A Validation of SPEP™ in Pennsylvania

PCCD Grant ID: 28121

Final Report

Edward P. Mulvey, PhD
Carol A. Schubert, MPH
Bobby Jones, PhD
University of Pittsburgh School of Medicine

Samuel Hawes, PhD
Florida International University

Updated February 21, 2020

Table of Contents

FIGURES

TABLES

I. INTRODUCTION

This report provides a validation test of the Standardized Program Evaluation Protocol (SPEP™) as it has been applied in Pennsylvania since 2013. The analyses presented here have integrated two data sources: 1) centralized records regarding service provision kept as part of the SPEP™ implementation process in multiple counties across the state, and 2) individual level information about adolescents who were enrolled in services that took part in the SPEP™ process. This set of analyses contribute to a larger effort, the Juvenile Justice Systems Enhancement Strategy (JJSES), to build an infrastructure for more effective juvenile justice services statewide. The current analyses provide information about the relations between SPEP™ ratings and recidivism, as well as suggestions for refining this process. This report represents one of only a few empirical examinations of the SPEP™ process and impact across the nation.

I.A. The SPEP™

The Standardized Program Evaluation Protocol (SPEP™) is a standardized measure of intervention effectiveness developed by Dr. Mark Lipsey and his colleagues. This approach is based on research findings from quantitative analyses ("meta-analyses") of the literature on the aspects of service provision for juvenile offenders that are associated with reduced likelihood of re-arrest (Lipsey, 2009; Lipsey and Howell, 2012). The meta-analyses underpinning this approach a) code information from a large data base of reports on programs aimed at reducing recidivism in juvenile justice-involved youths, and b) quantitatively analyze the created data to identify the distinguishing features of programs that have the largest impact on criminal offending after program involvement. The identified features (e.g., quality of service delivery) then serve as the framework for a detailed rating system that can be applied consistently to a wide range of programs. The logical belief is that programs that adopt and practice more of the identified features in their program approach should have a larger impact on recidivism. It is thus also expected that programs that improve their SPEP™ ratings over time on these factors should see a concomitant improvement in reducing recidivism of their youths served.

This system has been put into the field and refined over the last two decades to promote quality improvement in juvenile justice services. Using direct observations, interviews, and file reviews, trained raters assess service provision and program operations regarding adherence to the identified practices (Lipsey, 2009). The dimensions of program operations that are assessed include: a) program philosophy (service type), b) amount of service (dosage and duration), c) quality of service, and d) the risk level of youth served by the program. Using the SPEP™ scoring system, the raters assign points to reflect how closely each characteristic aligns with similar programs shown to have the best recidivism outcomes. A SPEP™ total score is then derived to reflect the overall performance of the service. An action plan is devised for the service provider to improve in areas with low ratings.

In sum, the SPEP™ gives service providers a research-based profile of their service as it compares to other similar programs along several specific dimensions related to recidivism

reduction. For juvenile justice system administrators, the SPEP™ provides an overview of the array of services adopted at the system level, a "best practice" standard against which to evaluate those programs, and a roadmap for improving system-level outcomes related to recidivism.

I.B. Existing research regarding SPEP™

As noted above, the SPEP™ is rooted in extensive, sound analyses of the outcome literature regarding juvenile justice interventions. In addition, the SPEP™ approach to quality improvement on dimensions of service provision is broadly applicable to a range of service protocols and makes sense to practitioners and policy makers alike. It is a systematic, empirically grounded, common sense approach. Nonetheless, perhaps because of the sustained commitment it takes at a policy level to put this system into place, there are few assessments of its application. However, higher program scores have been associated with greater recidivism reductions in statewide evaluations conducted in Arizona and North Carolina (Lipsey, 2008; Lipsey, Howell, & Tidd, 2007; Lipsey, Howell, Kelly, Chapman, & Carver, 2010; Redpath & Brandner, 2010).

In a 2008 report to the Juvenile Justice Service Division (JJSD) of the Arizona Courts, Lipsey summarized results from an analysis of services for JJSD in five pilot counties in the year February 2005-2006. The goal of the investigation was to determine if SPEP™ ratings of these programs (absent quality indicators which had not been developed at the time of the report; Lipsey, 2018) were related to recidivism outcomes for the juveniles they served. Taking pre-existing risk into account, Lipsey found statistically significant and relatively strong relationships with recidivism outcomes for the juveniles served. Juveniles served by providers with SPEP™ scores greater than 50 had recidivism rates 12-13 percent lower than predicted on the basis of their assessed level of risk. Juveniles served by providers with lower SPEP™ scores, however, recidivated at rates that were not different (within one percentage point) of their predicted recidivism rate.

Bolstered by this early validation of the SPEP™, Arizona expanded the implementation of SPEP™ to all fifteen counties. A 2010 report (Redpath & Brandner, 2010) indicated continued support for the initial observations reported by Lipsey. Within this larger sample, youth served by providers with higher SPEP™ scores had lower risk-adjusted recidivism rates, while youth served by providers with lower SPEP™ scores recidivated at a higher rate than their risk-adjusted predicted rate.

Both Arizona reports note design and data limitations of their studies. First, there was limited sampling of services across the spectrum of possible SPEP™ scores. The SPEP™ scores in the Arizona samples skewed somewhat to the lower end of the possible scores. There were relatively small numbers of high scoring services (about 75% of the juveniles in a SPEP™ rated program were in programs with ratings of less than 50), making the estimates of differences along the higher end of the scale less reliable. Second, there were some gaps in the available data. Juvenile service records were missing for a sizable proportion of closed cases, and the SPEP™ rating system used at the time did not include a rating for service quality (now a

standard dimension of the SPEP™ process). Finally, the indicator for measuring recidivism was a new complaint for a delinquency or status offense within six or twelve months of the service end date. This meant that cases included in the analysis had to have completed the service at least one year prior to their 18th birthday, when they would have left the jurisdiction of the juvenile court. As a result, the sample of adolescents followed up were slightly younger (mean = 15.6 years old) than a representative sample of adolescents who generally qualify for court services.

The North Carolina SPEP™ Project was initiated in October 2001 to evaluate state-funded juvenile offender programs for continued funding. After an initial pilot phase, the SPEP™ was implemented statewide in 2006. Each juvenile offender program was classified and rated using the SPEP™ definitions and existing electronic tools available in North Carolina (i.e., a client-tracking system, a validated risk assessment instrument, and offender management tools). A validation of the overall SPEP™ scores with recidivism was conducted with 113 community-based programs for court-supervised juvenile offenders and 50 prevention programs. Lipsey and colleagues (2007) computed risk-adjusted recidivism rates based on risk and prior delinquency history. The validation study found that the SPEP™ scores were moderately correlated with the risk-adjusted recidivism rates, with larger relationships found for the court supervision cases than for the prevention cases.

These studies provided initial validation of the SPEP™, and additional locales began to adopt this approach. The unavoidable limitations of these studies, however, highlighted the need for continued attempts to validate the protocol. Toward that end, the Office of Juvenile Justice & Delinquency Prevention (OJJDP) issued grant funds in 2012 to the Urban Institute to evaluate outcomes related to the implementation of the SPEP™ in three demonstration sites in Delaware, Iowa, and Wisconsin.

These analyses never came to fruition. Difficulties across the sites in mounting a successful implementation of SPEP™ and significant issues with the validity and completeness of the available data made it impossible for the investigators to complete their planned validations. Instead, the evaluation focused on insights about lessons learned during the evaluation process. The major points emphasized were the importance of established local support to successful implementation, the need for adequate time to get the SPEP™ process in place and operating with integrity, and the ongoing need for a high level of technical support to ensure operational success and valid data. The evaluators stressed this latter point emphatically; attempts to assess the recidivism impact of services at various levels of SPEP™ ratings requires reliable data for a large cohort of youths (Liberman & Hussemann, 2016; 2017).

We have been unable to find documentation of subsequent research efforts to validate the total SPEP™ ratings. However, some aspects of the SPEP™ ratings have been examined. Baglivio and colleagues (2018) report on the relationship of service quality ratings (one component of the SPEP™TM score) to recidivism of adolescents served by 56 residential programs for juvenile offenders in Florida. The state of Florida has a centralized, statewide

juvenile justice system with uniform standards and accountability measures as well as a group of qualified, active researchers (thus making it a favorable place for this type of research).

Quality ratings were assigned annually by state agency staff and were determined using a standardized measure of treatment quality. The Florida quality rating is more extensive but subsumes many of the aspects of care rated in the SPEP™ quality score. Adolescents in the follow-up sample participated in one of the selected programs during the one-year period reflected in the quality rating. These investigators found a strong relationship between the quality score and three recidivism indicators; higher quality scores decreased the likelihood of arrest, conviction and reincarceration in the year following release. For every point increase in the average treatment quality, the odds of recidivism were reduced by 11%.

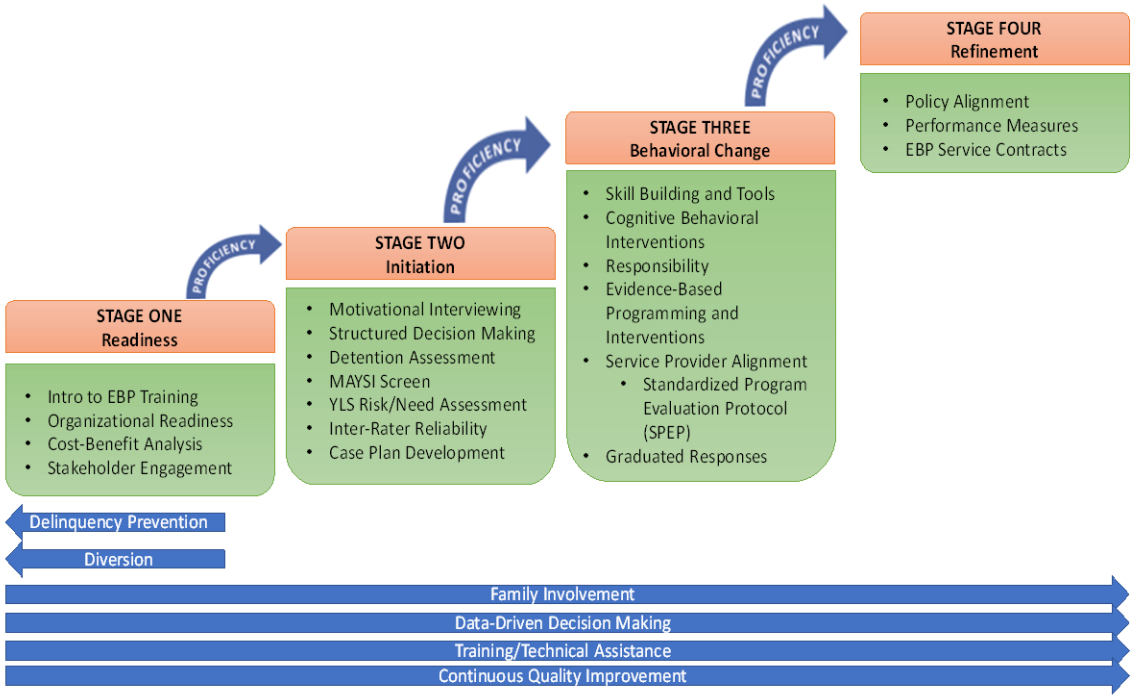
In summary, the existing literature suggests that the SPEP™ holds considerable promise as a method for rating services along dimensions related to more favorable recidivism outcomes. As mentioned earlier, this approach distinguishes itself as a method for program improvement in juvenile justice because it is derived from analyses of large bodies of research. In addition, it has made the translation of these findings into a usable set of practices in the field. It is an empirically sound, practical, and "scalable" method for evaluating juvenile justice services.

This process still needs closer examination in the field, however. The evidence of its utility so far validates this approach (showing programs with higher overall SPEP™ scores or higher quality scores with better recidivism outcomes), but the evidence is still somewhat scant. Examining whether these findings hold up in other applications of the methods and samples of adolescents examined would be worthwhile in expanding the base of evidence to support its general utility. In addition, further research on this approach can illuminate the limitations and potentials of SPEP™. It could point toward ways to refine this impressive approach to assessing juvenile justice services into effective operations for program improvement.

I.C. The SPEP™ initiative in Pennsylvania

In Pennsylvania, the SPEP™ is a component of a larger statewide strategy to improve juvenile justice services called the Juvenile Justice Systems Enhancement Strategy (JJSES). The goal of the JJSES is to create a more data-driven and effective system of juvenile justice practices and service provision. In the early 2000's, juvenile justice leadership in Pennsylvania began developing this strategic plan to implement the principles inherent in the state's Balanced and Restorative Justice model. As shown in Figure 1 below, an overlapping set of initiatives were identified as critical components to the reform of the Pennsylvania juvenile justice system.

Figure 1. The Pennsylvania Juvenile Justice System Enhancement Strategy



There are four stages in this system enhancement strategy, with the successful completion of one stage laying the groundwork for implementation of the next stage. Stage One focuses on the careful planning and training necessary to promote the successful implementation of new strategies and programs. New approaches to assessment and intervention were only going to be successful to the extent that juvenile justice professionals and service providers endorsed the overall effort and direction of change and were familiar with the tools that would be needed to move ahead. Stage Two involves the initiation of new procedures and the use of assessment tools by juvenile justice departments to prepare for behavioral change practices that are effective in reducing the risk to reoffend. The framework and adoption of systematic assessment in Stage Two is a necessary precursor for focusing resources on the "right" adolescents. Stage Three builds from the information amassed from the assessment tools and new procedures established in Stage Two. Clearly, the widespread adoption and implementation of risk/need assessment in juvenile justice jurisdictions across the state of Pennsylvania was essential to move toward implementation of SPEP™ throughout the state as part of Stage Three.

The implementation of the SPEP™ in Pennsylvania was coordinated and staffed by the Evidence-based Prevention and Intervention Support Center (EPISCenter) at Pennsylvania State University. The EPISCenter has worked closely with juvenile probation departments and service providers to prepare for, conduct, and interpret results from the SPEP™. In conjunction with an advisory group, personnel from the EPISCenter have trained individuals to conduct the SPEP™ evaluations, provided technical assistance to counties and providers, tracked the progress of these efforts, and organized the data collected about both the SPEP™ scoring and the

implementation process. This project has required continued commitment and effort from a large group of professionals statewide.

The EPISCenter reports indicate that 102 community based-services and 166 residential services completed an initial SPEP™ assessment as of late August 2019. The current evaluation uses available data compiled as part of these assessments. A small subset of these assessed services has undergone a second SPEP™ evaluation ("reassessment 1", n = 42). The available data from these reassessments are used to examine the effects of change over time in the same service. Although providing the potential for a more stringent analysis of the possibilities of using SPEP™ as a quality improvement tool, the small number of services with repeated measures limits the types of analyses that can be conducted with these data. We provide figures later in this document regarding the number of initial assessments examined and the number of within-service comparisons that can be conducted (see Section III).

I.D. The SPEP™ validation study in Pennsylvania

The validation study was initiated by Pennsylvania Council of Chief Juvenile Probation Officers, the Juvenile Court Judges Commission (JCJC), and the Pennsylvania Commission on Crime and Delinquency (PCCD) to inform juvenile justice stakeholders about the impact of the statewide implementation of SPEP™. In July 2017, this group enlisted the assistance of Edward P. Mulvey, Ph.D., a Professor of Psychiatry from the University of Pittsburgh, to examine how the efforts at implementing the SPEP™ might be related to reduced recidivism in Pennsylvania. Dr. Mulvey has worked with juvenile and adult criminal justice professionals throughout the state to examine and plan alternative services. He has also conducted large scale research studies involving individuals with mental health problems, violence, and justice system involvement (both juveniles and adults).

Ed Mulvey enlisted three other experienced researchers to assist in this project. Carol Schubert, MPH, oversaw the management and direction of the study as well as the specifics of data management for the project. Carol has extensive experience with project management, administrative data bases, data base management, data analysis, and research publication of project results. Bobby Jones, Ph.D., conducted statistical analyses of the effects of SPEP™ on recidivism. Bobby has served as a research scientist statistician for a wide range of projects (from behavioral genetics to criminal offending patterns) at both Carnegie-Mellon University and the University of Pittsburgh; he has also provided statistical consulting services to researchers around the world over the last 20 years. Samuel Hawes, PhD, an assistant professor in the Department of Psychology at Florida International University, also contributed to several aspects of the analyses. Sam has worked closely with Ed Mulvey and Carol Schubert on prior PCCD-funded research and other projects. His current program of research focuses the onset, maintenance and desistance from maladaptive, high-risk behaviors (e.g., antisocial behaviors, substance use) in adolescence, using advanced statistical modeling techniques. This group constituted the validation study team.

The validation study team consulted with Mark Lipsey, PhD, regarding several aspects of SPEP™ scoring and the analytic techniques used. Mark provided valuable insights about the origins of many of the concepts used in the development of the SPEP™ approach as well as the statistical approaches taken in some earlier validation efforts. While the team discussed some of the preliminary findings from the analyses with Mark Lipsey, Mark had no access to the data and did not produce any of the findings presented here. He served as an invaluable consulting resource and generous colleague.

I.E. Goals of the validation study

The primary question for the validation project was to determine if, and how, SPEP™ program ratings are related to recidivism of the adolescents receiving the rated programs. As noted earlier, this question has been addressed in some prior research, but only to a limited extent. None of the existing analyses have examined the potentially differential effects of the ratings of the dimensions of the SPEP™ score or the effects of service improvement on either the overall rating or the dimension scores on recidivism. This project provides more extensive information than currently available about the general associations of SPEP™ ratings and individual adolescent outcomes. More relevant to current Pennsylvania efforts, though, the analyses document how SPEP™ ratings are related to outcomes in the juvenile justice systems in counties across the Commonwealth and identify ways to focus SPEP™ practices in ongoing, future JJSES efforts in Pennsylvania in particular.

II. DATA SOURCES AND VARIABLES

II.A. Available data sets

Construction of samples regarding SPEP™ scores and cohorts of individual youths who were served by the rated services required the efforts and cooperation of two university-based groups and a state agency. Data for this study came from the EPISCenter at Pennsylvania State University and the Pennsylvania Juvenile Case Management System (JCMS), which is managed by the Juvenile Court Judges Commission (JCJC). The validation study team at the University of Pittsburgh performed the data analysis. The consolidation of data from the EPISCenter and JCJC produced a single data set containing information about a range of services and the characteristics and outcomes of groups of individuals who received those services at a particular time. This was only possible through the collaborative work of the institutions and their affiliated staff over an extended period.

In late September 2018, the EPISCenter provided the validation team with a data file that included service-level SPEP™ information (see Appendix A for a list of variables contained in the initial dataset). At the same time, the EPISCenter provided JCJC with identifying information for youths participating in the services at the time the SPEP™ ratings of these services were assigned (the *cohorts* of youths in a service at a particular time). JCJC pulled an extensive set of background and outcomes variables from JCMS for the youth in the EPISCenter data whom they could match (n=2496; 93%) in their files; identifiers for 177 youth in the

EPISCenter file could not be matched in JCMS. JCJC then provided de-identified court record information from JCMS on the identified sample to the study validation team at Pitt March 2019 (see Appendix B for a list of data received). The validation team then merged the court record information from JCJC with the service-level data provided by the EPISCenter, using a newly constructed identification number generated by JCJC. The validation study team did not receive identifying information about the youths composing the sample.

Creating a single, usable data file reflecting SPEP™ rating data and the appropriate youth background characteristics and outcomes (i.e., accurately combining the information provided by JCJC and the EPISCenter) was an involved and laborious data management task. We share this information regarding the difficulties of (and time required for) the data management tasks to alert future researchers and funders to this reality for planning purposes. It is not a veiled criticism of JCJC or the EPISCenter; all parties taking part in this validation underestimated the complexity of the work that would be required in the data cleaning, reorganization, and consolidation phases of this project.

Complications and delays encountered resulted largely from the inherent problems of constructing usable research data sets from data that were not collected for research purposes. To illustrate, the information provided to the validation team by JCJC was contained in 21 password-protected excel worksheets and there was variability in the structure of these data worksheets (e.g., multiple rows per unique youth matching the number of different residential placement episodes, multiple rows for a unique youth to capture distinct charges on a referral). Thus, it was necessary to complete multiple data management tasks along the way toward the creation of a single analysis file. In summary form, these steps included:

- generating summary scores for essential constructs (e.g. number of prior placements, number of and days to recidivism events)
- exporting the data into SPSS format
- restructuring the SPSS files to reflect one row per unique youth/service start date combination
- merging the various background and outcomes data files with the SPEP™ data into a single file for analysis.

