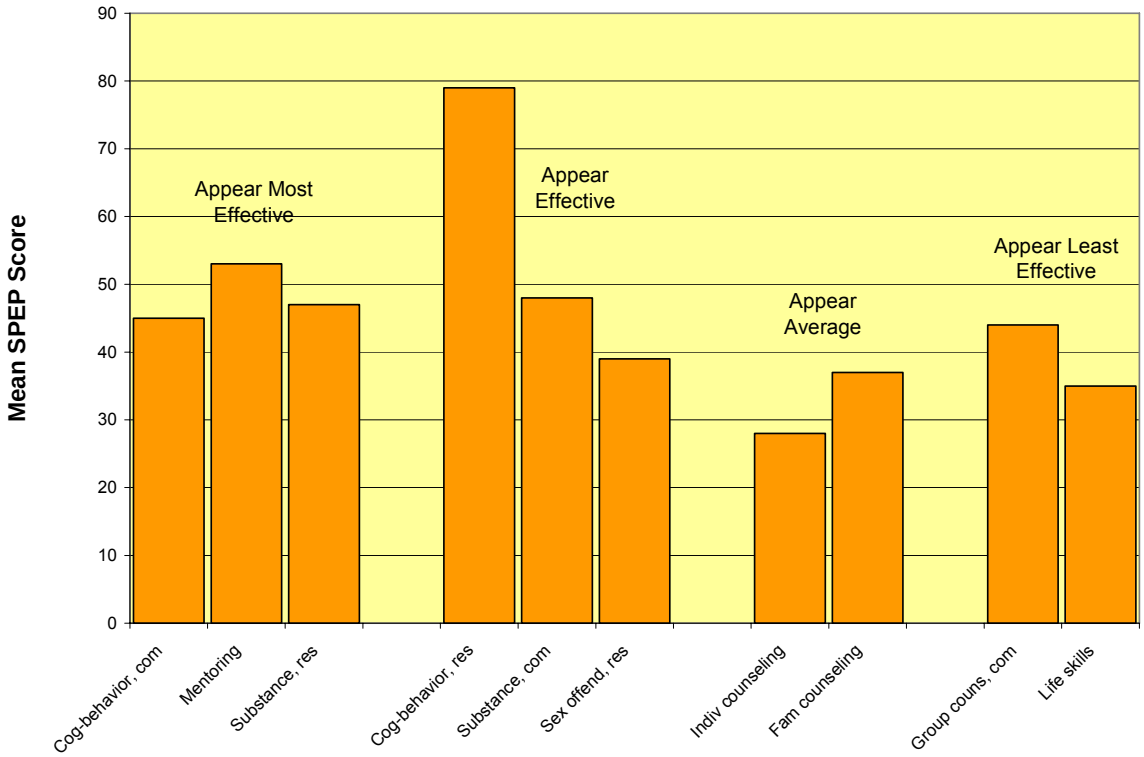
Service Category	Mean SPEP Score	6- Month Recidivism				12-Month Recidivism			
		Number of Cases	Actual Recidivism	Predicted Recidivism	Difference	Number of Cases	Actual Recidivism	Predicted Recidivism	Difference
SPEP rated programs									
Individual counseling	28	274	.25	.25	**.00**	179	.44	.46	**-.02**
Group counseling, community	44	93	.27	.11	**.16**	39	.44	.28	**.16**
Group counseling, residential	61	88	.41	.59	**-.18**	47	.64	.62	**.02**
Family counseling/therapy	37	339	.26	.26	**.00**	194	.39	.40	**-.01**
Life skills training	35	68	.18	.03	**.15**	30	.37	.27	**.10**
Mentoring	53	49	.31	.41	**-.10**	27	.37	.59	**-.22**
Cognitive behavioral, community	45	40	.25	.65	**-.40**	22	.55	.86	**-.31**
Cognitive behavioral, residential	79	39	.36	.56	**-.20**	19	.58	.63	**-.05**
Substance abuse, community	48	360	.24	.33	**-.09**	153	.48	.58	**-.10**
Substance abuse, residential	47	72	.47	.64	**-.17**	30	.60	.70	**-.10**
Sex offender, community	41	90	.08	.04	**.04**	37	.08	.22	**-.14**
Sex offender, residential	39	33	.06	.15	**-.09**	20	.05	.15	**-.10**
Programs not SPEP rated									
Behavior specific	NA	1024	.25	.18	**.07**	576	.44	.38	**.06**