This final step of merging the EPISCenter and JCJC data sets required multiple meetings with EPISCenter staff to understand the nuances of the data provided and to discuss aspects of data cleaning (e.g., creating consistency in organization/program/service names used by SPEP™ raters over time and location). All of these efforts have provided valuable "lessons learned," many of which have now been incorporated (or are being incorporated) into the ongoing SPEP™ processes for data collection used by the EPISCenter.

II.B. Caveats about the data sets

The end-product of these efforts is a highly unusual resource in juvenile justice research. The consolidated data sets provide information about both dimensions describing numerous services and the characteristics and recidivism of youths who received each particular service (by creating data organized by each unique youth/service start date combination). These data,

however, are not flawless. Completion of the tasks discussed above revealed several limitations of the existing data sets that had implications for later analytic approaches and interpretations of findings. These include the following:

- *YLS/CMI Assessment Date*. The data provided regarding the YLS/CMI scores assigned to a particular youth does not necessarily reflect the date the assessment was conducted. It instead reflects the date that the probation staff entered the score into the JCMS system. There are currently no dates available in the JCJC data bases that capture the actual date the YLS assessment was conducted. The EPISCenter identified and provided the evaluation team with the YLS scores closest to the date of the beginning of a service. In certain analyses, the validation team made the assumption that this score reflects the risk of recidivism observed for each youth at the start of the service with a SPEP™ score. The validity of this assumption could be questioned. In addition, though, it is assumed that this score would also have been factored into the SPEP™ subscore reflecting the proportion of high-risk youth enrolled in the service.

- *YLS/CMI Scores.* The YLS/CMI data presented challenging missing data issues as well as some concerns about the inconsistencies between the EPISCenter data file and data provided by JCJC. YLS *total score* is present in the EPISCenter data for 78% of the unique youth/service start data combinations (3,945 of 5,065 such combinations), and YLS/CMI *risk category* is present for 83% of these combinations (4,225 of 5,065). Descriptive statistics indicate little variability in the risk category variable (most youths are medium level risk), so we used total YLS/CMI score for descriptive purposes. However, despite several (unsuccessful) efforts to recover missing information by merging JCJC and EPISCenter data, the amount of missing data ultimately prevented us from using the YLS/CMI scores for other analyses because its use would have substantially reduced, and possibly biased, the available sample.

- *Services embedded within programs and organizations:* The SPEP™ makes a distinction between programs and services and the formats or organizational framework within which services are delivered. In other words, the "unit" to which a SPEP™ score is applied is the service, but services occur within programs and programs are delivered within an organization. In statistical parlance, this translates to "nesting issues" and means that analytic approaches must account for this aspect of the data. In many situations like this, multi-level hierarchical models are used to account for the nonrandom assignment of observations to the next level up (e.g., youths with certain characteristics are likely to be sent to particular types of programs). In this situation, however, these models were not seen as appropriate, given the nature of the outcome variables. Adjustments for the risk of cohort members are used instead and explained in the following sections on data analysis.

- *Cohort size.* There are 162 distinct cohorts of youths connected with corresponding services in the data set. On average, there are 31 youths per cohort, but the number of youths in a cohort can vary widely (they range from 4 to 146 youths in a cohort). Cohorts with large numbers of youths provide more stable estimates of values for statistical analyses; cohorts with smaller numbers provide less stable estimates. In line with the guidelines provided in the SPEP™ materials and to increase analytic accuracy, we only use cohorts with ten or more youths in analyses.

- *Service types:* Some service types (e.g., mediation) have a small number of associated cohorts. If the number of cohorts for a particular service are too small, it is difficult to get an accurate assessment of an effect for that service type. If any estimate of an effect is based on a small number of cohorts observed, the estimate will be unreliable. As a result, we did not test services with less than ten cohorts connected to its category. This meant that we tested the effects for seven out of the thirteen possible primary service types (specific services tested are indicated in Table 2).

- *Youth in multiple cohorts.* Some individual youths receive more than one service that was assigned a SPEP™ score. As a result, these youths are members of more than one cohort. This creates an analytic issue, since any differences observed between services in recidivism will be partially affected by the characteristics of the youths who appear in the cohorts used to test the outcomes of that service involvement. The issue of repeated appearances of youth in different cohorts is not trivial; 2,496 unique youths are represented in 5065 rows of data. On average, youths appeared in two (s.d. = 1.65) of the SPEP™ service cohorts present in the data, but most of the youth (61%) appeared in only one cohort. Three youths appeared in eleven service cohorts (the maximum observed). This issue is addressed in the analyses when it is relevant and the methods for doing so are presented in the results of these analyses (see Section III)

- *Missing data*. Some variables of potential interest have a sizable proportion of their values missing. For example, the county in which the youth resides is missing in 30.7% of the unique cases. Since different variables are appropriate for particular analyses and suitable sample sizes will vary depending on the analysis conducted, decisions about which variables to include in each analysis were made on an individualized basis, depending on the purpose of the analysis. There was no blanket rule used for exclusion of a variable.

These data limitations are important to keep in mind in assessing the findings presented. Some of the listed limitations are simply related to how a particular variable of interest is defined and these issues can be addressed when interpreting specific results. Other aspects of the data go beyond definitional limitations, however, requiring multiple and varied analyses to determine the impact of the limiting issue; we did not do extensive sensitivity

analyses of this type. The strategies taken to address particular aspects of the data will be presented in the relevant analysis sections below.

II.C. Variable Definitions

II.C.1. Recidivism

Recidivism is the primary outcome of interest. For the purposes of this validation study, we define a recidivating event as either an adjudication or conviction for a misdemeanor or felony offense. This definition does not reflect arrests, but instead only includes incidents which reach the point of court processing, either as a juvenile or adult. In addition, this definition does not include summary offenses.

We calculate this figure for four time periods after the SPEP™ service end date: 6-months (1-180 days), 12-months (1-365 days), 18-months (1-545 days) and 24-months (1-730 days). We chose the definition of recidivism used here to be identical to the definition used in prior reports generated by JCJC. It is important to note, however, that the meta-analyses which underpin the SPEP™ approach used 12-month rearrests as their outcome measure (Lipsey, 2009). Since not all arrests lead to an adjudication/conviction, the use of court appearances rather than arrests in this validation necessarily produces a lower rate of recidivism than the SPEP™ background research and prior validation studies.

The calculation of each follow-up period started when the individual's involvement with the service with the SPEP™ score ended. Obviously, the date of service involvement varied across individuals in the sample, e.g., one individual might have ended involvement with Service A on 01/12/2011 and another individual might have ended involvement with Service B on 2/28/18. The recidivism data, however, was pulled on the same day for all individuals in the sample, i.e., 12/6/18. This means that the youth in the first example given above would have over two years within which recidivism might be observed (the "recidivism window") while the second youth would only have a 281-day period for which recidivism might be observed. As a result, it makes sense to calculate recidivism figures for each case based on the time available for observation. In operation, that would mean that the first example above would be able to provide a value for each of the four recidivism time periods, while the second example would only be able to provide a value for the six-month period.

A "recidivism window" (representing the number of days between the service end date and the date recidivism record information was pulled) was calculated for each youth. A youth would need to have a recidivism window of at least 180 days for a valid indication of recidivism status at 6 months, a recidivism window of at least 365 days for a valid indication of recidivism status at 12 months, and so on through two years. If the recidivism window did not meet the threshold for the time period indicator and the youth had not recidivated, the case was considered "ineligible" to provide data for the specific time period in question.

However, if the recidivism window was less than the time period threshold and the youth had an adjudication or conviction sometime within the time period observed, they would be coded as having recidivated during that time period. The logic of this coding is that, even though the person had not been observed during the entire time period (e.g., for only 22 of the 24-month time period), they had already demonstrated that they would be a recidivist even if the entire time period until the end of that recidivism window had been observed. The non-recidivists who were not observed for the entire time period, however, present a different scenario. It is not logical to code these cases as non-recidivists since it is not certain that they would not recidivate in the unobserved remainder of the recidivism window. To extend the example above, we do not know who of the observed non-recidivists at 22-months would have recidivated in month 23 or 24.

We can see how this logic plays out in the subsequent tallies of recidivism. Take for example, a youth who is only observed for 120 days and does not recidivate during that time period. This youth would ineligible to provide any data for the 6-, 12-, 18- and 24-month recidivism indicators. If, however, the youth recidivated on day 110 of the observed 120 days, this youth would provide a case of positive recidivism for the 6-, 12-, 18- and 24-month recidivism indicators. Conversely, if a youth had an observation period of 340 days and did not recidivate during that time, he/she would be a non-recidivist for the 6-month recidivism indicator and ineligible for the 12-, 18- and 24-month recidivism indicators.

This approach obviously inflates the recidivism rate for a given period by an unknown amount by counting as recidivists some individuals who were not observed for the entire time period. This method of counting recidivism was chosen, however, because there are no other unbiased or arbitrary methods for calculating this figure. Moreover, this method is the one used by JCJC in calculating the figures presented in their reports. Thus, the results reported here are comparable to those figures.

II.C.2. SPEP™ Scores

Information related to the SPEP™ Total Score and the SPEP™ scores for the components of the rating were provided by the EPISCenter and were not changed or recoded by the validation study team. The components used in the analyses are defined in the Standardized Program Evaluation Protocol (SPEP™): A Users Guide (Lipsey & Chapman, January 2017). Below is an overview of the SPEP™ scoring process as well as a list of the SPEP™ core components and the definition provided in the User's Guide.

Overview of SPEP™ scoring. The SPEP™ is configured so that the maximum overall score is 100, with 100 points representing what research has shown to be most effective in reducing the recidivism of juvenile offenders. Ratings on the individual SPEP™ components each have a maximum value assigned in proportion to the strength of that factor for predicting recidivism effects in the statistical models used in Dr. Lipsey's meta-analysis. To generate a total SPEP™ score for a particular service, the data collected for each of the SPEP™ components are reviewed and used to assign points for that component per the SPEP™ scoring procedures; the

points are then summed across all the components to produce the Total Score (User's guide, page 18).

SPEP™ Components

Service type: For the *Primary Service Type*, the points assigned to a service are proportionate to the average overall magnitude of recidivism effects found in the research for that service type. The program service types covered in the meta-analysis database have been assigned to one of five groups according to their average level of effectiveness for reducing recidivism. Once a service is matched appropriately with a SPEP™ service category, the appropriate point rating for that service can be found in the scoring spreadsheets and supplementary documentation provided to authorized SPEP™ users (User's guide, page 19). A particular service can receive a maximum of 30 points for this component.

For the *Supplementary Service*, 5 additional points are added to the service type score if the program includes a secondary service that is frequently associated with the primary service and has been shown to be effective in reducing recidivism when delivered in conjunction with the primary service. For services that do not have any eligible supplementary services, 5 points are also added to the service type score so as not to penalize primary services that do not have sufficient research to identify relevant supplementary services. The most current identification of eligible supplementary services for each primary service is provided in the scoring spreadsheets and related documentation made available to authorized SPEP™ users (User's guide, page 19).

Service Quality points: The quality of service delivery section of the SPEP™ requires an overall rating of how well the provider organization supports and monitors the quality with which the services being assessed by the SPEP™ are delivered. The quality of service delivery score for the respective provider is derived from ratings on each of four elements of this SPEP™ component: (1) a written service protocol, (2) credentials and training for that service, (3) procedures for monitoring adherence to the protocol, and (4) procedures for taking corrective action when there are unwarranted departures from the protocol. Ratings on each of these elements are combined into a total score that is further divided into a low, medium, and high range and used to determine the overall point value given for the quality of service delivery component of the SPEP™ (User's guide, page 19). Quality points are either 5 points (low quality), 10 points (medium quality) or 20 points (high quality).

Amount of Service (Duration and Dosage points): Amount of Service elements integrate the amount of service provided and the optimal amount for a service of that type. Target values for each type of service are determined from Dr. Lipsey's meta-analysis. These target values represent the median amount of service provided by the services found to be effective in the meta-analysis. The dosage data collected for the weeks of service duration and total contact hours for the eligible juveniles in the selected cohort are used to determine the percentage of juveniles who received the target amounts of service. Those percentages, in turn, determine the points awarded to the service for this SPEP™ component. The target values for service

duration and contact hours are different for each service type and are periodically updated on the basis of new research added to the meta-analysis database. The most current values will be provided to authorized SPEP™ users (User's guide, page 20). Dosage and duration points are separately reflected as:

0 = 0% target hours of service
2 = 20% target hours of service
4 = 40% target hours of service
6 = 60% target hours of service
8 = 80% target hours of service
10 = 99% target hours of service

Service Risk points: For the final SPEP™ component, Risk Level of Youth Served, target values for classifying juveniles into risk categories that are appropriate to the risk assessment instrument used to determine risk are used (User's guide, page 20). In Pennsylvania, this variable represents the percentage of youth in the cohort that are medium or high risk as determined by the Youth Level of Service (YLS).

Total SPEP™ Score (Total raw Service points earned): The above points for the above components are added together to generate the Total SPEP™ Score.

Program Optimization Percentage (POP score): For some purposes, the SPEP™ Total Score may not be the most informative score for assessing the performance of individual providers and services. For example, a particular service may be the most appropriate one for addressing the needs of a certain segment of juvenile offenders, but the type of service it provides might not be classified in a service group that receives the highest score for type of service in the SPEP™ scheme. As such, its SPEP™ Total Score can never reach the maximum of 100 points. Similarly, a program or service may be specifically tasked with serving low or moderate risk offenders. In that circumstance, again, it is not possible for that service to reach the maximum 100 points of the SPEP™ Total Score. If the respective role for the service provider is clearly understood to be the provision of a lower rated service or a service to lower risk juveniles, it is possibly misleading to report only the SPEP™ Total Score as an assessment of that service.

For situations such as these, there is another SPEP™ score that reflects the performance of the service relative to what is expected of it—the Program Optimization Percentage or POP Score. To compute the POP Score, the points across each of the SPEP™ components are summed, just as for the Total Score. Instead of simply presenting that total, however, that number of points is divided by the maximum number of points possible given the service type and/or risk level appropriate to that type of service. The result is a percentage value that indicates how well the program scores in relation to the maximum number of points possible for a program service in that role; the POP score essentially standardizes the SPEP™ Total Score to account for the type of service provided (see the SPEP™ User's Guide, page 22).

II.C.3. Youth Risk Level (from YLS/CMI)

In stage 2 of the JJSES, Pennsylvania embraced the regular and consistent use of the Youth Level of Service/Case Management Inventory (YLS/CMI ™) as the primary measure of youth risk for recidivism. The YLS/CMI ™ is a validated and widely used structured risk/needs assessment which assesses risk for recidivism by measuring 42 risk/need factors over the following eight domains: prior and current offenses, family circumstances/parenting, education/employment, peer relations, substance abuse, leisure/recreation, personality/behavior, and attitudes/orientation. Both the overall score ("total risk score") and the domain scores from the YLS/CMI ™ administered nearest to the SPEP™ service start date were collected for use in the validation study (see note on page 9 about challenges encountered with the YLS/CMI ™ data). Per the YLS/CMI ™ manual, the total risk score can range in value from 0-42. This score is translated to a risk category using the following cut-offs: Low risk: 0-8; Moderate risk: 9-22; High risk: 23-34; Very high risk: 35-42.

III. SAMPLE DESCRIPTION

III.A. Data Structure

Before we begin describing the sample, it is important to explain how the EPISCenter data regarding the services are structured. The primary unit of analysis of the EPISCenter data is the service that has been given SPEP™ service scores. Each service has a start and end date indicating the time period of that service that produced the SPEP™ ratings. Each service that has received a SPEP™ score, however, also has a linked group of individual youths who received the service of interest during the time period reflected by the SPEP™ score (i.e., a "cohort" of youth connected to that service). The recidivism outcomes for this cohort of youth provides an estimate of that service's impact on the youths' subsequent criminal offending.

Each youth in a cohort can have different start and end dates for exposure to the rated service. One youth might be in an initial group that goes through the service and another youth might be in the next group getting that service. These youth would both be in the same cohort assigned to that service (i.e., they received that same service with a given SPEP™ score), but they would have different start and end dates of their involvement. Also, a single youth may be in more than one service that has a SPEP™ rating, and those services can have different start/end dates.

Each row of data in the EPISCenter data file (n=5,065 rows) reflects a combination of a unique youth and the start date of a rated service. There are 2,496 unique youths producing the cohorts for the services with SPEP™ ratings (see section below for a description of these youths). As noted earlier, 39% of the youth appear in more than one cohort.

It may be apparent by now that the reality of individual youth being nested in particular services creates considerable problems for data analysis. Youths are not randomly assigned to these different services, and any estimation of the "average" effect of a service has to take the

characteristics of the cohort of youth receiving it into account. This becomes more complicated when we consider that an individual youth may be contributing to the estimates of only one or several services, and at least part of the estimate obtained of the "average" effects of the service will be affected by the characteristics of that youth (e.g., a cohort with a large proportion of high-risk adolescents could be expected to generate a relatively higher average recidivism rate). These structural features of the data and their implications will be explored in more depth when we present the analytic approaches later in this document. For our purposes now, however, it is simply important to remember that there are several lenses that can be used to describe these data and that it is important to note which lens is being used in any particular description or analysis.

III.B. Services and outcomes

There are two relevant samples to consider in the following analyses. First, there is the sample of unique youths who are the recipients of the SPEP™ services; these individuals form the cohorts for each assessed service. Second, there are the cohorts of youths who received a particular type of service with a SPEP™ assessment. The tables below show characteristics of the sample from each "viewpoint."

III.B.1. Youths in the cohorts

There are 2,496 unique youth who constitute the cohorts connected to the services with SPEP™ ratings data drawn from the EPISCenter data base. Table 1 below shows the basic demographic characteristics of these youths.

Table 1. Characteristics of the individual youths (n=2,496) constituting the cohorts

Characteristic		**Average/Prevalence** Mean (s.d.) or Number (%)
Age at SPEP™ service start date		16.19 (s.d. 1.4)
Gender	Male	2076 (83.2)
	Female	420 (16.8)
Race/Ethnicity	White	864 (34.6)
	Black	1113 (44.6)
	Hispanic	423 (16.9)
	Other	96 (3.8)
Age at first referral	(n = 421 missing)	13.89 (s.d.=1.76)
Number of Prior Written Allegations (referrals to a probation department)		3.31 (s.d. = 2.50)
Prior removal from the home	(n = 421 missing)	1651 (79.6)
SVC Indicators	Serious	190 (7.6)
	Violent	109 (4.4)
	Chronic	244 (9.8)

	S, V or C	301 (12.1)
Child Offender	(n = 421 missing)	234 (11.1)

As shown in Table 1, the youth receiving the services in this sample are primarily minority (65%) males (83%) who were about 16 years old on average when they began their involvement with a service (mean age = 16.19 years old; s.d. = 1.4). Age was not able to be calculated for 142 youth who were missing their SPEP™ service start date. Although predominantly urban youths, this sample has some representation from almost all counties in Pennsylvania. About half (49%) of the youths are from the juvenile courts in Allegheny, Philadelphia, Lehigh or Berks counties.

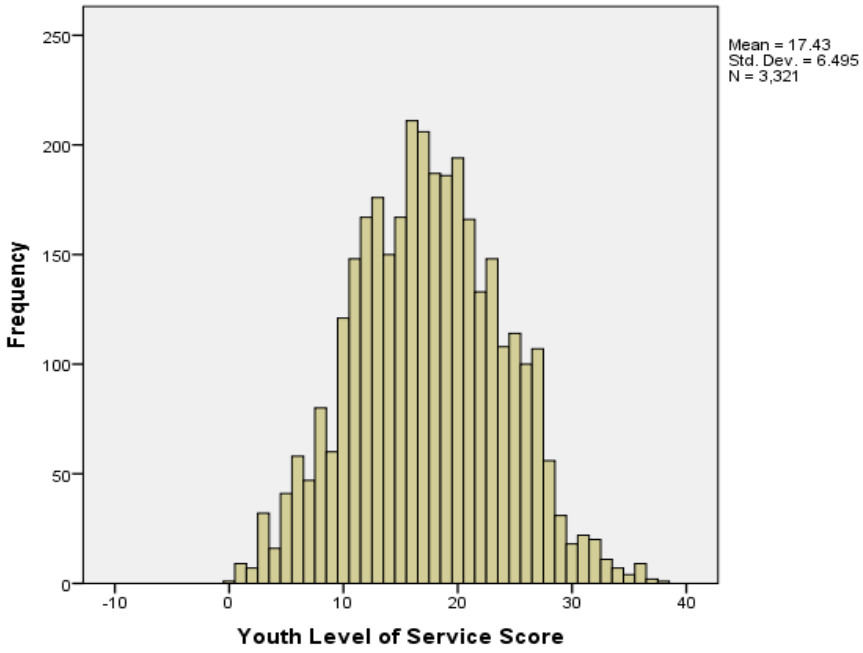
III.B.2. ***YLS/CMI Risk scores and levels***

The YLS/CMI scores and risk level designations of the sample are of most interest if they indicate the values of these variables at the time of youth's entry into a service. Obtaining this view required a bit of data manipulation. As noted earlier, some of the 2,496 unique youth in the sample described above are in more than one cohort; meaning they received more than one of the services with a SPEP™ rating. Since services can start and end on different dates and the risk level of an individual youth may change over time, a single youth might have more than one total risk score, representing different risk levels at the different times of enrollment in a service.

To capture the range of risk scores at service enrollment, therefore, the sample examined is the number of unique youth/service start data combinations (n=4,165). When missing YLS/CMI data is taken into account (see page 9), the sample of unique youth/start date combinations is reduced considerably. The YLS/CMI total score is available for 3,321 unique youth/service start date combinations ("cases"). This amount of missing information would have seriously restricted the analyses and we present the scores here for descriptive purposes only. The analyses presented later use an "expected recidivism rate" derived from other youth background characteristics that was developed in response to the level of missing YLS/CMI scores (see page 27).

The YLS/CMI total mean score for this sample is 17.43 (s.d. = 6.50; range 0-38). Figure 2 displays the distribution of YLS total scores in these combinations. These scores are impressively normally distributed, meaning that the data hovers around the mean score with few cases falling to the extreme right or left.

Figure 2: Distribution of YLS/CMI™ Total Scores (unique youth/service start date combinations)



YLS/CMI risk category is available for a slightly larger sample of unique case/start date combinations (n= 3,457). Of these, 9% are classified as low risk, 68% as moderate risk, 23% as high risk, and less than 1% as very high risk.

III.B.3. *Service Cohorts*

Youths who were engaged in a particular service during the time period covered by the SPEP™ ratings were considered for inclusion into the "cohort" of youths who might be followed regarding their subsequent recidivism connected with that service involvement. There are 162 distinct cohorts included in the data with an average of 31 youths in a cohort (s.d. =29.3). The number of youths in a cohort ranged from 4-146. There is variability among cohorts in the characteristics of the youths participating in the service at the time of the SPEP™ rating. A table of youth characteristics in each cohort is provided in Appendix C.

III.B.4. Service Types

The SPEP™ has a well-developed protocol for identifying and classifying services, and these procedures were followed in the SPEP™ effort in Pennsylvania. Only services that are identified as therapeutic (those oriented mainly toward facilitating constructive internalized and sustained changes in behavior) were eligible to engage in the SPEP™ process. Qualifying services were then categorized into defined service types. The dimensions of program operations used to categorize a service type are as follows:

- *Therapeutic Category (definitions from the SPEP™ User's Guide manual, pg 8)*
    - *Restorative services*. Services that aim to repair the harm done by the juvenile's delinquent behavior by requiring some compensation to victims, reparations via community service, or reconciliation between victims and offenders
    - *Counseling and its variants*. Services characterized by a personal relationship between the offender and a responsible adult who attempts to exercise influence on the juvenile's feelings, cognitions, and behavior; family members or peers may also be involved
    - *Skill building services*. Services that provide instruction, practice, incentives, and other such activities and inducements aimed at developing skills that will help the juvenile control his/her behavior and/or enhance the ability to participate in normative prosocial functions
- R*esearch base*: distinguishing evidence-based services from those which are locally developed
- *Setting*: services delivered in the community versus those delivered in a residential setting

These broad designations of service type provide a method for categorizing services into validated groups for analyses of particular policy interest. For example, it is of clear interest to funders and policy makers if community-based services have a better overall outcome than residential care in terms of recidivism (once controlling for the risk of recidivism of the samples considered). These service types provide a characterization of a service in terms of its general therapeutic orientation, its evidence-based nature, and where the service is delivered.

III.B.5. Primary Service Groups

As mentioned above, services are also assessed for their expected "dosage" (e.g., number of sessions to be attended) and "duration" (e.g., the weeks spent in the service). Different types of services logically have different expected target amounts of duration and dosage. The User's Guide notes that target amounts for each service are meaningful only in the context of the full set of SPEP™ ratings. That is, the expected effects of the given amounts of service for any service category depend on the quality of service delivery and the risk level of the youth served as defined in the SPEP™. There are five primary service group types that each have a different assigned number of points contributing to their overall SPEP™ score, depending on whether the youths in the service meet the target levels of involvement for that service. The different service types and their target service involvements are:

Group 5 services (Score=30)

Cognitive-behavioral therapy (Target weeks=15; target hours=45)

Group 4 services (Score=25)

Group counseling (Target weeks=24; target hours=40)

*Mentoring (*Target weeks=26; target hours=78)

Behavioral contracting; contingency management (Target weeks=24; Target hours=72)

Group 3 service (Score=15)

Family counseling (Target weeks=20; target hours=30)
Family crisis counseling (Target weeks=4; target hours=8)
Mixed counseling (Target weeks=25; target hours=25)
Social skills training (Target weeks=16; target hours=24)
Challenge programs (Target weeks=4; target hours=60)
Mediation (Target weeks=4; target hours=8)

Group 2 services (Score=10)

Restitution; community service (Target weeks=12; target hours=60)
Remedial academic program (Target weeks=26; target hours=100)

Group 1 service (Score=5)

Individual counseling (Target weeks=25; target hours=30)
Job-related training; Vocational counseling (Target weeks=20; target hours=40)
Job training (Target weeks= 25; target hours=400)
Work experience (Target weeks=26; target hours=520)

As mentioned above, these categories are used primarily to determine the correct reference score for scaling the SPEP™ ratings for the amount (dosage) and length of time (duration) of the services provided to the youths taking part in that service. As such, the five-category breakdown does not provide a conceptually useful way of clustering services. As a result, the five category assignments are not used in any of the presented analyses. They are presented here to provide background for consideration of results related to analyses that use the SPEP™-POP score as a metric for comparison or those that examine the effects of ratings of duration and dosage in terms of recidivism.

III.B.6. Primary Service Type

A more useful, and conceptually coherent, method for differentiating types of services can be constructed from the distinctions subsumed in the five-group categorization system presented above. The Primary Service Type differentiates the five service groups outlined above into thirteen types of services. This set of categories allows for the service provided to be characterized as one of thirteen possible types of service, i.e., individual counseling, job-related training, remedial academic program, restitution/community service, challenge program, family counseling, mediation, mixed counseling, social skills training, behavioral contracting/contingency management, group counseling, mentoring, or cognitive-behavioral therapy. These descriptors are applied to the primary service provided and comparisons of the impact of services can be made between and among these types of services in later analyses.

Table 2 provides a summary of the number of youths represented in the different service type categories presented above. For this purpose, a "case" indicates a match of a service with a SPEP™ score and an individual youth in a cohort connected to that rated service. This means that the same youth may contribute to multiple "cases" counted below, since a youth may be in more than one cohort.

Table 2: Frequency of cases by Service Type categories (n=5,057 cases; 8 are unclassified)

Service Type Category	Number of cases	(%) of sample of individual level observations
Therapeutic Category		
Counseling	1890	37.4
Restorative	350	6.9
Skill-building	2817	55.7
Primary Service Group Type		
Group 1	739	14.6
Group 2	748	14.8
Group 3	1146	22.7
Group 4	1121	22.2
Group 5	1303	25.8
Primary Service Type		
Group 1: Individual Counseling	484	9.6
Group 1: Job Related Training	255	5.0
Group 2: Remedial Academic Program	417	8.2
Group 2: Restitution; Community Service*	331	6.5
Group 3: Challenge Programs*	159	3.1
Group 3: Family Counseling	372	7.5
Group 3: Mediation*	19	.4
Group 3: Mixed Counseling*	60	1.2
Group 3: Social Skills Training	536	10.6
Group 4: Behavioral Contracting; Contingency Management*	147	2.9
Group 4: Group Counseling	741	14.7
Group 4: Mentoring*	233	4.6
Group 5: Cognitive-behavioral Therapy	1303	25.8
Research base		
Evidence-based	3852	76.2

Locally Developed	1205	23.8
Setting		
Community setting	1544	30.5
Residential setting	3513	69.5

**There are fewer than 10* SPEP™ *cohorts for these service types and they are, therefore, not included in certain analyses.*

III.B.7. Length of Service

Different types of service can require different periods of involvement, and different youths can participate in services for different lengths of time (e.g., a youth is moved to a new facility mid-way through the program). Youths could therefore have different lengths of service involvement that are not totally dependent on the type of service. The length of service for each case (SPEP™ service and unique youth combination) was calculated using the service start and end dates provided.

Based on this calculation for the available sample, length of service involvement appears to vary considerably. Youths in the sample were in a service for an average of 128 days (s.d. = 111). As might be expected, services provided in a residential setting were statistically significantly longer than services provided in the community although not by a great amount of time (residential service average = 132 days, s.d.=116; community service average = 117 days, s.d. = 96; $t = 4.05(4,852)$, $p<.001$). Locally-developed programs have significantly longer lengths of service than evidence-based programs (locally developed = 139 days, s.d. 120; EBP = 91 days, s.d. = 67; $t = 13.13$ (4,852), $p<.001$).

III.B.8. SPEP™ rating scales

As described earlier (pages 13-15), ratings of six program components are summed to determine the total SPEP™ score. Each of these component scores has a minimum and maximum score that is provided in the User's Guide. Table 3 shows the descriptive statistics for the component and total scores for the entire sample (n=5,057 cases, 8 unclassified). There were some missing values that affected the figures presented in this table. The necessary values were available for 4,854 cases; 211 cases are considered missing for this calculation. Specifically, service start date was missing for 142 cases, the recorded service start date was prior to the service end date for 44 cases, and the same date was recorded as the start and end date for 25 cases.

Table 3: Descriptive information for SPEP™ Scores and components scores

SPEP™ Components and Scores	Min	Max	Mean (SD)
Primary Service Type Points	5	30	18.87 (9.07)
Supplemental Service Type Points	0	5	4.42 (1.60)
Service Quality Points	5	20	14.80 (5.77)
Service Duration Points	0	10	4.08 (3.31)
Service Dosage Points	0	10	3.49 (3.49)
Service Risk Points	2	25	16.22 (5.20)
Total SPEP Score (Total raw Service points earned)	23	100	61.88 (15.48)
SPEP Score Percentage point or POP Score (Total raw points divided by max possible points)	.310	1.000	.69 (.137)

Since SPEP™ scores are assigned to a service, all youth in the service cohort have the same values for the SPEP™ components and total score (since they are all "nested" within a specific service for which the SPEP™ score was assigned). Therefore, the most appropriate way to view the distribution of SPEP™ scores is to do so across the 162 cohorts. The number of individuals in a particular cohort, the service type of that cohort, and the component and total SPEP™ scores for that service are all presented in Appendix D.

III.C. Recidivism

The definition for observed recidivism and the rules for determining recidivism rates at different follow-up points are presented earlier in the report (see page 11). Here we provide some descriptive information about the findings regarding this outcome indicator.

III.C.1. Observed recidivism rates at each follow-up window

We first examined the length of the recidivism window for each unique youth/service start date combination ("case"). Table 4 shows the number and percentage of cases that reached each of the demarcated periods of follow-up for recidivism.

Table 4: Number (%) of eligible cases reaching each recidivism follow-up period

Recidivism Window	Number (% of 5065)
6-month recidivism (cases with at least 180 days in the recidivism window)	5065 (100%)
12-month recidivism (cases with at least 365 days in the recidivism window)	5025 (99%)
18-month recidivism (cases with at least 545 days in the recidivism window)	4666 (92%)
24-month recidivism (cases with at least 730 days in the recidivism window)	3872 (76%)

As shown in Table 4, a substantial number of youths are available at all four recidivism windows, but nearly 10% did not reach the 18-month recidivism window and a quarter of cases did not reach the 730-day (2-year) mark. We should note, however, that some of the cases that did not have a recidivism window of 730 days are still counted as a "recidivist" for the 24-month indicator if they had a recidivism event at any point prior (see earlier explanation on page 11). More specifically, while 1,193 cases did not have a period of 730 days after the end of their service involvement, 284 of those cases nevertheless have a positive value for the 24-month (730 days) recidivism marker because they had a recidivism event that fell somewhere during the time that they were examined for a recidivism event. Table 5 below provides a summary of the recidivism rates at each follow-up period for all of the cases in the sample.

Table 5: Recidivism Rates for each Time Period Indicator

Recidivism Time Period Indicator	Overall (adjudication or conviction)	Adjudication	Conviction
6 month (recidivated between day 1-180)	9% (476/5065)	6% (314/5065)	3% (164/5065)
12 month (recidivated between day 1 - 365)	23% (1143/5030; 35 ineligible)	15% (733/5030)	9% (436/5030)
18 month (recidivated between day 1 - 545)	33% (1550/4728; 337 ineligible)	21% (971/4728)	15% (712/4728)
24 month (recidivated between day 1 - 730)	43% (1786/4156; 909 ineligible)	28% (1182/4156)	24% (1006/4156)

We chose to focus on six and 12-month recidivism outcomes for the analysis reported below. This strategy was taken for two reasons: 1) the possibility for bias in results when a sizable percentage of cases are excluded and 2) the likelihood that the positive effects of the service would deteriorate substantially as intervening events occur in the lives of the youth. Expecting a service of the type examined here to have a detectable effect on recidivism eighteen months of two years after program involvement is a high bar to meet. Although analyses of impact from an institutional stay often use a two-year recidivism measure as a logical outcome (Harris, Lockwood, & Mengers, 2009), a one-year follow-up period is also used in many studies. The range of intensity provided by the services assessed here was broad, making the use of a shorter recidivism window seem to be a fairer standard for comparison.

It is also important to comment on the low observed recidivism rate we find at six months (9%) and, to a lesser degree, twelve months (23%). These base rates of recidivism are lower than those seen in much of recidivism research. While possibly an impressive metric of the performance of Pennsylvania services, it might also be an indicator of the stringency of the definition of recidivism applied. In either case, a low base rate for the occurrence of an outcome presents a formidable challenge for research because it makes it very difficult to detect treatment effects. Even a variable with a high level of accuracy in identifying the occurrence of an outcome may not be strong enough to overcome a low base rate.

This is because the ability of a variable to perform above chance (and thus be statistically significant) must be better than the demonstrated performance of predicting the non-occurrence of the outcome; in this instance, about 90% at the six-month indicator. In other words, to obtain statistical significance, the effect of any variable must exceed the accuracy of just predicting that recidivism would not occur (true in 90% of the individual cases at 6 months and about 75% of the cases at the 12-month follow-up). These relatively low base rates cannot be ignored as a potential factor explaining some analyses that do not show significant effects on recidivism rates. Moreover, with a recidivism rate this low, there's little even the most effective programs can do to push it lower. A service lowering the recidivism rate from approximately 10% to 5% would be achieving a different, and arguably much more difficult, objective than a program lowering a recidivism rate from 50% to 45%. Conversely, analyses that do find an effect for the SPEP™ are remarkable since the effect must be large in order to emerge.

It is important to remember, however, that the analyses of most interest will be conducted by examining the recidivism rates at the level of the cohorts. The questions central to this investigation have to do with how much the cohorts who receive certain types of services compare, in terms of recidivism, to cohorts who receive other types of services. The "cohort level" observed recidivism used to make these comparisons is the rate of recidivism for cohorts of adolescents who receive the same type of service. This is simply the prevalence rate of recidivism for the cohort; i.e., the percent of the cohort who recidivated by the specified follow-up period. These values may or may not mirror the patterns of the individual level recidivism figures, depending on which adolescents are members of the cohorts affiliated with the services examined.

When examined by cohort, we find that the average cohort-specific observed recidivism rate is 9% (s.d. .08) at six months and 22% (s.d. .12) at 12-months. Figures 3 and 4 shows the distribution of cohort-specific recidivism rates. A table with the specific recidivism rate for each of the 162 cohorts can be found in Appendix C. The calculation and application of these rates will be explained later in this report.

Figure 3. Distribution of cohort-specific observed recidivism rates – 6 months

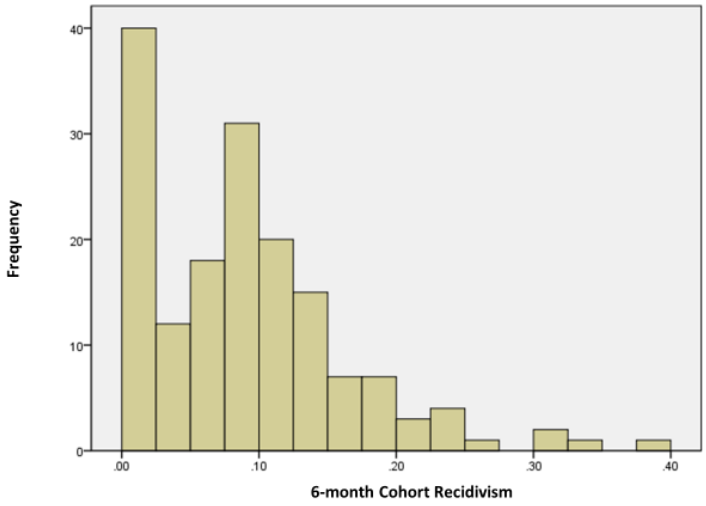
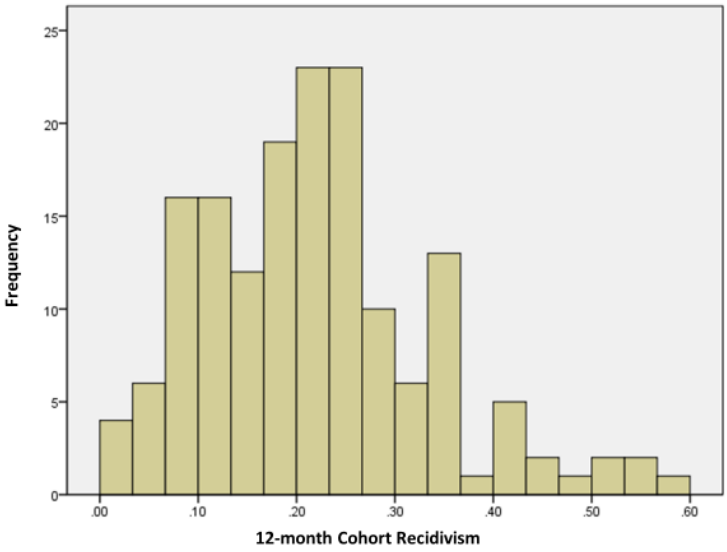


Figure 4: Distribution of cohort-specific observed recidivism rates – 12 months



III.C.2. *Expected Recidivism Risk*

As mentioned previously, there were reservations about the use of the YLS/CMI scores as indicators of the likelihood of reoffending at the time of entry into a service. Specifically, the amount of missing data for the YLS/CMI and concerns about the accuracy of that score for depicting risk of reoffending at the time of service entry made it unwise to use these scores to depict the risk of reoffending for the youths in the sample. As a result, an alternative approach was taken to provide an estimate of a youth's risk for rearrest at the time of entry into a service.

A score was calculated for each adolescent in every cohort to reflect that adolescent's likelihood of recidivating during a particular follow-up period after the involvement with a service (we call this the *expected recidivism risk* for that individual). This score is based on the characteristics of that adolescent at the time of entry into the service being assessed. It is determined by a formula based on a regression equation predicting observed recidivism in this sample at each of the follow up points. If the regression equation has an acceptable overall level of accuracy in identifying the adolescents who do recidivate, it can then be used as a calculation applied to each individual case to assign a score indicating how much a particular adolescent "looks like" a case that will recidivate. It is a value that indicates the inherent "expected recidivism risk" of the adolescent at the time of service entry. A higher score indicates an adolescent with a higher chance of recidivating.

The variables used to develop the regression-based formula for determining the expected likelihood of recidivism were: gender, race/ethnicity, whether the youth had ever been adjudicated prior to the start of the SPEP™ service, age at the start of the SPEP™ service, whether the youth was classified as a child offender, the number of days in court-ordered placement or detention prior to the SPEP™ service start date, count of prior serious offenses, count of prior violent offenses, count of chronic offenses, and whether the youth fit the definition of a serious, violent or chronic offender. It is worth noting that these variables heavily reflect aspects of a youth's prior history; they are not indicative of the range of factors that might be considered relevant to continued offending. The purpose here, though, is not to construct a theoretically sound set of predictors and to compare and interpret their relative explanatory power. The purpose of this analytic task is to simply generate a suitably predictive combination of variables that, when taken together, identify how likely each individual is to re-offend.

In order to use the expected recidivism risk score for each case in the subsequent analyses, we need confidence that this score really discriminates those adolescents who reoffend from those who do not. If a large number of cases with low expected recidivism risk scores actually recidivated or if a large number of individuals with high likelihood scores don't recidivate, then the predictive equation is not working well. We look at a standard metric used

to assess clinical tests in medicine and psychology to see how well this calculated score performs.