How SPEP Scores for the Primary Program Services Relate to the Recidivism Results

Table 4 also shows the mean SPEP ratings for the programs included in each primary service category for which SPEP ratings can be made. Because the SPEP rating scheme is derived from research about program effectiveness, we would expect the mean SPEP ratings to be higher for program services that show lower than predicted recidivism and thus appear to be more effective for the juveniles in the analysis sample. Comparing the mean SPEP ratings with the recidivism difference scores across the primary services, therefore, should give some indication of the validity of the SPEP ratings.

Figure 3 shows the mean SPEP ratings for the program service categories identified above as those that appeared to be most and least effective on the basis of the difference between actual and predicted recidivism. As can be seen, the program services that appeared more effective did generally have higher SPEP ratings than those that appeared less effective. The picture was not entirely consistent, however. Some of these program categories had higher mean SPEP scores than their recidivism results seem to warrant (e.g., residential cognitive behavior and community-based group counseling). Others had lower mean SPEP scores than the recidivism results suggest (e.g., community-based cognitive behavior and residential substance abuse services). The number of cases on which the analysis is based is small for some of these program categories, however, so these patterns may not be stable.

Figure 3: Mean SPEP Scores for Program Service Categories that Appeared Most and Least Effective in the Recidivism Analysis



Recidivism for Service Providers with SPEP Scores

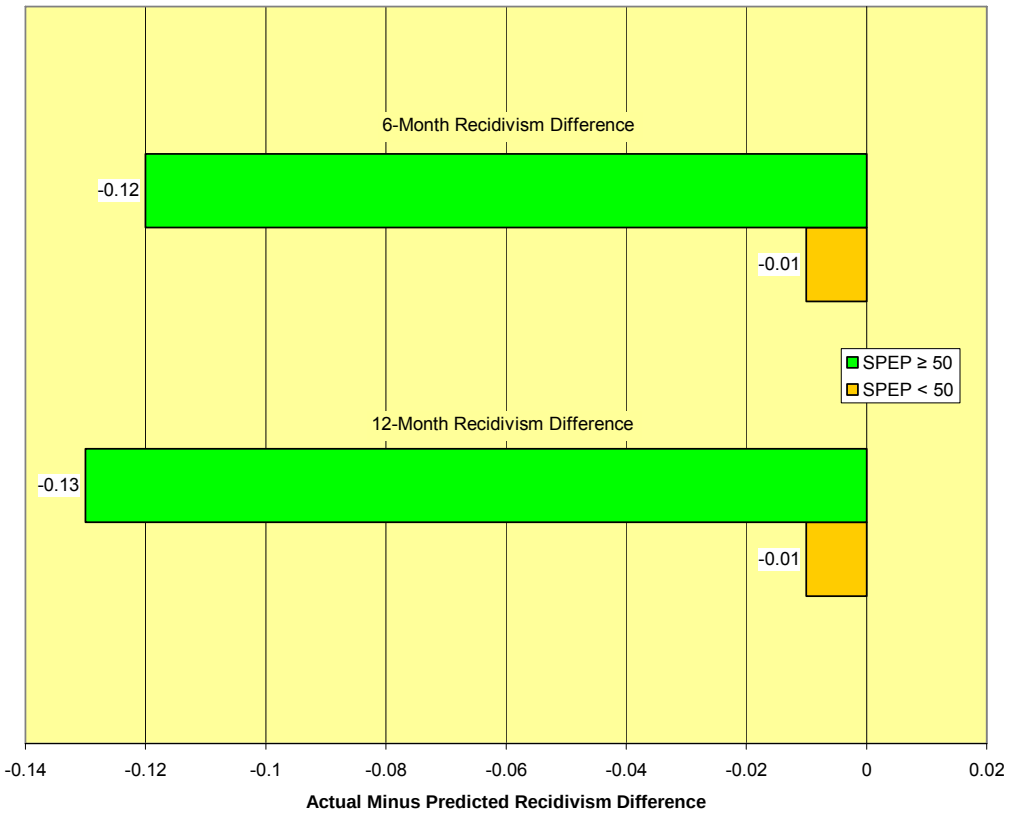
SPEP scores rate individual service providers based on the services they provide to their juvenile clients. The most informative analysis of the extent to which SPEP scores are valid indicators of program effectiveness, therefore, is one that examines the relationship between the scores for individual providers and the recidivism of the juveniles they serve. As described earlier, there are limitations to any such analysis that must be kept in mind. The SPEP ratings with which we are working are incomplete—they do not include the important component that represents the quality of the services provided. In addition, our ability to predict the expected recidivism from the available data is modest. Omitted variables in the statistical prediction model can cause over or under prediction with corresponding mis-estimation of program effectiveness. Finally, the current database provides rather small numbers of juveniles with recidivism data who were served by many of the SPEP rated programs, making any analysis of them relatively unstable. If the SPEP ratings are valid indications of program effectiveness, however, we would expect the available data to show some indication that service providers with higher SPEP scores have better risk-adjusted recidivism outcomes despite these limitations.