We examine the overall *sensitivity* of the scale (how many of the actual recidivists get higher scores) and the *specificity* of the scale (how many of those with higher scores actually recidivate). Essentially in this step we are comparing the "true positive rate" against the "false positive rate" at different scores and judging the overall performance of the scale accordingly. These comparisons of sensitivity and specificity are graphed in what is called a Receiver Operating Characteristic curve or "ROC curve." The ROC curves for this calculation of expected recidivism risk for six-months and twelve-months after the end of service are shown in Figures 5 and 6 below.

Figure 5. Recidivist vs. non-recidivist ROC curve for equation predicting recidivism for the whole sample at six months after service exit

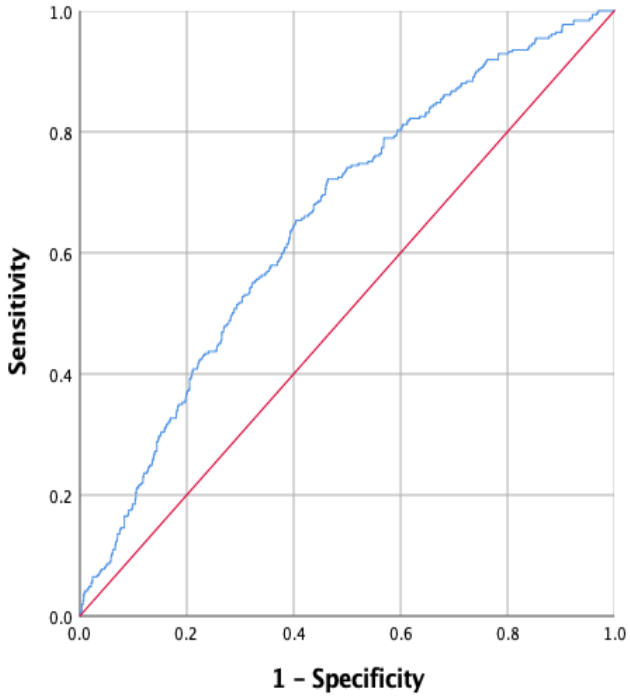
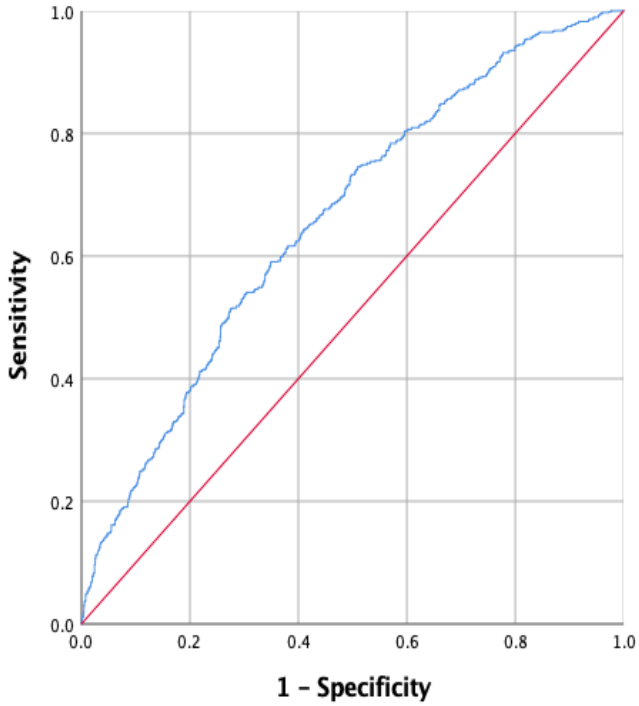


Figure 6. Recidivist vs. non-recidivist ROC curve for equation predicting recidivism for the whole sample at twelve months after service exit



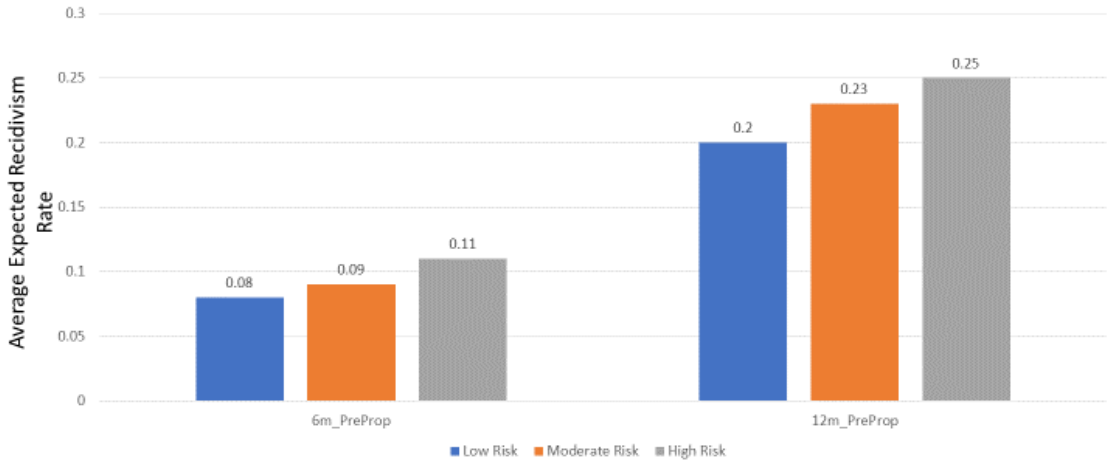
The overall performance of an instrument for differentiating cases effectively is calculated by the Area under the Curve (AUC) on the ROC graph. The diagonal (red) line in the middle of the graph is what would be obtained if the score had no ability to discriminate above pure chance; if it was just like flipping a coin. The blue line in each figure represents the performance of the calculated regression equation. The better the performance of the instrument (in this case, each of the equations with the background characteristics), the more the blue line moves away from the diagonal and toward the upper left corner of the graph. As it moves farther away from what it would do by chance, it then generates a larger area under the curve (AUC) value; a measure of its overall distance from the chance diagonal line. That is, it shows higher sensitivity and specificity; it is better at distinguishing recidivating from non-recidivating individuals.

The area under the curve (AUC) values in the above figures are .65 for six-month recidivism and .66 for twelve-month recidivism. This is about the same level of accuracy generally obtained in actuarial instruments for assessing the likelihood of future violence in individuals with mental illness or likely re-arrest in criminal populations**.** It is also only slightly lower than the level of accuracy obtained with the Youth Level of Service (YLS/CMI) in its development research, and is nearly identical to the AUC found in applications of the YLS/CMI across a range of studies (Schwalbe, 2007). Finally, it is within the range of accuracy of risk of recidivism estimates used in other research done to validate SPEP™ (see, for example, Lipsey, 2008; AUC = .65 for predicting rearrest at 6 & 12 months and Onifade & colleagues, 2008; AUC for predicting a new charge within 12 month = .62). The expected recidivism risk score seems

to be doing an adequate job of providing an estimate of the chances that a particular adolescent in the sample will recidivate in the next six months or one year after service involvement.

It is reasonable to wonder how the predicted probability values generated for this sample correspond to the YLS/CMI, the most widely used risk assessment tool used in PA. As illustrated in Figure 7 below, the expected recidivism rate increases as the YLS/CMI risk level gets higher and the differences between these groups in the calculated expected recidivism rate is significant (ANOVA, 6 months, $F = 16.46(3)$, $p<.001$; 12 months $F = 16.01$ (3), $p<.001$). These patterns combined with the similar AUC values found between our data and extant literature lead us to feel confident that our approach to approximate the risk of recidivism for each youth is valid. We also believe that our findings are likely consistent with what we would have seen had we been able to use the YLS/CMI.

Figure 7: Average Expected Recidivism by YLS/CMI Risk Category



As pointed out by Lipsey (2008), it is important to remember that this expected recidivism score does *not* indicate the risk of the adolescent recidivating if given no services. This evaluation does not have a control group of adolescents who received no services. The regression equation is based on an analysis of the entire of sample of adolescents who received at least one of the services with a SPEP™ score. As a result, the expected recidivism risk indicates the likelihood of that adolescent recidivating if they received some type of the services being examined. This score provides an estimate of how likely this youth is to recidivate based on background characteristics and assuming that he/she received the average treatment effect across all the services tested.

IV. ANALYTIC RESULTS

IV.A. General analytic approach

The key question for the proposed project was to determine if, and how, SPEP™ program ratings are related to recidivism of the adolescents receiving the rated programs. As noted in the introduction, this has been addressed in some prior research, but the applicability of these findings to practice in Pennsylvania has not been examined. This project analyzes how SPEP™ ratings are related to outcomes in the juvenile justice systems in counties across the Commonwealth. In doing so, these data help to identify ways to focus SPEP™ practices in ongoing, future JJSES efforts in Pennsylvania.

The analyses of the SPEP™ ratings and recidivism reported here use the cohort of adolescents receiving a particular service as the unit of analysis. There are 162 cohorts in the data set receiving 13 different types of services. In the majority of the analyses outlined below, the effect of interest is the difference between the observed *recidivism* in a particular cohort compared to the *mean expected recidivism* for that cohort. It is worth reviewing how we determine these values for comparison.

The *observed recidivism rate for a cohort* is simply the percentage of the cohort who actually recidivate. If there is a hypothetical cohort composed of ten youths and 4 them end up with an adjudication or conviction within six months of leaving the service, then the observed six-month recidivism rate for that cohort would be 4/10, or .4. This is simply the prevalence rate of recidivism for that cohort. This value is calculated for each of the cohorts examined.

Now recall that each adolescent in the sample has been assigned a score that indicates their (that individual adolescent's) likelihood (chances) of recidivating, based on their particular characteristics when they enter a program (e.g., number of prior offenses, age). This is each youth's expected recidivism likelihood. The *expected recidivism of a cohort* is calculated as the mean (average) of the expected likelihoods of recidivism of the adolescents in that particular cohort. For example, think again of our hypothetical cohort of 10 youths who received a particular service. Let's assume that these 10 cohort members had the following individual expected likelihood scores for recidivism: .2, .2, .9, .6, .6, .4, .7, .8, .7, and .5. The expected likelihood of recidivism for that cohort would be the sum of each adolescent's individual likelihoods (.2 + .2 + .9 + .6 + .6 + .4 + .7 + .8 + .7 + .5 = 5.6) divided by the number of adolescents in the cohort (n=10). This would give the mean expected likelihood score of .56 for that cohort. Each of the 162 cohorts connected to a service has such an expected recidivism value based on the expected recidivism values of the youths in that cohort.

For our hypothetical cohort, then, the difference between the *observed* likelihood of recidivism and the *expected* recidivism would be .40 minus .56, or -.16. This would indicate that the observed recidivism was 16% lower than we would have thought it would be, based on the characteristics of the individual youth who made up that cohort. In this hypothetical cohort, the service was having an effect over and above what we would have expected. Obviously, the

mean differences calculated for each cohort can indicate a better than expected or worse than expected effect (with negative values indicating a better than expected effect and positive values indicating a worse than expected effect).

The difference values for each service cohort can then be examined as a sample of observations indicating the general positive or negative effect of that type of service. To extend our example, let's say that the difference calculated above (the value of -.16) was for a service classified as "individual counseling." It would then be one of a set of cohort level observations of the effectiveness of individual counseling, since there is more than one cohort of each service type to which the calculations described above are applied. We can then compare the magnitude of the effects (either increasing or decreasing recidivism) of the different types of services, or other categories of service provision (e.g., community or residential services) by using the difference scores associated with each cohort in the groups of interest. There are statistical issues that must be addressed in doing these comparisons among cohorts of different sizes and services with different requirements, and these are addressed as each particular analysis is discussed.

IV.B. Questions Addressed

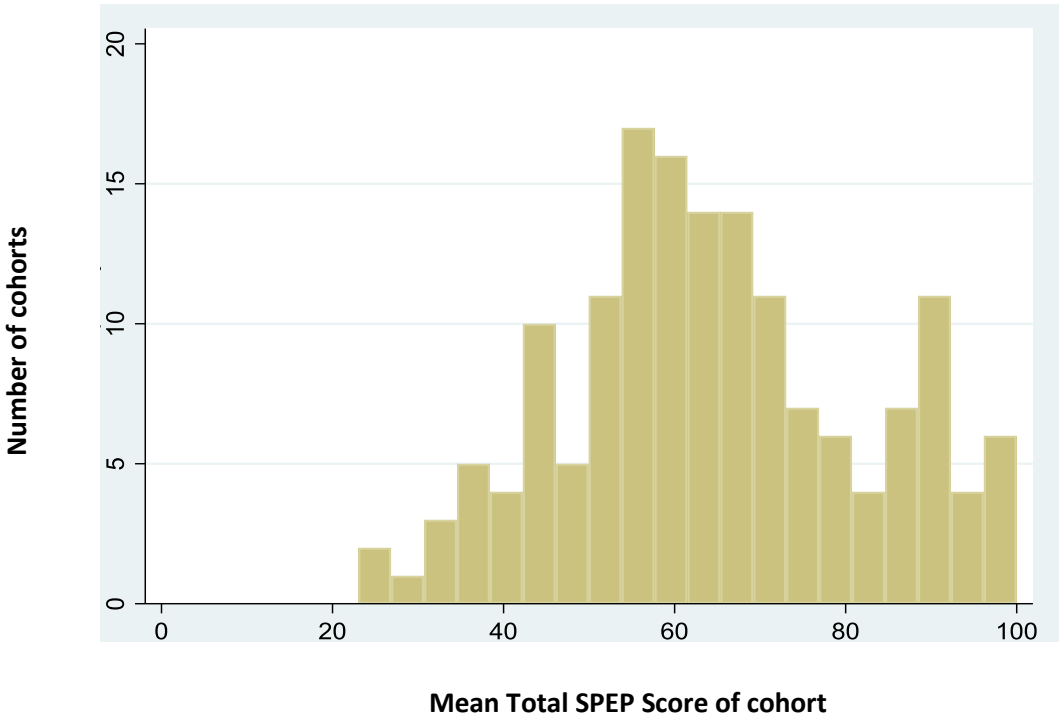
Data analyses focused on four basic questions. These questions reflect central issues related to the effectiveness of the SPEP™ approach as it has been applied in Pennsylvania. These are presented below.

1. ***What is the overall relationship between the SPEP™ scores and recidivism outcomes? Are higher SPEP™ scores related to larger differences between observed and expected recidivism rates?***

The basic issue here is whether better performance on the SPEP™ rating system translates into better performance at reducing recidivism. It is important to know if higher SPEP™ scores indicate services that generally perform better in terms of recidivism if this metric is going to be used to determine overall service effectiveness for a variety of services across the juvenile justice system.

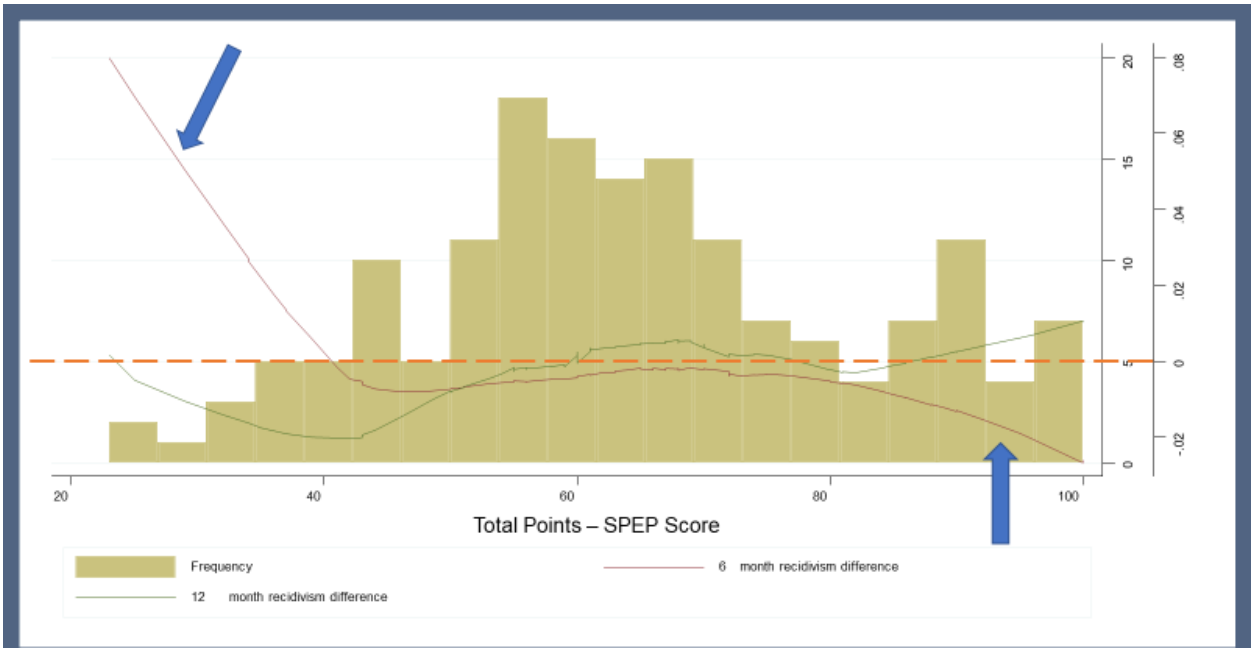
We first examined the range of average SPEP™ scores across all the cohorts. This is necessary to see if there is an adequate distribution SPEP™ scores across cohorts to make the test of an association informative. As can be seen from Figure 8 below, cohort scores (N = 158) are distributed in a relatively (but not totally) normal fashion, from a score of 23 to 100, with a mean of 61.9 (s.d. = 15.4).

Figure 8. Frequencies of Mean SPEP™ Total Score of cohorts (N=158)



We then examined the relations between the SPEP™ total scores and the differences between the observed and expected recidivism rates across the cohorts. Figure 9 below illustrate this relationship. The bars on the graph indicate the number of cohorts with the corresponding level of SPEP™ total scores indicated across the bottom of the figure (these values are indicated on the first Y-axis scale on the right of the graph, going from 0 to 20). The lines indicate the smoothed curve of the recidivism differences observed at each level of the SPEP™ total score (the values of the recidivism differences are indicated on the second Y-axis scale on the right of the graph, going from -.02 to .06). The maroon line indicates the recidivism differences for the cohorts at the six-month follow-up point; the green line indicates the recidivism difference scores at the twelve-month follow-up point. The dashed red line indicates the value of zero for the recidivism differences. The dashed zero line indicates where the recidivism difference for the cohort would be the same as expected; in effect, no positive impact from the service. Being above the dashed red line would mean doing worse than expected; being below the dashed red line would mean doing better than expected.

Figure 9. SPEP™ Total Scores with Raw Recidivism Difference Scores



The general pattern of these relationships are in the expected direction. For both the six-month and twelve-month outcomes, the recidivism differences for the cohorts scoring low on the SPEP™ ratings (pointed out by the left blue arrow) are markedly positive (indicating poorer performance than expected). The six-month line then indicates that the services at the top end of the SPEP™ ratings (pointed out by the right blue arrow) go downward (indicating better performance than expected). There do not appear to be clearly observable differences among the services scoring in the mid-range of the scale.

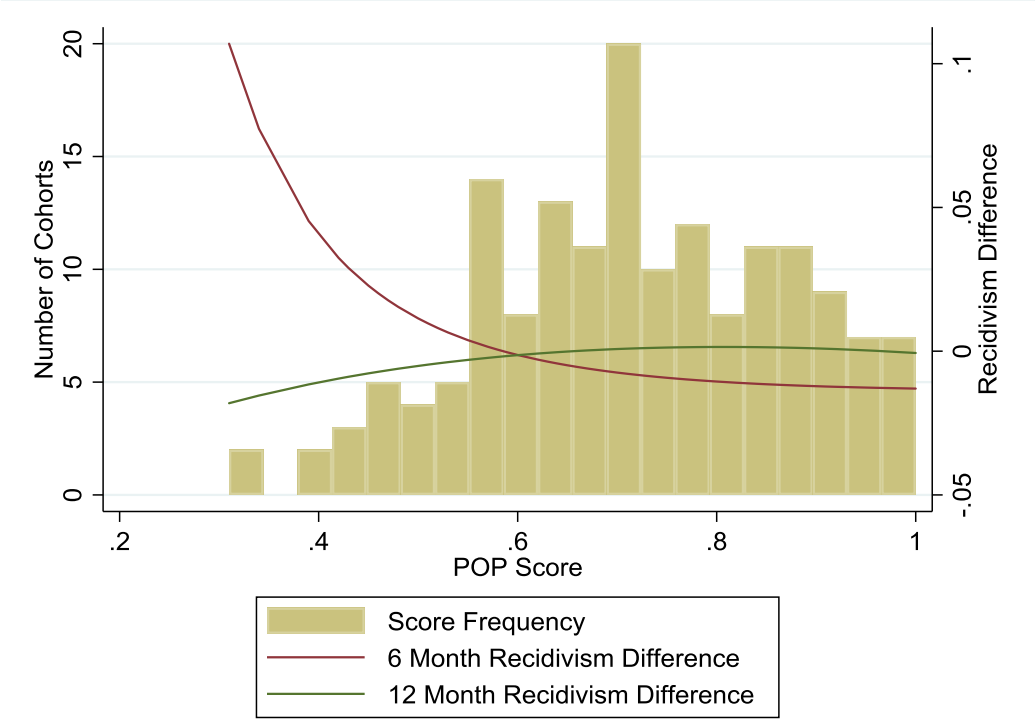
The overall association between the SPEP™ total scores and the difference between the expected and observed recidivism rates, however, was low and not statistically significant. The Pearson correlation between SPEP™ Total Score and the recidivism difference was not statistically significant at the six-month follow-up (r = -.10; n.s.; n=141) or the twelve-month follow-up (r = .05; n.s.; n= 140). In addition, there were no statistically significant difference in the recidivism difference scores for either the six-month or twelve-month follow-up points when the scores were tested based on a median split. That is, the cohorts scoring in the lower half of the distribution (those below 62) were not significantly different in recidivism differences than those scoring in the upper half.

These tests of the association between the SPEP™ ratings and the difference scores indicates that, over the full range of the scale, the correlations essentially fluctuate randomly around zero. There is no *strong linear relationship* up or down over the full range of the SPEP total score. The large number of cases scored in the mid-range of the scale probably drives this

overall test of the association. This flattening out of the difference values in the mid-range of the scale contributes a substantial number of nonsignificant observations to the overall test of the relationship of the SPEP™ scores and the recidivism differences.

We also examined the association between the average Program Optimization Percentage, or POP, score (see description of this score on page 14 above) and the average recidivism difference scores across the cohorts. It is reasonable to examine this association because the POP score provides a related, but slightly different, metric on the relative performance of a primary service. It essentially scales the SPEP™ total score to account for the type of service being assessed. The POP score provides a value between 0 and 1 that indicates the SPEP™ score value of a service relative to the total possible score that a primary service of that group/type can receive. The POP score thus provides a figure reflecting how much that service (according to the SPEP™ assessed dimensions) reflects the ideal for that type of service. The distribution of the POP scores of the cohorts and the smoothed curves of the recidivism differences are shown in Figure 10 below (portrayed in the same way as the preceding figure).

Figure 10. POP Total Scores with recidivism difference scores for 6-month and 12-month recidivism



Analyses of the associations between the cohort POP score with a) the SPEP™ total score and b) the recidivism difference indicated significant relationships of interest. As would be expected, the correlation between the average POP score and the SPEP™ total score was high and significant ($r = .93$; $p < .001$). In addition, the correlation between the average POP

score of the cohort and the six-month recidivism difference was significant ($r = -.17$; $p < .05$, two-tailed). The association between the POP score and the twelve-month recidivism difference was not significant ($r = .02$, n.s.).

The line for the twelve-month recidivism differences is essentially flat, indicating no significant difference in the observed and expected scores as the POP score increases. The statistically significant relationship indicated by the smoothed curve for the six-month recidivism differences, however, indicate a notable pattern. As can be seen in the figure, there is a generally downward, but curved relationship of the POP score with the magnitude of the recidivism rate differences for the six-month recidivism follow-up period (the maroon line). The difference between the observed recidivism and the expected recidivism for the six-month follow-up gets smaller and eventually is negative as the POP score increases (crossing the value of zero and indicating more favorable recidivism outcomes at about a POP score value of .6). The services with higher POP scores eventually reduce recidivism below the expected values at the higher end of the scale.

In summary, this initial examination of the relations between the SPEP™ total and POP scores with recidivism differences provides a very general, but generally positive, view of how these scores reflect program value. There are underlying dynamics behind these general relationships that have yet to be explored. There are, however, a few notable points about these observed associations.

First, the distributions of SPEP™ total scores and POP scores across the cohorts both indicate a reasonable "spread" of scores. It is clear that the SPEP™ process is making relevant distinctions among the services examined and that services are distributed rather normally across the possible range of scores. In some earlier research, there were often a small proportion of high performing services, making comparisons or tests across the full scale difficult to interpret. The rating system as applied here appears to be doing a sufficient job of differentiating among services.

Despite their high correlation with each other, the SPEP™ total score and the POP score appear to provide a slightly different level of accuracy in identifying services with higher likelihoods of reducing recidivism. The overall association between the SPEP™ total score and recidivism reduction is lower than might be expected, showing no statistical association with recidivism reduction across the range of the scale for either six- or twelve-month recidivism.

Scores on the POP score scale, on the other hand, are significantly related to recidivism differences for the six-month follow-up period. One can see from the illustration of these relationships (Figure 10) that the services at the low end of the POP score have marked differences in recidivism rates, with the observed rates higher than the expected rates. These differences reduce and reverse over the course of the scale, until the observed differences are lower than the expected differences, indicating better service performance. This seems to argue that use of both the SPEP™ total score and the POP score provides the most complete picture of the relationship to recidivism.

The lack of a strong, statistically significant relationship between the SPEP™ scores and the recidivism outcomes should not be taken as a demonstration of a lack of valid information or the utility of the SPEP™ method. It instead indicates that the picture of how SPEP™ is working at differentiating service performance is more complicated than simply providing a unitary metric of program effectiveness. In these data, the pattern is more complicated than a simple linear one where an equivalent increase on the scale produces a uniform shift in outcomes. The overall patterns of how the SPEP™ total score and POP score are related to recidivism differences provides a clear illustration of the relationship of these scores and program effectiveness. In both patterns, it is clear that low scoring services do much worse than expected, that there is little differentiation in the performance of services receiving mid-range scores, and that higher scoring services have more of an impact on recidivism.

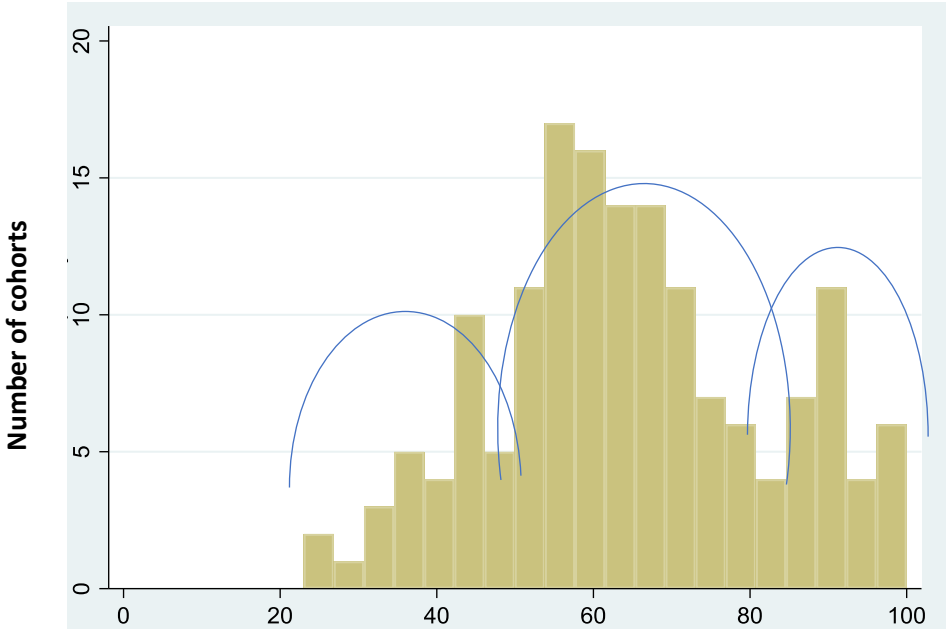
2. *Moving beyond the overall relationship between the SPEP™ total score and recidivism, can we identify ranges of scores along the continuum of the SPEP™ total scores that are related to reductions in recidivism?*

Previous work on "cutoffs" for determining a minimal standard for SPEP™ ratings have been rather blunt (e.g. above a total score of 50) and somewhat arbitrary. The approach taken here is to see if there are more nuanced and data-driven underlying groups of services with SPEP™ total or POP scores that might provide alternative goals for services to try to achieve. The value of this information is that it could help establish benchmarks for service providers to reach in their performance ratings.

There are two activities related to this set of analyses. The first one is to see whether there appear to be identifiable, underlying groups of service cohorts with SPEP™ total or POP scores within particular ranges. Second, if there are such groups, are the differences between the observed and expected recidivism rates statistically significant across these groups. The starting point for us is to simply do an "eyeball" examination of the distribution of the SPEP™ total scores to see if there appear to be some distinct underlying sets of scores. What we are looking for are some "natural breaks" in the scores that could indicate different underlying, "latent" groups, that would be valuable to consider in segmenting the SPEP™ total (and later the POP) distribution of scores.

Subgroups on SPEP™ Total Scores. The distribution of SPEP™ total scores across the cohorts has already been illustrated in Figure 8 and is shown again below. This time, however, we are showing our original intuition about the possible groups that might exist in the pattern of the overall distribution. These are marked by the three blue lines below, indicating sets of scores that might hang together apart from the overall distribution. If this intuition can be verified, we can then see if reductions in recidivism are different between the identified groups. For example, one group with scores within a certain range might all have rather similar differences in recidivism reduction while another group might show a much higher reduction in recidivism or a pattern of increasing effect as their SPEP™ total score rises.

Figure 11. Frequencies of mean total SPEP™ score of cohorts with possible subgroups



The initial step in this analysis is to determine if we can identify possibly informative groupings in the overall distribution of the total scores for the cohorts. This is done using a type of analysis that examines different possible solutions for grouping cases and tests the relative suitability of these solutions to the patterns seen in the data. It essentially examines the range of scores to see if there are ways to split the scores into groups so that each identified group is composed of cases that look more like each other in the group than they do to other cases in the larger sample. It finds distinct groups containing cases with a significantly higher probability of being in one of the identified groups than any of the other identified groups. [1]

This analysis produced a three-group solution that had a good fit to the patterns seen in the 158 cohorts examined. The specifics of the solution are shown in Table 6 below. We can

[1] We provide the following information for those interested in the technical aspects of the procedure used. Finite mixture modeling was used to investigate possible informative groupings in the total scores data. We used the *traj* plugin (Jones & Nagin, 2013) in Stata to perform the modeling at a single time-point with intercept only models per group. Time is ignored in the intercept only model, yielding a cross-sectional analysis. The censored normal model was used, censored at 0 and 100. Bayesian Information Criterion (BIC) was used to test the goodness of fit to select the best number of groups for the final model. Estimation was performed adjusting for cohort size. Probabilities of individual membership in each group are provided, directly presenting the uncertainty of group membership and providing an additional measure of model fit. Average posterior probabilities are as follows for the SPEP solution: low (.78); medium (.84); high (.89).

think of these as cohort groups having low, medium, or high SPEP™ total scores, with accompanying average scores in the ranges indicated (minimum, maximum values).

Table 6. Three group solution for SPEP™ cohort scores

Group	% of cohorts	N(cohorts)	minimum value	maximum value	mean score
Low	16.5%	26	23	43	35.9
Medium	61.4%	97	44	77	59.5
High	22.2%	35	80	100	87.7

It is worth noting that this is roughly the split that would be made on the sample if one were to divide it with a cohort group surrounding the mean, one group a standard deviation below the mean, and another a standard deviation above the mean. This is thus roughly a split of the sample into the lower 16%, the middle 67%, and the higher 16%.

Next, we tested whether the magnitudes of the differences between observed recidivism and expected recidivism were significantly different across the three identified cohort groups; i.e., whether the discrepancies between the observed and expected values across the low, medium, and high cohort groups could have simply occurred by chance. A test of the variation in recidivism difference scores across the three cohort groups indicated that the differences in recidivism outcomes across the three groups were significantly different for the six month outcomes ($F(2,137) = 3.02$; $p = .05$), but not for the twelve month outcomes ($F(2,138) = 1.07$; $p = .35$). The amount of variability among the differences in the different groups, however, was quite low, essentially fluctuating close to zero. The mean scores for difference between the observed recidivism and expected recidivism across the three cohort groups are shown below:

Table 7: Mean of Difference Between Observed and Expected Recidivism

		6 months	12 months
Group	**Low**	.03204	.00219
	Medium	-.00611	.00085
	High	-.00519	.01220

Although a statistically significant amount of variation among these groups regarding the recidivism differences at 6-months, it is clear that the amount of recidivism reduction detected in this three-group model was still low; between 1% and 3%. It is also clear that the statistically significant finding for group differences at the 6-month follow-up is the result of the differences between the low scoring group and the medium and high groups. The two higher scoring groups have almost the same recidivism difference scores (in the favorable direction), while the low group has a more sizable and unfavorable difference score. This indicates that these low scoring cohorts are distinguishable because they are producing observable recidivism rates well above the expected rates (they are performing significantly poorer).

Subgroups on POP Scores. The same approach to identify the possibility of subgroups was applied to the distribution of the POP scores. Again, we explored for the presence of underlying, normally distributed groups within the sample of cohorts, using the same statistical approach presented above. This approach also identified a solution with good fit to the data, with three groups in the distribution of POP scores.[2]

The three cohort groups identified in the distribution of POP scores can also be thought of as having low, medium, or high POP scores. The specifics of this solution for the POP score distribution is shown in Table 8 below. The cohorts having average scores in the ranges indicated (minimum, maximum values) are assigned to the respective group.

Table 8. Three group solution for POP cohort scores

POP Group	% of cohorts	N(cohorts)	minimum value	maximum value	mean score
Low	8.3%	13	.31	.49	.43
Medium	56.3%	89	.50	.78	.66
High	35.4%	56	.79	1.00	.88

We also then tested whether the magnitude of the differences between observed and expected recidivism were significantly different across the three identified POP cohort groups. Unlike the results for the SPEP™ Total Score groups, there was no overall significant difference among the three POP score cohort groups in the discrepancy between their recidivism scores. This was the result for both the six-month and twelve-month follow up point.

The above analyses indicate that there are empirically identifiable subgroups in the distributions of SPEP™ total and POP scores. The cohort groups identified are not based just on intuition about where the boundaries of group membership start and stop. Instead, the analyses find data-specific, justifiable points for considering a SPEP™ total or POP score as low, medium, or high based on the distribution of scores in this sample. These analyses provide a clear picture of how these scores are assigned in a three-group pattern and where the markers for those "naturally" emerging groups are.

The argument for these groups indicating clear differences in effectiveness is less strong. There was significant variation in recidivism differences among the identified SPEP™ Total Score cohort groups for the twelve-month recidivism outcomes, but not for the six-month outcomes. There were no significant differences among the three POP subgroups. This would seem to indicate that there are not bright lines of effectiveness separating these groups. As mentioned

[2] The posterior probabilities for the three identified groups were as follows: low = .83, medium = .92, high = .89.

above, however, some of this lack of statistical significance may be attributable to low base rates of recidivism and limited variability in recidivism outcomes across both scales.

Overlap of SPEP™ and POP scores with dimensions of program operations. Our initial examination of the data convinced us that, at least in this sample, certain types of program operations were more likely to be present in services that are assigned a higher or lower SPEP™ Total Score and/or higher POP scores. This is not a surprise since that is exactly what the SPEP™ is designed to show; i.e., services with features validated to have strong associations with recidivism reductions should, by design, receive higher ratings. Nevertheless, this is an important point to keep in mind when interpreting and assessing the utility of the SPEP™ process for future application.

An examination of the mean SPEP™ total and POP score by dimensions of program operations (i.e., primary service type, theoretical orientation, evidence-based/locally developed, and residential/community) provides one indication of this overlap. Table 9 below shows the mean SPEP™ total and POP scores according to different dimensions of program operations for services with at least 10 youth.

Table 9: SPEP™ total and POP score by dimensions of program operations (n=158)

Dimension of Program Operation		SPEP™ Total Score Mean (SD)	POP Score Mean (SD)
Primary Service Type*	Individual Counseling (n=16)	48.88 (10.72)	.65 (.14)
	Job-related (n=11)	42.64 (13.79)	.57 (.18)
	Remedial academic (n=10)	58.50 (11.08)	.74 (.14)
	Family counseling (n=13)	59.31 (9.07)	.70 (.11)
	Social skills (n=10)	59.00 (11.36)	.70 (.14)
	Group counseling (n=17)	59.35 (8.61)	.63 (.08)
	CBT (n=49)	83.14 (12.05)	.83 (.12)

Theoretical Orientation* (significance is group vs. all else)	Counseling (n=57)	55.21 (10.59)	.64 (.12)
	Restorative (n=10)	52.00 (5.80)	.65 (.07)
	Skill-building (n=91)	71.70 (18.69)	.77 (.16)
Evidence-based**	Locally Developed (n=115)	59.10(16.02)	.68 (.15)
	EBP (n=43)	78.98 (13.44)	.81 (.12)
Setting** (significance is residential vs. community)	Residential (n=106)	67.9 (18.35)	.75 (.15)
	Community (n=52)	57.60 (14.11)	.65 (.13)

*ANOVA used to test mean differences. SPEP total score and POP score means are significantly different for both primary service type and theoretical orientation ($p<.001$)

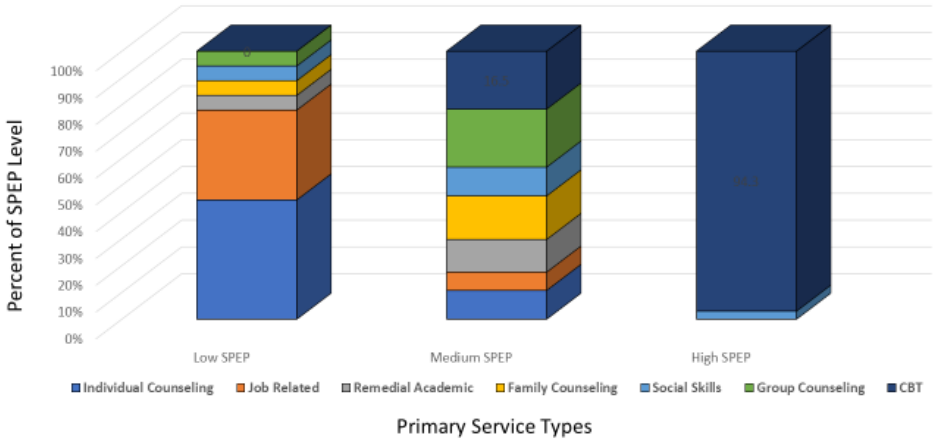
**t-test used to test for mean differences. SPEP total score and POP score means are significantly different for both evidence base and setting ($p<.001$)

As can be seen above, certain types of program operations score consistently higher on either or both the SPEP™ Total Score and the POP score. CBT services are rated considerably higher than any of the other primary service types; skill-building services are generally rated higher than counseling or restorative services; evidence-based services, on average, score higher than locally developed services; and, residential services are rated higher than community-based services. None of these differences are extremely large or particularly counterintuitive; for instance, services with more mature and defined protocols (e.g., EBP services) would seem to be more likely to have the necessary procedures in place to fit the SPEP™ criteria more closely.

At the same time, it is important to remember that the ratings given to any dimension of service are the result of the distribution of other dimensions of service within the categories examined. This can be seen in the difference in mean ratings of residential and community-based services. The mean ratings shown above indicate that residential services receive higher average scores than community-based services. This straight comparison (residential versus community), however, ignores the fact (as we will see below) that a large proportion of the residential services assessed were CBT programs (probably driving up the overall ratings of the residential services).

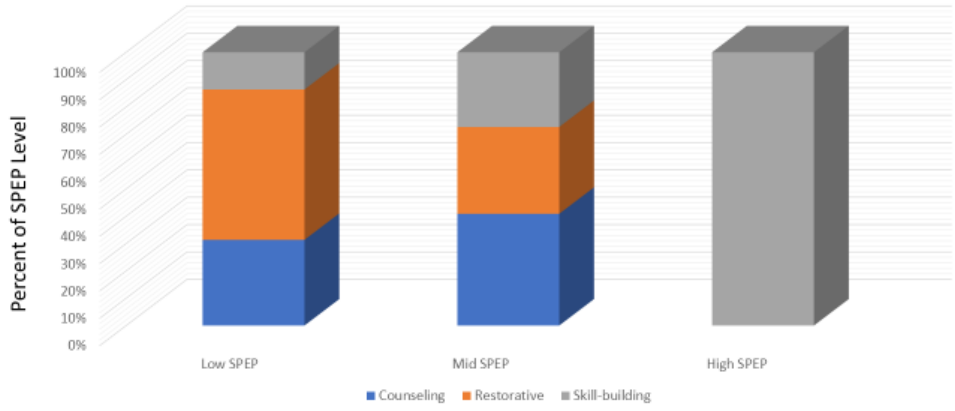
In the figures below (Figures 12– 15), we show the composition of each of the SPEP™ subgroups identified above (i.e., low, medium, high) in terms of the percentage of the cohorts in each subgroup that fit a particular dimension of service (e.g., primary service type, setting). Each of the overall associations between the SPEP™ total or POP score and the categorization of services examined is statistically significant.