As shown in Figure 2, presented earlier, more than 70% of the providers in the analysis sample had a total SPEP score below 50 (85 is currently the maximum possible score). Figure 4, below, shows the mean difference between actual and predicted recidivism rates for the service providers with SPEP scores of 50 or more compared with the difference for providers with SPEP scores of less than 50 (weighted by the number of cases available for each provider). For both 6- and 12-month recidivism, juveniles served by providers with SPEP scores of 50 or higher had recidivism rates 12 to 13 percentage points lower than predicted on the basis of their prior risk. In contrast, juveniles served by providers with SPEP scores lower than 50 had recidivism rates much closer to the predicted values—only one percentage point higher than predicted. For individual providers, therefore, higher SPEP scores were rather strongly associated with larger effects on recidivism.

The relative magnitude of the SPEP-related differences between actual and predicted recidivism can be illustrated with the 12-month recidivism rates. Juveniles served by providers with SPEP scores of 50 or more had a predicted recidivism rate of .61 and an actual rate of .48, implying a level of program effectiveness that produced a 21% reduction relative to the average effects for all services. The juveniles served by providers with SPEP scores under 50, on the other hand, had a predicted 12-month recidivism rate of .42 and an actual rate of .41 implying a level of program effectiveness that produced only a 2% reduction in recidivism below the average for all providers. While the limitations of the data and analysis must be acknowledged, these results give quite positive indications of the validity of the SPEP ratings for identifying programs that are effective in reducing recidivism.

To provide a more detailed analysis, correlations between providers' SPEP scores and their 6-month and 12-month relative recidivism reductions were also computed. The correlations with the total SPEP scores were examined along with those for the SPEP component scores relating to type of service, amount of service, and risk level of the juveniles served (see Table 3, presented earlier, for the distribution of the total and component scores).

Figure 4: Difference Between Actual and Predicted Recidivism for Service Providers with SPEP Scores of 50 or More vs. Scores of Less than 50



Within the limits of the data available, these correlations can be viewed as validity coefficients for the SPEP ratings. If the SPEP scores are related to program effectiveness as expected, we should find higher scores associated with recidivism rates lower than predicted from prior risk factors. That is, there should be significant negative correlations between the various SPEP scores and the actual-minus-predicted recidivism differences across service providers. Because the SPEP scores were not designed as interval scales, nonparametric Spearman rank order correlation coefficients were used for this analysis.