Figure 12. Primary service types within each SPEP™ total score subgroup



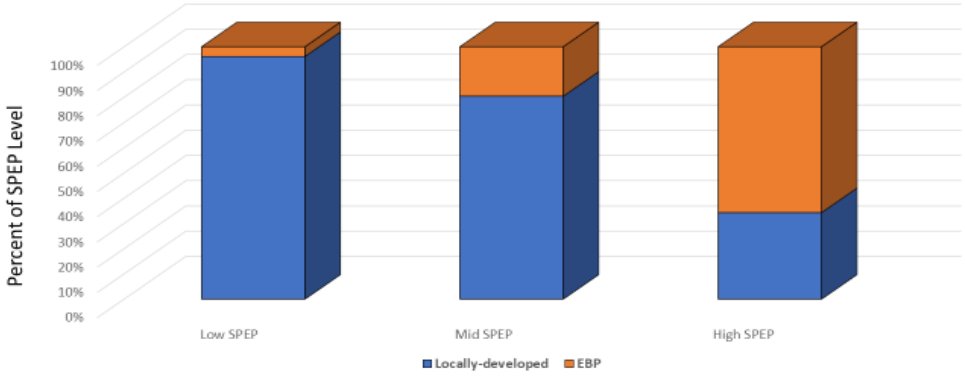
Chi square=122(24), p<.001

Figure 13. Theoretical orientation within each SPEP total score subgroup



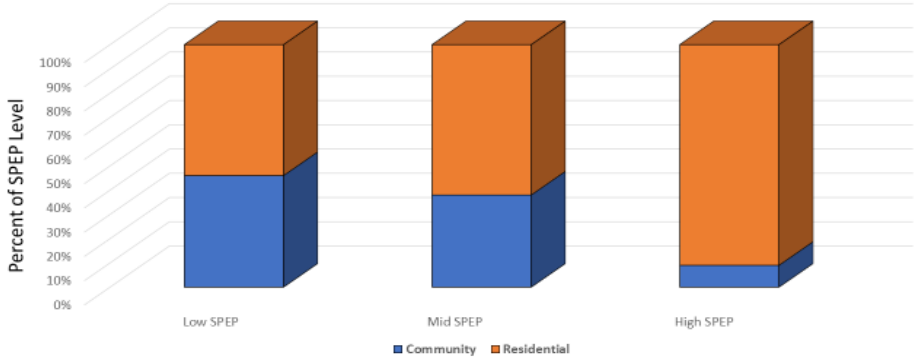
Chi-Square=36.65(4); p<.001

Figure 14. Evidence base within each SPEP total score subgroup



Chi-Square=36.21(2); p<.001

Figure 15. Setting within SPEP total score subgroup

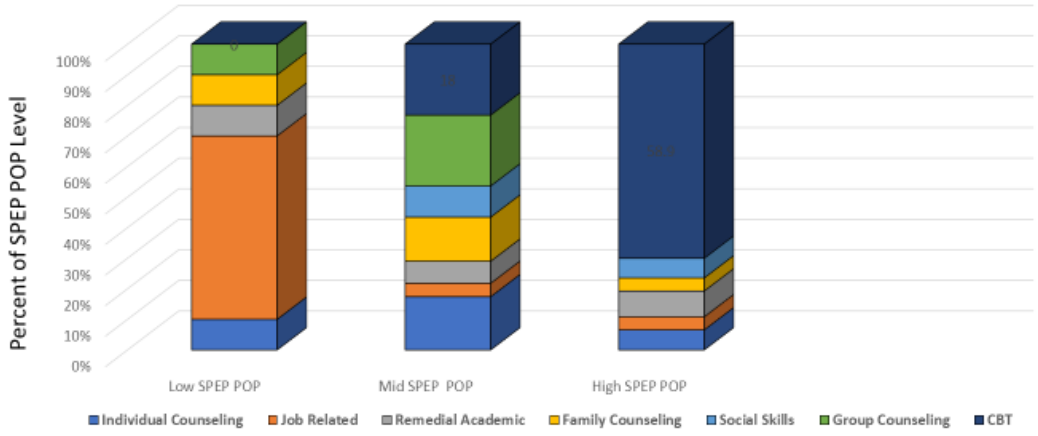


Chi-Square=12.66(2); p<.005

Figure 12 clearly shows that the cohorts receiving lower SPEP™ total scores are generally individual counseling and job-related services, while the cohorts receiving the highest SPEP™ Total Scores are almost all cognitive-behavioral therapy services. In addition, Figures 13 and 14 show that those cohorts receiving high SPEP™ total score are highly disproportionally skill-building and evidence-based programs. Finally, Figure 15 indicates that residential services are more likely to receive high SPEP™ total scores than services in community settings.

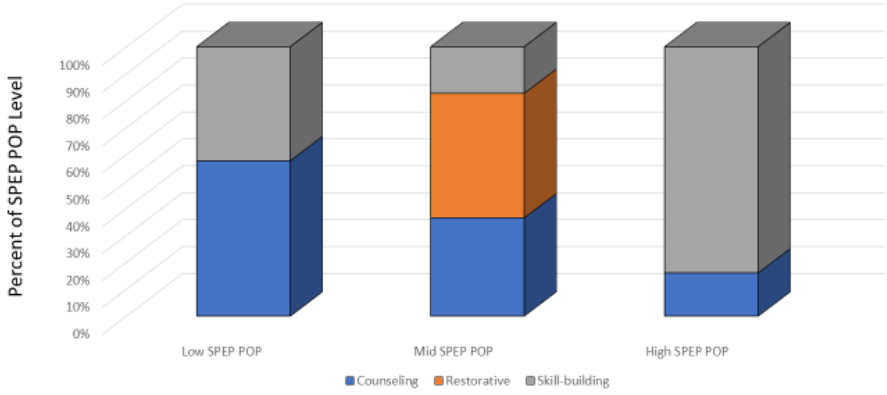
Figures showing the relationship of the dimensions of service provision and the POP score are provided below (Figures 16 – 19).

Figure 16. Service types within SPEP-POP score levels



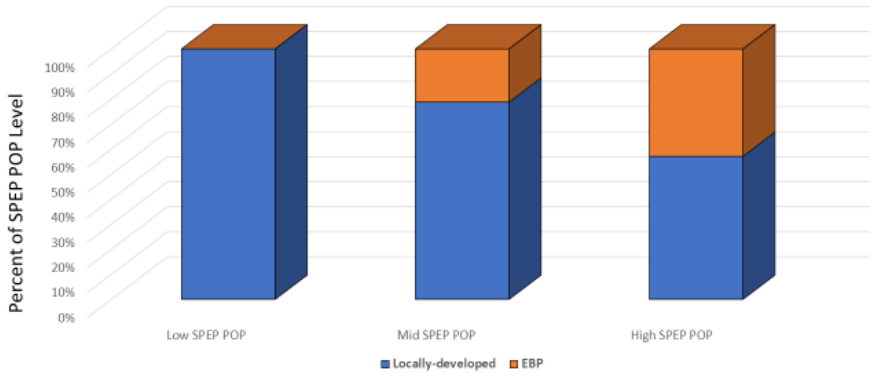
Chi-Square=97.04(24); p<.001

Figure 17. Theroretical orientation within SPEP-POP score levels



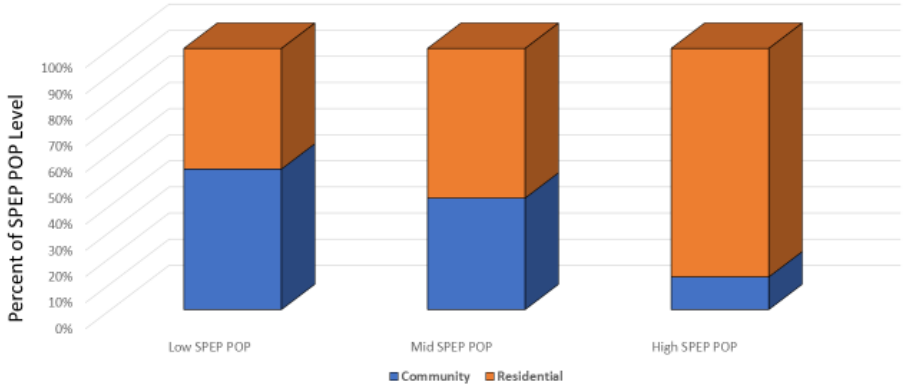
Chi-Square= 38.89 (4) p<.001

Figure 18. Evidence base within SPEP-POPscore levels



Chi-Square=13.32(2); p=.001

Figure 19. Service setting within SPEP-POP score levels



Chi-Square=17.01(2); p<.001

The uneven distribution of the dimensions of program operations across the POP score ratings are still apparent (and statistically significant), but the extent of the overlap is less pronounced with the POP score than with the SPEP™ score. This is reasonable because the POP

score introduces the type of service rated into its calculation, using a maximum score assigned to each service type as a denominator for that service's SPEP™ score. In a sense, the POP score partially compensates for the fact that, by the rules of the SPEP™ scoring procedure, some services have a predetermined ceiling on the number of SPEP™ points that they can receive. Even looking at the POP score, however, the tendency for certain types of services to be more or less likely to be rated within a particular range of scores is still evident. The POP score reduces, but does not eliminate, the regularity that types of services are significantly disproportionately represented at different levels of the SPEP™ and POP rating scales.

This overlap of dimensions of program operations with SPEP™ and POP score seems somewhat inevitable, given the goals and procedures of the SPEP™ process. If a high SPEP™ total score is meant to provide a metric for considering a service as more or less in conformity with indicators of best practices to reduce recidivism, then programs with well-developed protocols and procedures (and usually an accompanying restricted focus to intervention, such as thinking processes) would certainly be strong candidates for higher scoring. In addition, type of primary service in and of itself is a strong contributor to the SPEP Total Score. As a result, it seems unlikely that emerging or innovative services with less well-defined treatment targets and protocols would obtain a high SPEP™ total score. The POP score appears to adjust considerably for this association, thus again showing its possible utility for examining certain types of effects in future analyses. There still remains questions about whether particular aspects of program operations are related to recidivism outcomes in and of themselves.

3. Do particular components (e.g. primary service type, dosage) of the SPEP™ Total Score show significant relations to recidivism outcomes?

Services can be examined through the lens of various dimensions of program operations and the performance of these dimensions of the SPEP score can be compared to each other. For example, the sample of services contain several primary service types, and it is therefore possible to assess the relative impact of different types of services on recidivism. It may be that a particular primary service type (e.g., social skills training) has a stronger effect on the recidivism outcome when compared to the other primary service types.

We tested the relative impact of the levels of several dimensions of service. The dimensions examined were the same as those compared above: a) primary service type (7 of 13 indicators examined), b) theoretical orientation (i.e., counseling, restorative, skill building), c) evidence base (i.e., evidence-based versus local programs), and d) setting (i.e., residential/community). Comparing the relative performance of services along these dimensions provides some evidence that could be useful in formulating guidelines to promote more over less effective types of services at the state level.

Each of the dimensions were tested using the categories presented above as independent variables and the differences in the recidivism scores as the dependent variable. This allowed for consideration of whether a) a particular dimension overall (e.g., primary service type, theoretical orientation) was related to differences in recidivism and b) which

particular aspect of that categorization was most powerful compared to the others (e.g., do skill building services have a bigger effect on recidivism than counseling or restorative services?).[3]

Primary service types. Seven of the thirteen primary services types were examined in these analyses. These seven primary service types each had at least ten cohorts that contributed to the estimates of the expected and observed recidivism. As explained above, if a primary service type had less than ten cohorts connected with it, we considered the estimates of expected and observed recidivism to be unstable, and these services were not included as a result.

A series of regressions were conducted with the primary service type as the independent variable and the differences in recidivism as the dependent variable. The overall model including all seven primary service types as predictors did not reach statistical significance for the recidivism differences for the six-month follow-up point as the outcome. Also, none of the primary service types, when tested against each other, had a significantly greater effect on recidivism differences; none of the primary service types stood out as significantly more powerful than the others for this outcome. No primary service type independently showed an effect at reducing recidivism; they were basically indistinguishable in their relationship with the six-month recidivism outcome.

The examination of the twelve-month recidivism differences did show a statistically significant effect for the overall model including the seven primary service types ($F(6,106) = 2.97$; $p < .01$), indicating that there was significant variation in the twelve-month recidivism differences explained by considering the primary service types. The only primary service type showing significant differences in the recidivism outcome was the group providing remedial academic programming ($t=2.79$; $p < .01$). The difference, however, was not in the desired direction; the cohorts representing youths who received remedial academic programming showed more positive differences between observed and expected recidivism, indicating that these services more consistently raised the observed rate above the expected rate. In short, they have worse recidivism outcomes.

Theoretical orientation. The regressions for the effects of theoretical orientation showed no statistically significant differences with either the 6-month or 12-month recidivism differences as outcomes. No particular theoretical orientation proved to be significantly stronger in their relationship to the recidivism outcomes. Restorative practices did show a stronger positive effect with recidivism differences at the 6-month outcome point, but this

[3] The analyses described below were also conducted including the SPEP score as an additional independent variable. The purpose of these analyses was to see if the SPEP score was providing overlapping or independent information to that contained in the ratings of individual dimensions. In general, the inclusion of the SPEP score as a variable did not change the loadings or effects of the dimensions tested. The overall pattern of findings reported here were the same for these additional analyses. The general finding was that the SPEP score contributed some varying amount of information independent of the service dimension ratings and was not simply duplicating the effects seen by considering the dimensions alone. Reporting those findings in detail here seemed potentially more confusing than enlightening.

effect did not reach statistical significance ($p < .09$). Given the relatively smaller number of cohorts available to test for the effect of restorative practices (and the resulting reduction in statistical power to detect an effect), it is uncertain if this tendency is an indicator of a "real" difference.

Evidence-base. The regressions testing for the effects of whether a program was evidence-based or locally developed showed no significant findings with either the 6-month or 12-month recidivism outcomes. The recidivism difference scores between the evidence-based and locally developed programs were not statistically significantly different from each other at either point.

Setting. There were statistically significant effects for the setting in which the service was delivered. The comparison of service settings (residential and community) showed the most consistent results regarding recidivism differences. The recidivism difference scores of the community-based services and residential services were statistically significantly different at both the six-month ($t(1,139) = -3.89$; $p < .0002$) and the 12-month ($t(1,138) = -2.84$; $p < .005$) follow-up points. The recidivism differences for community-based services were significantly more favorable (lower observed rates than expected rates) at both follow-up points.

Amount and quality of service. The amount (i.e., dosage and duration) as well as the quality of the service could be critical to a service's success in reducing recidivism. If adolescents enrolled in a program do not get an adequate amount of time participating in program activities (dosage), it is unlikely that the program (regardless of its orientation or focus) could have a notable impact. Similarly, programs that do not maintain enough sustained contact with a youth (duration) or who deliver services with little integrity when they do have contact with a youth (quality) cannot be expected to have a powerful impact. These three aspects of program operations are rated separately in the SPEP™ process and therefore offer an opportunity for providers to examine their operations in light of these dimensions specifically and to focus improvement efforts in these areas. It is thus important to address whether differences in these aspects of program operations relate significantly to improvements in recidivism outcomes.

A series of regression equations were conducted to examine whether there was a significant association among scores on these three ratings in the SPEP™ improvement process and recidivism outcomes. All three ratings on these aspects of program operations were entered together to examine how much of the overall recidivism differences were explained by these particular aspects of program operations. The question is simply whether ratings of dosage, duration, and quality provide any significant information alone about possible recidivism reduction?

The regression equations testing the effects of dosage, duration, and quality on recidivism differences indicate some significant effects. The overall effects of these three aspects of programming, when taken together, do not have a statistically significant effect on recidivism differences at the six-month follow-up point. At the twelve-month follow-up point,

however, the overall effect for these three ratings is statistically significant ($F(3, 136) = 2.83$; $p < .04$), meaning that the three ratings taken together show a significant relationship to the variability in recidivism differences. This effect is primarily the result of the ratings for quality and dosage, not duration. The quality rating is significantly related to recidivism difference at this follow-up point ($t = -2.34$; $p < .02$), and the ratings for dosage approach statistical significance ($t = -1.89$; $p < .06$).

This set of findings highlights the importance of the ratings of dosage, duration, and quality as important factors to consider in the SPEP™ process. Even without the other dimensions of program operations taken into account, there is a systematic relationship between this trio of ratings and recidivism differences. A service's performance on these characteristics alone appears to be a telling barometer of the success of that service at reducing recidivism below what would be expected at the 12-month follow-up point.

All of the analyses reported up to this point test the associations between differences in observed and expected recidivism rates for services with different SPEP™ or POP ratings as well as the associations between ratings of various aspects of program operations and recidivism outcomes. This information is valuable to indicate how well SPEP™ or POP ratings reflect the subsequent performance of a group receiving a rated service. These analyses give us information about how SPEP™ ratings indicate better performance; testing the validity of our expectation that higher scale scores actually do indicate better performance on the outcome of most interest in juvenile justice. At this point, we take a slightly different approach and examine how a shift in SPEP™ Total Score in a service might or might not be related to better recidivism outcomes.

4. Is there a relation between improvement in SPEP™ scores and changes in recidivism rates for that same service?

The main purpose of the SPEP™ assessment process is to identify areas in which a service can be improved in order to make the most difference for the recidivism outcomes of the juveniles served. With this goal in mind, the SPEP™ process places a strong emphasis on communicating the findings of the review to the providers in a "performance improvement planning session." Each provider is given a Feedback Report that is reviewed together with the SPEP™ team and a response plan is developed with technical assistance provided as necessary. After a period of time, one or more additional service reviews by the SPEP™ team are conducted to determine if the service has improved.

An examination of SPEP™ scores over time provides a more stringent test of the value of the SPEP™ process. More specifically, examining whether increases in SPEP™ scores in the same service over time produce greater disparities between observed and expected recidivism rates would be a direct validation of the SPEP™ process (since quality improvement is the primary objective of gathering and using the ratings). Examining the effect of a service changing scores over time brings us a step closer to seeing a causal effect from improving

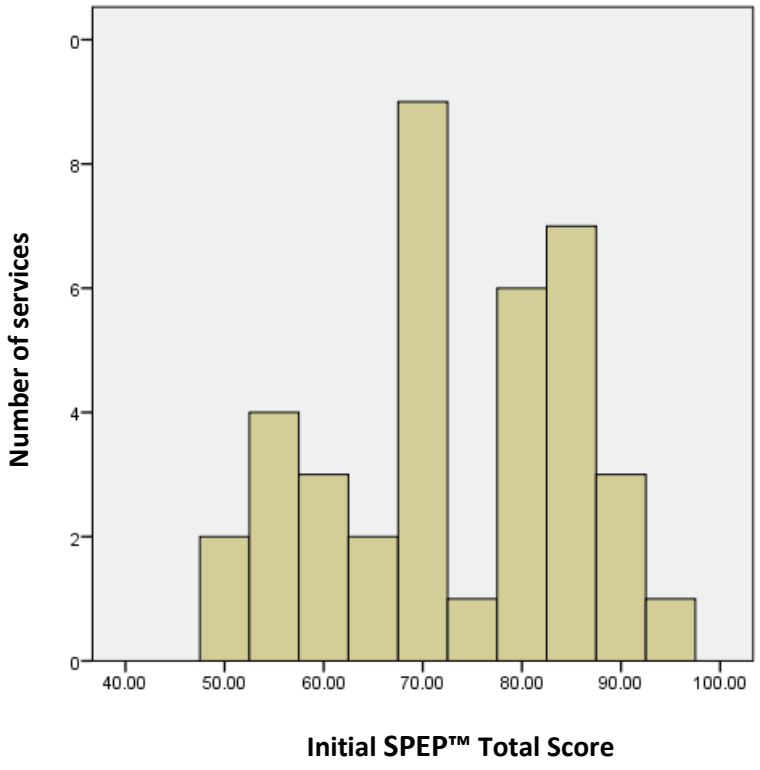
service performance, since it is the same service being measured twice (once before SPEP™ involvement and once after the performance improvement plan is implemented).

This evaluation provides a limited test of this sort. A sample of 38 services included in the initial sample had a second SPEP™ assessment after they reviewed their initial feedback report. Just as for the youth included in services at the time of the initial SPEP™ assessment (describe above), another cohort of adolescents who received the service during the time period reflected in the second SPEP™ rating was also drawn and their subsequent recidivism outcome data were obtained.

This allows for a comparison of outcomes in two different cohorts who received the same service rated during two different time periods. The question of interest is whether positive or negative changes in the SPEP™ or POP total scores for that service are related to positive or negative changes in the recidivism outcome. If services are improving in their scores, it seems that they should also show an impact on recidivism in a cohort of adolescents who receive the improved service.

Subsample of services examined. As noted, thirty-eight cohorts from the initial sample of 162 cohorts had a reassessment SPEP™ score. On average, the reassessment occurred 702 days (s.d. = 332 days) after the initial SPEP™ feedback report review date (range 170-1,714 days. Figure 20 below shows the distribution of SPEP™ scores across this sample of services. The average SPEP™ score for this sample is 73 (s.d. = 12.14).

Figure 20. Distribution of Initial SPEP™ Total Scores in sample rated twice



The SPEP™ total scores in the reassessment sample are generally normally distributed, ranging from scores in the 50's to those in the 90's. The reassessments have a slightly higher mean score (73) than the initial set of 162 cohorts (62), indicating that the reassessment cohorts have fewer services scoring in the lower range of total SPEP™ scores. The reassessment sample is composed of services from primarily three service types: cognitive behavioral programs (n = 14; 37%), family counseling (n = 6; 16%), and behavioral contracting (n = 4; 11%). Each of the remaining primary service types is represented by only 1, 2, or 3 cohorts. Regarding theoretical orientation, the cohorts in this sample are categorized as skill building services (n = 24; 63%) and counseling services (n = 14; 37%); there are no restorative services represented. Also, the cohorts in the reassessment sample are more likely to be locally developed services (n = 24; 63%) than evidence-based (n = 14; 27%). Finally, the reassessment sample is evenly split between residential and community services (n = 19, or 50% of each type).

It is difficult to characterize the overall representativeness of this subsample of cohorts to the larger group of cohorts that were part of the initial SPEP™ data. Based on the general descriptions above, however, there are a few differences between the groups that could be relevant to interpretation of the results of the subsequent analyses. First, the subsample with reassessments is skewed toward higher SPEP™ scores, indicating that any tests of change in this group might be indicating the effect of program improvements for services at the higher end of the SPEP™ scoring scale. Second, the reassessment sample is composed largely of primary service types that are associated with better records of reducing recidivism (CBT and family counseling). This could potentially limit the generalizability of the findings to services with more defined protocols and a higher level of structure. In sum, even though the sample is relatively small and somewhat limited, it still appears to have sufficient service variability to warrant an initial examination of the patterns of change and their impact on recidivism.

Examining the differences in SPEP™ and POP Total Scores. Figure 21 below shows the distribution of the difference on the SPEP™ Total Scores for the 38 cohorts and Figure 22 shows the same for the POP scores. Services at the zero point did not show any difference between their initial SPEP™ score and the subsequent SPEP™ rating. They received the same score at both times. The services scoring positively (above zero) showed an improvement in scores (the values on the X-axis indicate the number of SPEP™ Total Score points of improvement). The services scoring below zero went down in their ratings (again with the value indicating the number of SPEP™ Total Score points "lost" in the second rating).

Figure 21. SPEP™ Total Score differences in services rated twice

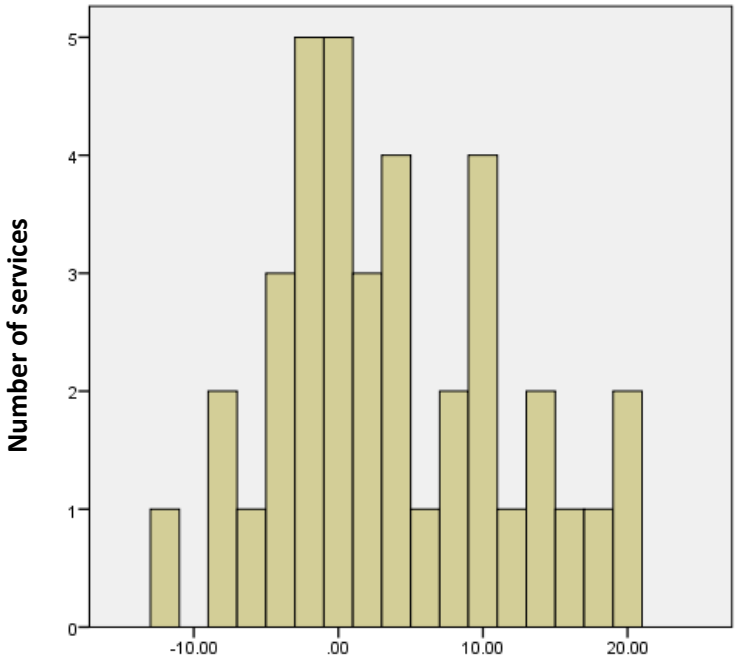
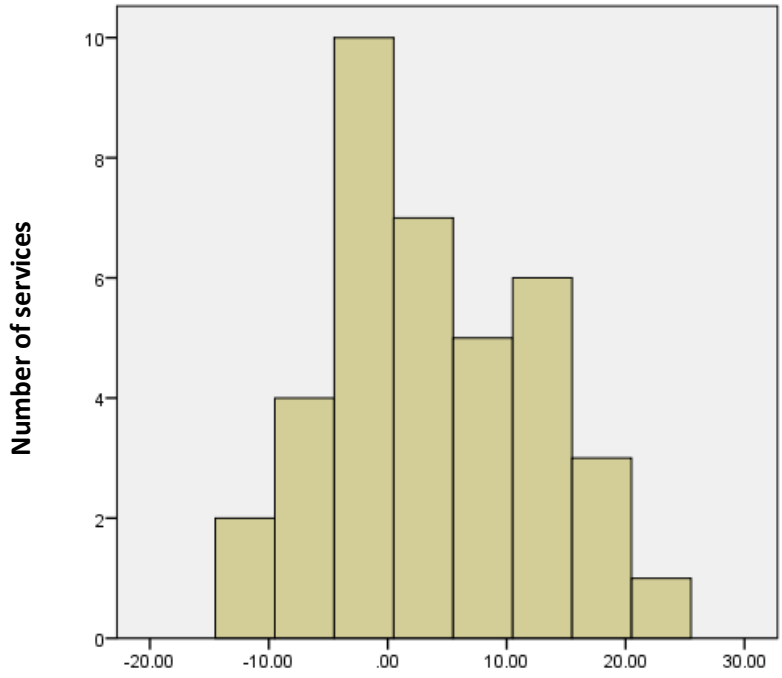


Figure 22. SPEP™ POP Total Score differences in services rated twice



It is interesting to note that not all the services improved in their SPEP™ total and corresponding POP score from the initial assessment to the reassessment (see Appendix E for changes seen in each cohort).

The overall mean SPEP™ Total Score difference is 3.16 points, a movement toward improvement for the whole group. However, 17 of the 38 (45%) services had no change or a negative change in their score (SPEP Total and POP Scores). Further analyses indicated that the services with positive changes (when compared to the group with no or negative change scores) were more likely to be locally developed (not evidence-based practices), community-based, counseling services. Although the current sample is just 38 cohorts from a single locale (i.e., Pennsylvania), it may indicate that the SPEP™ process holds greater promise for systematizing the practice of services that are still forming a routinized and documented approach to service delivery.

An analysis of the changes from the initial SPEP™ and POP Total Scores to the reassessment SPEP™ and POP scores for the whole group indicated that, despite some services getting lower ratings, these shifts were positive overall and statistically significant (SPEP™ score t-value (31) = 2.03, $p<.05$; POP t-value (31) = 2.09, $p < .05$). As a group, these services improved positively more than would be expected by chance. The shift toward positive score changes is greater than what would have happened if services were shifting randomly (in that case, the overall shift of scores would be zero).[4]

Examining the relationship between SPEP™ total and POP score changes and recidivism differences. An analysis was conducted to see whether an improvement (or decrement) in overall SPEP™ score for a service cohort was associated with a greater (or reduced) disparity between expected and observed recidivism rates for the cohort. The question examined here was whether change in the SPEP™ total score over time was related to a more positive recidivism result over the same time. If a service had a lower SPEP™ total score when it was reassessed, were the recidivism differences of the reassessment cohort less favorable than the ones seen at the initial assessment? If the service had a higher SPEP™ score upon reassessment, did the recidivism outcomes look more favorable for the reassessment cohort? For this analysis, we are ultimately interested in the association between a) the change in the SPEP™ total or POP score and b) the change in recidivism difference scores (the observed rate minus the expected rate) at the two assessment points.

It is important to remember that although the 38 services are the same for the initial and reassessment SPEP, the youths in each cohort are not. As noted earlier, reassessments took place an average of 702 days after the initial SPEP™ assessment and there is no overlap of youths in the two samples. This opens up the possibility for differences in the characteristic of

[4] Five T1 cohorts had $n < 10$ and one T2 cohort has $n < 10$. In line with other analyses, these cohorts were excluded, leaving 31 cohort pairs for this analysis. Since cohorts did not have the same size at T1 and T2, we made an adjustment for the unequal sizes by using the averaged cohort sizes at the two times and weighted regression on the difference of the SPEP and POP scores at T2 and T1. Then in the regression, a test for intercept = 0 is the test of interest.

the youth who were receiving the service concurrent to the SPEP™ assessment. One cohort, for example, could have more serious offenders than the cohort drawn for the other period. As a result, the differences between the observed and expected recidivism rates are useful as an outcome measure to compare because they "correct" for this possible difference between the cohorts for every service.

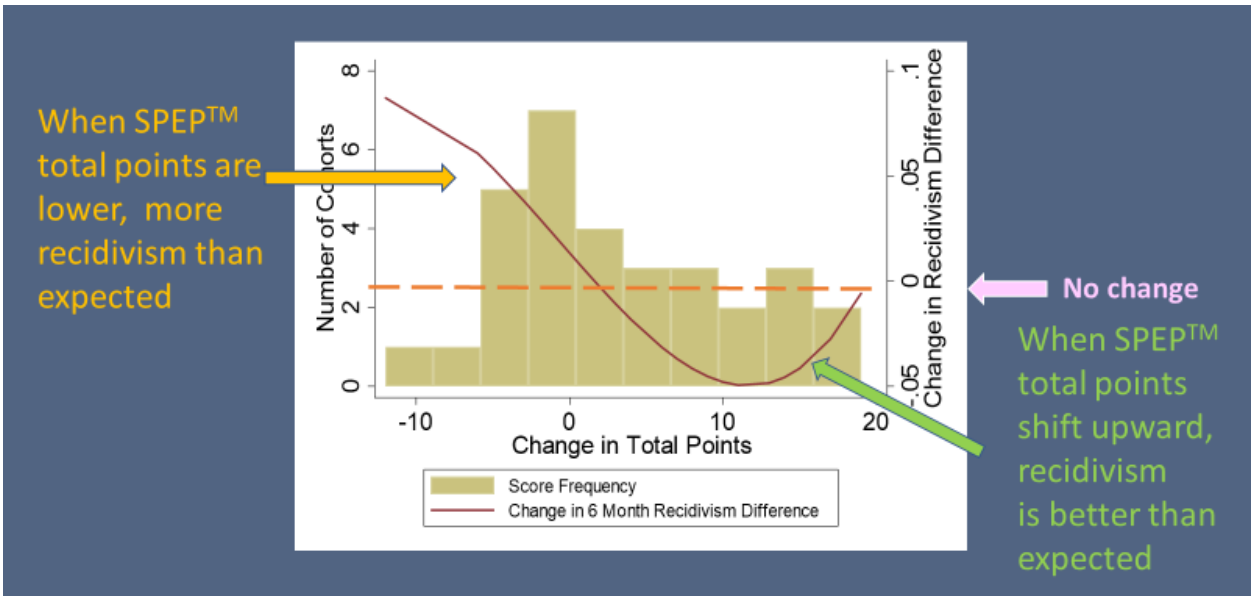
A series of regressions examined the relationship between a) the amount of change in the SPEP™ or POP Total Score from the initial assessment to the reassessment and b) the differences in the observed and expected recidivism rates of recidivism for the initial and reassessment cohorts. These analyses tested whether there was a significant relationship between these two types of change from the time of initial assessment to reassessment. These analyses examined whether the "difference" in the SPEP™ or POP Total Score was related to the "difference in the differences" in the recidivism rates at the two times. In short, was there a significant association between the magnitude of the shift in the SPEP™ and POP Total Score and the magnitude of the difference in the recidivism outcome measure?[5]

The results of these regressions indicated that some changes in the SPEP™ total and POP score were related to changes in recidivism over time. The amount of change in total SPEP™ points was significantly related to differences in the recidivism rates at the six-month point ($t(1, 29) = -2.03$; $p < .05$), but not at the twelve-month point for this sample of services. Similarly, the difference in the POP score approached significance ($p < .06$) at the six-month point, but not at the twelve-month point. These results indicate that the amount of shift in SPEP™ total and POP scores corresponds to a shift in the recidivism outcome differences in the short term (six-months after leaving the service).

Figure 23 below illustrates the relationship between shifts in the SPEP™ total score and the recidivism differences for the cohorts receiving that service. As indicated below, the recidivism difference scores improve (go below zero) as there assessed services improve their SPEP score. Services that achieve lower scores on their reassessment have recidivism difference scores going in the opposite (higher) direction; they are doing worse than expected.

[5] As in the previously reported analyses, we again accounted for the differences in cohort size at the different assessment times by using the averaged cohort sizes at the two times and weighting the differences in the regressions.

Figure 23. Differences in SPEP™ Total Score upon reassessment and the differences in recidivism outcomes



Effects in the change in recidivism outcomes related to service type. The small size of the reassessment sample precludes conducting a full set of analyses examining the relationship between service type and the changes in recidivism between initial assessment and reassessment. Many of these analyses would simply not provide a valid test of this relationship. The numbers of services considered become rather small when subgroups of a sample of 38 are considered (even before cohorts are eliminated because the number of individuals in the cohort are less than 10). For example, regarding the primary service types, the analyses seeing whether the recidivism outcomes differed in a group of fourteen cohorts receiving CBT versus six cohorts receiving individual counseling would be unstable and misleading. In addition, the size of the differences between or among such small groups would have to be extremely large to produce statistical significance. Similarly, it is not possible to do analyses of differences among services based on theoretical orientation, with only two of the three possible types represented.

Two exploratory comparisons were made, however, using two of the dimensions of program operations. The evidence base and setting were examined to see if these aspects of program operations were related to the magnitude of difference in the recidivism rates at the six-month and twelve-month follow-up points. In these analyses, the difference in the recidivism measure approached significance for the setting (residential/community) at both the six-month ($t = -1.87$; $p < .07$) and twelve-month ($t = -1.84$; $p < .08$) follow-up points. Community-based services showed a larger difference in recidivism rates (better performance than expected). The tests for the relationship between evidence-based/locally developed services were not significant.

Summary. The analyses reported above provide a previously unavailable view of the performance of the SPEP™ process over time. Using a relatively small sample of services with two SPEP™ assessments, these results indicate a positive relationship between an increase in ratings and recidivism outcomes. These analyses provide a view of what happens when services actually improve their performance, rather than just a view of the static relationship between SPEP™ rating and recidivism.

The analysis has some notable strengths and weaknesses. It has the strength of a sample of services examined at two time points and scored independently at each point. It also has an outcome measure of recidivism that is "corrected" for the expected recidivism rate of the adolescents included in the cohort examined. At that same time, the sample examined is rather small, and its representativeness of the full range of services of interest may be questionable. Yet because of its small size, the differences in outcome have to be rather substantial in order to reach statistical significance. Unfortunately, the small sample size does not allow for more elaborate tests of the relative influence of different dimensions of service on the differences in the assessment scores. While these findings are encouraging, they certainly require a larger replication to establish full confidence in their findings.

V. CONCLUSIONS AND RECOMMENDATIONS

The purpose of this validation study is to provide a picture of how the SPEP™ process is functioning in the state of Pennsylvania currently. There is no attempt to pass a singular, broad judgment on whether this process is "working" or not, since there is no clear, objective metric for making such a determination. The SPEP™ is not an intervention with expectations to reduce recidivism. Rather, the SPEP™ process is a method for moving juvenile justice services toward more uniform, empirically demonstrated practice that should improve recidivism outcomes. By combining data from several sources, this validation study has been able to provide a multifaceted view of how SPEP™ performs and how it might be improved as it moves forward.

Before addressing the points highlighted by this evaluation, it is first important to acknowledge the complexity of getting the SPEP™ assessment process started and sustained. A large number of interacting parts and a larger number of individuals are required to put the SPEP™ into operation statewide. Political and juvenile justice leadership, service provider buy-in, sustained funding, and competent technical assistance are all needed to launch this type of effort and to keep it afloat. SPEP™ is an innovative practice that few locales have been able to "pull off" successfully. Pennsylvania has served as an exemplar of how to achieve the goal of improving juvenile justice services on a large, rather than piecemeal, approach as part of a systematic statewide plan. This makes this initial look at its operations important for guiding where such efforts might go.

The process of this validation study and the analyses presented above lead to several general conclusions, each with implications for the future operation and monitoring of the SPEP™ process. These conclusions and recommendations are presented below.

CONCLUSION #1. Pennsylvania is a leader in the implementation of the SPEP™ process and, as such, has an opportunity to foster nationwide progress in quality improvement in juvenile justice services. In order to fulfill this potential, some improvements in data management and analysis tailored to the SPEP™ process would be desirable.

As pointed out in the first part of this report, few states have successfully implemented, documented, and evaluated their use of the SPEP™ process. Pennsylvania, along with a few other states, is at the forefront of this innovative practice to systematically improve service quality. The above analysis of the SPEP™ process indicates that the state is moving in the right direction to incorporate SPEP™ into the ongoing juvenile justice service environment; promoting a generally consistent assessment process and generally demonstrating expected associations between SPEP™ ratings and recidivism outcomes.

The challenge of collecting and integrating relevant data to assess the implementation of SPEP™ and its performance was discussed in some detail earlier in this report as background for the discussion of the analytic plan. However, the issues noted also highlight some of the challenges that must be faced to improve and monitor the SPEP™ process as it moves forward. With the assumption that there will be additional validations of SPEP™ and continuing partnerships with researchers in the future, there are several data management issues that, if addressed, could ease the process and increase the value of future work. These suggestions are rooted in our own experience and conveyed with an understanding that Pennsylvania is always looking to improve and lead the nation in smart, data-driven policies.

First, increased effort could be directed toward ensuring valid data on the youth receiving services. The validation of SPEP necessarily relies on the ability to examine the outcomes for youths in the services that are assessed. Historically (or at least prior to this validation), the EPISCenter has reasonably had a primary focus on collecting and analyzing the SPEP data elements, with less emphasis on monitoring the quality of the youth-specific information. Going forward, both the EPISCenter and service providers could apply more resources to maintaining good records regarding the youth in the services (e.g. identifiers, risk/needs, strengths). Doing so would allow more complete and accurate matching with information provided by JCJC regarding youth outcomes. Regular, ongoing communication between the EPISCenter and JCJC regarding the identification of youth represented in the SPEP™ services would allow for monitoring who is getting which services as well as detecting and avoiding problems that could impair attempts to match the two data sources for future research inquiries.

Some of the issues of data coordination could be addressed by systematic record keeping of data structures and variable definitions. For example, the EPISCenter could develop documentation of existing data sets, variables, and values that would serve as a resource for outside investigators. This documentation could be valuable for ongoing data quality checks and working with service providers and county probation departments to ensure full and complete data. Completing these tasks on an ongoing basis could ease the process of data transfer and documentation for future validations studies.

Second, improvements to the collection and organization of the JCMS system could allow this resource to reach its full potential for improving services statewide. JCJC could help to ensure complete and valid youth background data by taking steps to increase the range and validity of the fields in JCMS. Currently, as noted earlier in the report, there are a few shortcomings regarding the YLS/CMI data (e.g. missing data, definition of assessment date, inconsistent reporting of youths' strengths), and it is our understanding that these are being worked on currently. It is our sense that many issues like these can be addressed by making certain fields "required" in JCMS, conducting training session for JCMS users, developing documentation (to include clear definitions of each JCMS field), and completing regular data quality checks.

JCMS holds a wealth of information that could be tapped to address many questions related to policies and practices in the Commonwealth. Currently, however, getting information out of JCMS and into a form usable by researchers is quite cumbersome. To foster and prepare for partnerships with researchers, JCJC might consider investing resources in staff or consultants who are familiar with the statistical packages frequently used by researchers and the data requirements for conducting statistical analyses.

CONCLUSION #2. As implemented currently statewide, the SPEP™ rating process is producing seemingly valid scores across a variety of services.