Table 5 shows the resulting correlations. The first thing to notice about these correlations is that nearly all are negative, showing the expected direction of effects between higher SPEP scores and lower than expected recidivism, and many of these are statistically significant. The correlations with the total SPEP score range from -.18 to -.29, with the one involving 12-month recidivism reductions attaining statistical significance. Once again we see that juveniles served by providers with higher overall SPEP scores show lower than predicted recidivism. These results provide good support for the assumption on which the SPEP is based—that Arizona programs with characteristics more closely matching those shown effective in research studies will in fact be more effective in reducing recidivism.

The correlations in Table 5 that involve the component scores that contribute to the total SPEP ratings generally show that each of them is separately associated with relative recidivism reductions. The relationships of the scores for risk and for amount of service contact are especially consistent—all are in the right direction and many are statistically significant. The component scores for type of service are consistent and significant for supplemental service but mixed for the primary service component, showing the expected negative correlations only with 12-month recidivism reductions, not with the 6-month ones.

Table 5: Correlations across Service Providers of the SPEP Component and Total Scores with the Difference between Actual and Predicted Recidivism

	Weighted Correlations[a] with Actual-Predicted Recidivism Difference	
SPEP Scores	**6-Month Recidivism (N=66)**	**12-Month Recidivism (N=63)**
Type of Service Subtotal	**.06**	**-.19**
Primary	.08	-.13
Supplemental	-.21*	-.30**
Amount of Contact Subtotal	**-.15**	**-.30****
Weeks	-.24**	-.29**
Hours	-.03	-.20
Risk Subtotal	**-.24****	**-.13**
Lower risk range	-.27**	-.14
Upper risk range	-.23*	-.10
Total SPEP Score	**-.18**	**-.29****

** p<.05 * p<.10

(a) Spearman rank-order correlation weighted by number of juveniles served by each provider.

Conclusions

The primary purpose of the analyses summarized in this report was to make a preliminary investigation of the validity of the SPEP rating scheme as an evidence-based assessment of the effectiveness of programs for reducing the recidivism of juvenile offenders. It is only a preliminary investigation because the data available were limited to baseline SPEP ratings of providers in the five pilot counties during the first year of SPEP initiation, most of which were based on relatively modest numbers of juvenile service records for closed cases. Moreover, complete recidivism data were not available for all of the juveniles represented in those service records. SPEP ratings are generated for individual service providers from their service records and neither those ratings nor the recidivism rates for the juveniles served by those providers are fully representative of program performance when based on relatively small numbers of closed cases.

Aside from the data limitations, there are inherent limitations in the analyses conducted for this report. Juveniles are not randomly assigned to service providers so we cannot assume that the youth they serve have equal risk of recidivism. Lower recidivism rates, therefore, do not necessarily indicate better program performance—they may only indicate that a provider served

lower risk juveniles. To help level the playing field for recidivism comparisons across programs with different SPEP scores, the expected recidivism for the youth served was predicted from a battery of prior risk and demographic factors and actual recidivism was then compared with predicted recidivism. The recidivism rates for programs more effective than average should be notably lower than those predicted for the juveniles they serve, and the recidivism for programs less effective than average should be the same or higher than predicted. Though this technique should make recidivism outcomes more comparable across programs, it is far from perfect. Recidivism is difficult to predict and we have no assurance that all the relevant risk factors have been accounted for in the procedures used in the analyses reported here. On the other hand, there is little reason to believe that whatever bias remains in the estimated recidivism outcomes is likely to be systematically related to SPEP scores. That would require juveniles less likely to recidivate than predicted to be more likely to be assigned to service providers with high SPEP scores. SPEP scores were not available at the time these juveniles were assigned to services and there is no obvious characteristic of higher scoring services that would stimulate referrals of juveniles with unmeasured favorable risk profiles across the multiple counties and many providers represented in the available data.

In light of these various considerations and limitations, the results of the analyses reported here are quite encouraging. The SPEP scores do show statistically significant and relatively strong relationships with the risk-adjusted recidivism outcomes for the juveniles served by the respective service providers. ***Juvenile offenders served by providers with higher SPEP scores had lower than predicted recidivism; juveniles served by providers with lower SPEP scores recidivated at a rate closer to what was predicted. The main conclusion of this report, therefore, is that the SPEP scores are working as expected and do show promising empirical validity as guides to effective programming for juvenile offenders.***

Given indications that the SPEP ratings have sufficient validity to guide programming, it is relevant to consider how well the Arizona programs score on the SPEP. With a maximum possible score of 85 at present (15 points remain for service quality, but this component is not yet scored in the SPEP), more than 70% of the programs that could be rated scored below 50, with most between 30 and 49. It is not surprising that real world programs do not match the characteristics of the best programs represented in research studies of program effects, so these ratings should not be taken to indicate that the Arizona programs are especially weak. Nonetheless, there is clearly room for improvement and the preliminary validity results for the SPEP ratings suggest that improving the SPEP scores should translate into greater effects on recidivism. The breakdown of the SPEP component scores shows that the Arizona programs are generally using effective types of services and serving some juveniles of sufficiently high risk to warrant service. Though there is room for improvement in these areas, especially with regard to the targeting of high risk juveniles, the greatest shortfall in the SPEP scoring appears for the amount of service—the duration of service and, especially, the number of service contact hours with the juveniles served. These components of the SPEP showed solid correlations with relative recidivism reductions so there is every reason to believe that increases in the amount of service provided will yield larger effects on recidivism. ***The second major conclusion of this report, therefore, is that there is ample room for improvement in the effectiveness of the Arizona programs as indexed by the SPEP ratings. Increases in the amount of service provided, along***

with more focus on high risk cases, while maintaining an emphasis on the more effective types of service is likely to yield the largest effects on recidivism reduction.

Not all of the Arizona programs could be rated in the SPEP scheme. The SPEP assessment is derived from research studies of each program type and for some programs there is not a sufficient body of research to support development of a SPEP. Relatively few of the general types of programs used in Arizona fall outside the scope of the SPEP but one category of those, the behavior specific programs, was notable for the large number of juveniles served—more than one-third of all the cases in the research sample analyzed for this report. Youth in these programs participated in one or two class sessions, as might be appropriate for low risk offenders assigned to relatively perfunctory services. The risk scores provided by the probation officers on the risk assessment instrument for these juveniles, however, showed ratings that were as high as those for many juveniles referred to more intensive SPEP-rated services with nearly half rated in the high risk category. Moreover, the analysis of the difference between the actual and predicted recidivism of these juveniles suggested that the behavior specific programs were not especially effective. ***The final major conclusion of this report, therefore, is that the brief behavior specific programs that seem to be designed for low risk youth are most likely not very effective in reducing the recidivism of the large number of high risk juveniles referred to them. The better SPEP-rated programs showed stronger indications of effectiveness with such youth.***

The staff of the Juvenile Justice Services Division has vigorously implemented the SPEP assessment scheme and has firmly and constructively encouraged service providers to use those assessments to guide program improvements. Their impressive efforts in this regard are in large part motivated by the belief that this process will lead to better use of the programs currently available, incremental development of more effective programs, and, ultimately, lower recidivism rates and greater public safety. Nothing in the preliminary investigation of the relationship of the SPEP to recidivism reductions summarized in this report suggests that these objectives are unrealistic. On the contrary, the findings of this investigation are very encouraging in their support of the potential value of the SPEP assessment for identifying effective programs and providing guidance for program improvements.

Appendix A: Computing Predicted Recidivism

Predicted recidivism values for each juvenile were created with a logistical regression analysis in which a selected set of predictor variables was regressed, in separate analyses, on the dichotomous 6-month and 12-month recidivism variables. The following variables were found collectively to be the best predictors:

- Number of prior service events (about 20% of the juveniles had one or two service events prior to the one analyzed; those with more were dropped from the analysis);
- Total cumulative months of prior service;
- County—dummy codes for Counties 11, 12, 14, and 15 with County 8 omitted as the reference group;
- Age at end of service;
- Sex;
- Race—dummy codes for Black, Hispanic, and Indian with White omitted as the reference group;
- Risk rating from the Risk Assessment Instrument completed by the probation officer (with maximum likelihood imputation of the missing values);
- Selected individual items from the Risk Rating instrument-- Juvenile's Relationship with Family, Drug Involvement, Truancy, School, Runaway, Probation Officers' ratings of the likelihood of reoffense;
- Age at first prior offense;
- Log of the total number of prior complaints (including VOP);
- Log of the number of prior VOP counts;
- Mean Severity index of prior offenses (Type x Class codes rescaled);
- Mean Severity of the prior complaint dispositions;
- Total number of prior days detained.

These predictors were used in separate logistic regression analyses to predict 6-month and 12-month recidivism. The predicted values (predicted probability of recidivism) that resulted from these analyses were then dichotomized at the points that provided the same recidivism proportions as the actual recidivism being predicted and recoded for each juvenile as 1 if recidivism was predicted and 0 if it was not. That dichotomized variable was then used to generate the predicted recidivism rates for the various groups of juveniles examined in the analysis. The correlation of the predicted and actual recidivism values in each case were statistically significant with correlation coefficients as follows

- 6-month recidivism: .24 (correctly predicted for 71% of the cases)
- 12-month recidivism: .29 (correctly predicted for 65% of the cases).

As an example, the results of the logistic regression analysis for predictors of 12-month recidivism are presented in Table A1 below. This analysis includes 1542 records. The overall prediction model is statistically significant with a Chi-Square of 188.8 (df=24), $p<.001$. It is notable that one of the best single predictors in this model is the risk score from the Risk Assessment instrument completed by the probation officers and used in the SPEP ratings. (The PO's risk ratings by themselves correlate .18 with 6-month recidivism and .19 with 12-month recidivism).

Table A1: Logistic Regression Results for Predicting 12-Month Recidivism

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	priorservice_number	.204	.203	1.009	1	.315	1.227
	priorservice_months	-.148	.069	4.566	1	.033	.863
	County11	.570	.201	8.036	1	.005	1.769
	County12	-.426	.241	3.132	1	.077	.653
	County14	-.191	.243	.620	1	.431	.826
	County15	.189	.224	.716	1	.397	1.208
	Age	-.161	.061	7.016	1	.008	.851
	R_Sex	.914	.136	45.380	1	.000	2.495
	Race_black	.399	.198	4.064	1	.044	1.490
	Race_hispanic	.248	.124	4.024	1	.045	1.282
	Race_indian	.474	.328	2.092	1	.148	1.606
	RiskScore	2.155	.654	10.854	1	.001	8.624
	juvrel	-.047	.142	.109	1	.742	.954
	drugs	-.091	.156	.339	1	.561	.913
	truancy	-.394	.144	7.485	1	.006	.674
	school	-.048	.123	.152	1	.697	.953
	runaway	.187	.154	1.480	1	.224	1.206
	reoffend	-.058	.100	.337	1	.561	.943
	P_CAge_first	.035	.039	.788	1	.375	1.036
	log_NTotComplaints	.661	.176	14.055	1	.000	1.937
	log_VOPcts	-.181	.090	4.080	1	.043	.834
	P_CSev_index_mean	-.002	.025	.008	1	.930	.998
	P_CDispo_Sev_mean	-.082	.061	1.783	1	.182	.922
	P_Days_Detained	-.005	.001	14.295	1	.000	.995
	Constant	-1.290	1.056	1.492	1	.222	.275