As seen in the initial description of the SPEP™ scoring across an array of services, the range of services assessed in Pennsylvania is rather broad. The distributions of SPEP™ total and POP scores indicate a reasonable "spread," suggesting that relevant distinctions are being made among services; the range of the scales is being used. The rating system as applied here appears to be doing an adequate job of differentiating among services, spreading out scoring in a near normal distribution across the entire scoring scale.

The initial analyses presented provide an understandable depiction of the relationship of the SPEP™ total and POP scores and the recidivism outcome (observed minus expected recidivism rate) of the cohorts associated with services at different score values. For the six-month recidivism outcome, it appears that a) services rated at the low end of both scores show dramatically worse recidivism outcomes, b) the services in the middle scoring range show little variability and limited recidivism outcomes, and c) the services at the top of the rating scale perform better as their SPEP™ scores increase. The pattern for the 12-month recidivism outcome is similar for both, but the recidivism outcome does not show a drop off at the high end of either score.

While these patterns of the relationship between the SPEP™ total and POP scores and the recidivism outcomes seem to indicate a strong desired effect for services scoring in the low range and some positive impact on recidivism for high scoring services, the tests of statistical significance show limited support for this "eyeball" interpretation of the association. The only statistically significant test of overall association between the SPEP™ total and POP score and

the recidivism outcome is the one between POP score and recidivism differences for the six-month follow-up period. While an obviously important factor to consider, tests of statistical significance might not be a definitive test of the validity of these scoring systems.

There are several factors that affect the statistical associations obtained between SPEP™ total and POP scores and the recidivism outcomes. First, the observable effect between higher scores and better recidivism outcomes appear to operate most clearly at the low and high ends of the range of scores; there is little shift in recidivism outcomes in the middle range. In the current analyses, however, the vast majority of the services examined received scores near the middle range. When tested statistically, then, this large proportion of cases would contribute heavily to the test of any overall association between the ratings and recidivism. Second, the sample has a relatively low recidivism base rate, making statistical association between any measure (SPEP™ and POP scores included) and the outcome difficult to obtain. As mentioned earlier, the low base rate of recidivism at the six-month and twelve-month follow-up points sets a high bar for any statistical test to achieve, even if the scale of interest shows relatively strong performance. As a result, while a demonstration of statistical significance may have been desirable in these analyses, it would seem unwise to interpret the lack of statistical significance to mean that there is no relationship between the SPEP™ total and POP scores and recidivism reduction. Common sense would tell us that there is a readily apparent, powerful one; just one that is not as simple as we might have expected.

CONCLUSION #3. There are discernable subgroups of scores within the full continuum of SPEP™ and POP Total Scores that are associated with recidivism outcomes.

Analysis were performed to determine if there are underlying subgroups of service cohorts (based on the SPEP™ Total Score) and, if so, whether these score-based groupings are associated with better or worse than expected recidivism outcomes. If such subgroups are identified and they are shown to be associated with better recidivism outcomes, service providers will have a benchmark as to the "amount" of program improvement (defined as the amount of change in the SPEP™ total or POP score) necessary for the service to realize improved recidivism outcomes. Importantly, the analytic approach used allows for the groups to be identified from *within* the data, rather than having an "imposed" cut-off (e.g. below/above 50) that is based on another sample.

Three data-driven, score-based groups emerged from this work and the subgroups with higher SPEP™ total scores are associated with better recidivism rates at 12 months (but not at six months). This is a partial validation of the utility of these subgroups delineated by cut-off scores (presented in the text). It is not a resounding demonstration of a "bright line" marking where services have a markedly different effect. It is, however, some justification to see if setting marks for service providers at these points might have some positive effects. These markers are grounded in the existing data and have some demonstrated validity. They may be more useful to the SPEP™ process and make more sense to providers than the current, seemingly almost arbitrary, distinction made at a SPEP™ Total Score of 50.

CONCLUSION #4: Dimensions of program operations demonstrate varying influence on recidivism outcomes.

Moving beyond the SPEP™ total score, a series of analyses were conducted to look at the relations between each of the five dimensions of program operations (primary service type, theoretical orientation, evidence-base, setting, and amount/quality of service) and recidivism outcomes. The question examined was whether the difference between observed and expected recidivism is related to the "types" identified in the dimension (e.g., CBT vs. mentoring; residential vs. community-based) or the score for certain aspects of service provision (i.e., quality, duration, and dosage). The risk level of the youths was accounted for in all analyses (by virtue of the fact that the outcome is defined as the difference between the observed and expected recidivism rate for the cohort). Each dimension of program operations was tested independently, and the results varied across the dimensions. A few general findings emerged.

First, there does not seem to be any large difference in recidivism outcomes attributable to primary service type. There were no primary service type differences in observed versus expected recidivism at six months and just one difference at 12 months. Also, the difference found was not because one service type looked a lot better than the others, but instead because one service type (remedial education programs) looked so much worse.

Second, the dimensions of theoretical orientation and evidence base showed no strong effects on recidivism. Restorative services were associated with marginally better recidivism outcomes at six months, but these were just short of statistical significance. Evidence-based services showed no significantly better outcomes than locally developed services. While seeming surprising, there is some evidence about effects often being similar for services that are clearly evidence based ("name brand programs") and those developed by local providers (see Lipsey, et al, 2010).

Third, the most consistent and strong findings were those regarding community-based and residential services. The recidivism differences were significantly more favorable for the community-based services at both six and twelve months. Given that the outcome measure for recidivism incorporates the risk level of the cohort of youths served, this result is not the product of residential services just simply working with "more difficult" adolescents.

Fourth, quality, duration, and dosage ratings do, when taken together, have a relationship to the recidivism outcome at 12-months. This relationship is explained by the quality and dosage ratings, not the duration ratings. Given that programs can often have a direct managerial influence on assuring that these aspects of service provision are implemented, efforts to do so seem well grounded. At a policy level, although taxing, it also seems reasonable to keep working with services to monitor and increase their efforts in this regard.

CONCLUSION #5. Overall there was improvement in SPEP™ Total Scores for services having an initial and reassessment rating, and positive change was associated with an improved 6-month recidivism outcome.

A small number of services (n = 38 cohorts) had both an initial and reassessment using the SPEP™. Compared to services with just an initial assessment, services with a SPEP™ reassessment were less likely to be evidence based, more restricted in terms of the primary service types, and evenly split between residential and community-based services. The SPEP™ total reassessment score for these services tended to be skewed toward the higher end of the rating scale. These services showed an overall improvement in SPEP™ Total Score; on average, these services had a 3.16 point increase in the SPEP™ total score. However, 45% of the services had the same or a lower SPEP™ total score upon reassessment.

We also noted substantial variation in the number of days (range 170 to 1,714) between the date of the initial feedback report review and the SPEP™ reassessment of services. It is logical to think that amount of elapsed time to implement an improvement plan will have some bearing on a program's reassessment score (as might other factors such as service type). Although the current, limited data is not sufficient to examine the specific impact of elapsed time between assessments as a factor in the effect of the change scores, this topic seems worthy of further consideration. Such work could help establish guidelines for the optimal time for reassessments to ensure the consistency and validity of comparisons in program improvement.

Analyses of the change in scores over time indicated that the amount of change in the total SPEP™ score was significantly related to recidivism differences at the six-month point, but not at the twelve-month point. The importance of this finding should not be understated. *This is an initial finding indicating that when a service makes improvements to align with the SPEP™, there is a significant reduction in the six-month recidivism outcome for youth who completed the improved service.* The ability to test the actual improvement and recidivism in youths completing the same service at two different times makes a strong argument for the practical utility of the SPEP™ approach as it operates in Pennsylvania.

Because only a small number of service cohorts had both an initial and SPEP™ reassessment, these findings, although very encouraging, should be considered preliminary. It is not clear how representative the services in this sample are of the domain of services across Pennsylvania. It is also unclear if the youths served provide an adequate sampling of the youths involved in juvenile justice statewide. More work needs to be done to replicate these findings on a larger scale as the SPEP™ process moves forward.

Conclusion #6: Looking across the whole set of findings, it appears that efforts to implement the SPEP™ in Pennsylvania are well placed. There is preliminary support for the idea that program improvements in the SPEP™ framework should reduce recidivism by systematically improving service provision.

The findings presented here provide evidence that the implementation of SPEP™ across Pennsylvania is providing valid ratings of dimensions with demonstrated connections to reduced recidivism. In addition, the analysis of services receiving SPEP™ reassessments provide evidence that changes in these scores within a service over time are related to positive recidivism outcomes. Nonetheless, the SPEP™ process would be well served by ongoing monitoring of its activities and investigations examining its operations in terms of recidivism reduction. The results here are preliminary, and there is a need to validate and test additional aspects of the SPEP™ process with larger, more representative samples in the future.

VI. REFERENCES

Baglivio, M.T., Wolff, K.T., Jackowski, K., Chapman, G. Greenwald, M.A. Gomez, K. (2018). Does treatment quality matter? Multilevel examination of the effects of intervention quality on recidivism of adolescents completing long-term juvenile justice residential placement. *Criminology & Public Policy, 17(1),* 147-180

Harris, P.W., Lockwood, B., & Mengers, L. (2009). A CJCA white paper: Defining and measuring recidivism [White paper]. Retrieved from http://www.cjca.net

Hoge, R.D., Andrews, D.A., (2011). Youth Level of Service/Case Management Inventory. User's Manual. New York: Multi-Health Systems, Inc.

Jones, B.L, Nagin, D.S. (2013). A note on a Stata plugin for estimating group-based trajectory models. *Sociological Methods & Research, 42(4)*, 608-613.

Juvenile Court Judges' Commission (2012). *Pennsylvania's Juvenile Justice Systems Enhancement Strategy: Achieving Our Balanced and Restorative Justice Mission Through Evidence-based Policy and Practice*. Harrisburg, PA: Juvenile Court Judges' Commission. Available online: http://www.episcenter.psu.edu/juvenile/jjses

Liberman, A., Hussemann, J (2016). *Implementing the Standardized Program Evaluation Protocol™ to Rate Juvenile Justice Programs: Lessons from OJJDP's Juvenile Justice Report and Reinvestment Initiative.* Washington, DC: Urban Institute

Liberman, A., Hussemann, J (2017). *Local Validation of the SPEP™ Rating of Juvenile Justice Program Effectiveness: A Case Study*. Washington, DC: Urban Institute

Lipsey, M.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. A Report to the Juvenile Justice Services Division, Administrative Office of the Courts, State of Arizona*. Nashville, TN: Center for Evaluation Research and Methodology, Vanderbilt Institute for Public Policy Studies.

Lipsey, M. (2009). The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Victims & Offenders, 4*, 124-147.

Lipsey, M.W. (2018). Effective use of the large body of research on the effectiveness of programs for juvenile offenders and the failure of the model program approach. *Criminology & Public Policy, 17(1)*, 189 – 198.

Lipsey, M.W. & Chapman, G.L. (2017, January*) Standardized Program Evaluation Protocol (SPEP™): A Users Guide.* Nashville, TN: Vanderbilt University Peabody Research Institute.

Lipsey, M.W., Howell, J.D. (2012). A broader view of evidence-based programs reveals more options for state juvenile justice systems. *Criminology & Public Policy, 11(3),* 515-524

Lipsey, M.W., Howell, J.C., Kelly, M.R., Chapman, G., Carver, D. (2010). *Improving the Effectiveness of Juvenile Justice Programs: A New Perspective on Evidence-Based Practice*. Washington, DC, Georgetown Public Policy Institute, Center for Juvenile

Lipsey, M. W., Howell, J.C., Tidd, S.T. (2007). The Standardized Program Evaluation Protocol (SPEP): A practical approach to evaluating and improving juvenile justice programs in North Carolina. Final evaluation report. Nashville, TN: Center for Evaluation Research and Methodology, Vanderbilt Institute for Public Policy Studies.

National Research Council. (2013). *Reforming Juvenile Justice: A Developmental Approach.* Committee on Assessing Juvenile Justice Reform, R.J. Bonnie, R.L. Johnson, B.M. Chemers, J.A. Schuck, Eds. Committee on Law and Justice, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Onifade, E., Davidson, W.S., Campbell, C. (2008). Predicting recidivism in probationers with the Youth Level of Service Case Management Inventory (YLS/CMI). *Criminal Justice and Behavior, 35(4)*, 474-483.

Redpath, D.P., Brandner, J.K. (April, 2010). *SPEP Rating and Relative Recidivism Reduction: An update to the January 2008 report by Dr. Mark Lipsey*. Phoenix, AZ: Arizona Supreme Court, Administrative Office of the Courts, Juvenile Justice Service Division.

Schwalbe C. S. (2007). Risk assessment for juvenile justice: A meta-analysis. Law and Human Behavior, 31, 449-462.

VII. APPENDICES

APPENDIX A

Service and SPEP™ Data provided by the EPISCenter

Variable	Variable label
SPEP™ID_COH	SPEP™ Cohort File ID
SPEP™ID	SPEP™ Service ID
AssmtType	SPEP™ Assessment Occasion Type (Baseline or Reassessment)
AssmtTime	SPEP™ Assessment Timepoint (1, 2, 3…n)
S_ComRes	SPEP™ Service Setting Type (Residential, Community, Mixed)
S_County	Service (or Program) Lead County
S_EBP	Program determined as Locally Developed or Evidence-based Program
S_NumCohort	Total number of Youth in Cohort (Number of Youth Calculated Risk Score)
C_CohortNum	Youth Number in Cohort
C_YID	Youth Juvenile Justice ID Number
C_YID_Ref	Reference Type for Youth Juvenile Justice ID Number
C_Gender	Youth Gender
C_Race	Youth Race
C_Ethnicity	Youth Ethnicity
C_DOB	Youth Date of Birth
C_ComCounty	Youth Committing County
C_StrDate	Youth Service Start Date
C_EndDate	Youth Service End Date
C_SvcWeeks	Youth total number of Service Weeks: Duration
C_SvcHours	Youth total number of Service Hours: Dosage
C_YLSScore	Youth Level of Service Risk Score
C_YLSRiskLevel	Youth Level of Service Risk Level
C_YLSDate	Date Youth Level of Service (YLS) was Finalized
C_YLSOverride	Youth Level of Service Override
C_SvcInter	Service Interruption for Youth
C_Edischarge	Youth Early Discharge
C_Notes	Notes regarding YLS override, service interruption, and/or early discharge
C_YLSFlag	Youth Level of Service Timing: Early, On-time, or Late
P_SvcCat	Primary Therapeutic Category -Restorative, Counseling, Skill-Building
P_SvcGrp	Primary Service Group Type
P_SvcTyp	Primary Service SPEP™ Match Sub - Category (Service Type)
P_SvcSuppTyp	Supplemental Service SPEP™ Type (if applicable)
S_SvcTyp_pts	Primary Service type points
S_SuppSvc_pts	Supplemental Service points
S_Quality_pts	Service Quality points
S_Duration_pts	Service Duration points

S_Dosage_pts	Service Dosage points
S_Risk_pts	Service Risk points
S_Total_pts	Total SPEP™ Score (Total raw Service points earned)
S_MaxPoss_pts	Service Maximum possible points (Denominator for POP score)
S_POPScore	SPEP™ Score Percentage point (Total raw points divided by max possible points)
S_WP1	Written Protocol: 1. Does a manual or written protocol exist that describes the service?
S_WP2	Written Protocol: If a manual or written protocol exists, does it outline in specific detail (process, topic, lesson, session) what should be addressed during service delivery?
S_WP3	Written Protocol: Is the type of youth most appropriate for this service identified in writing? (YLS Risk Factors, Criminogenic Needs, etc.)
S_WP4	Written Protocol: Is there documentation that the manual or written protocol describing the service is used or referenced during service delivery?
S_WP5	Written Protocol: 5. Is there documentation that the manual/written protocol reviewed and updated at predetermined timeframes?
S_WPSum	Written Protocol Sum Score
S_TRN1	Staff Training: Is there a documented minimum education or equivalent experience requirement to deliver the service?
S_TRN2	Staff Training: 7. Is there a written policy that identifies any specialized training or certification required to deliver the service?
S_TRN3	Staff Training: 8. Is there documentation that all staff who deliver the service received the specialized training or certification?
S_TRN4	Staff Training:9. Is there documentation that ongoing or booster training occurs at predetermined timeframes?
S_TRN5	Staff Training: 10. Is there documentation that the supervisor is trained to deliver the service?
S_TRNSum	Staff Training Sum Score
S_SPV1	Staff Supervision: Do supervisors monitor staff delivering the service to assess fidelity and quality?
S_SPV2	Staff Supervision: Is there documentation that the supervisor is monitoring service delivery?
S_SPV3	Staff Supervision: Is there documentation that monitoring occurs at predetermined timeframes?
S_SPV4	Staff Supervision: Do all staff receive written feedback regarding service delivery?
S_SPV5	Staff Supervision: Do performance evaluations, in part, reference fidelity and quality of service delivery?
S_SPVSum	Staff Supervision Sum Score
S_DRF1	Response to Drift: Are there written policies/procedures in place to identify departure from the fidelity and quality of service delivery?
S_DRF2	Response to Drift: If written policies/procedures exist, is there documentation that they are utilized?
S_DRF3	Response to Drift: If written policies/procedures exist, do they include an “if-then” approach or specific corrective action steps to address departure from the fidelity and quality of service delivery?

S_DRF4	Response to Drift: Is data collected on the fidelity and quality of service delivery?
S_DRF5	Response to Drift: If data is collected on the fidelity and quality of service delivery, is it evaluated and used to adapt or improve the service delivery?
S_DRFSum	Response to Drift Sum Score
P_FPPDate	Date of Full Program Profile
P_SICDate	Date of SPEP™ Interview (Categorization)
P_QMIDate	Date of Quality Measures Interview
P_FRRDate	Date of Feedback Report Review
P_PIPDate	Date of Completion of Written Performance Improvement Plan
S_FRCDate	Date of Feedback Report Completion
P_OrgName	Name of Service Provider
P_ProgName	Name of Program (if applicable)
P_SvcName	Name of Service
S_EPISPOC	EPISCenter Lead Staff or Person of Contact
P_JPO	SPEP™ers: JPO (followed by comma), list others if involved with commas in between

APPENDIX B

Youth Background Characteristics and Recidivism Outcome from JCMS

Variable	Time period
DOB (to calculate age at time of service)	
Gender	
Race/Ethnicity	
Base case offense type	Base case for this purpose will be defined as the referral that is closest (but prior to) the SPEP™ service start date
Placement Disposition at base case	Base case for this purpose will be defined as the referral that is closest (but prior to) the SPEP™ service start date
County of base case	County of the juvenile probation department with which the youth was involved for the base case
Age at first referral	From start of JJ involvement
Ever adjudicated delinquent	From start of JJ involvement through SPEP™ service start date
Number of written allegations	From start of JJ involvement through SPEP™ service start date
Number of prior service events with start and end dates and type	From start of JJ involvement through SPEP™ service start date
Residential placements (start and end dates; used to calculate days in placement and number of distinct placement episodes)	From start of JJ involvement through SPEP™ service start date
SVC Indicator	As of start of SPEP™ Service start date
YLS total score and all subscales; YLS assessment date	All from start of JJ involvement to current date (we will select the one closest to the SPEP™ service start date and the one closest to SPEP™ service end date
Responsivity indicators (not mandated, so will not be there for all)	All from start of JJ involvement to current date
Strength indicators (not mandated, so will not be there for all)	All from start of JJ involvement to current date
Over-ride indicator (will indicate when the YLS administrator over-rode the score to classify the youth as H, M or L risk)	All from start of JJ involvement to current date
OUTCOMES	
Offense date	From SPEP™ service end date through the date the data was pulled
Adjudication/conviction date	From SPEP™ service end date through the date the data was pulled
Substantiated Charges/grade	From SPEP™ service end date through the date the data was pulled

Residential Placement dates	From SPEP™ service end date through the date the data was pulled

APPENDIX C

Youth Characteristics by Cohort

Pitt Cohort ID	# IN COH	% Male	% White	Mean Age (mean, SD)	Mean YLS (mean, SD)	Recidivism Rate N yes/N with a valid response, %			
						6	12	18	24
82_c	11	100	9.1	16.18, 1.54	18.45, 4.03	2/11, 18.2%	6/11, 54.5%	8/11, 72.7%	9/11, 81.8%
162_c	38	76.3	21.1	15.26, 1.91	12.05, 5.98	0/38, 0%	3/38, 7.9%	6/38, 15.8%	9/38, 23.7%
81_c	31	71	5	15.97, 1.08	NA	6/31, 19.4%	9/31, 29%	15/31, 48.4%	16/31, 51.6%
1_c	24	66.7	33.3	15.58, 1.35	NA	2/24, 8.3%	6/24, 25%	8/24, 33.3%	9/24, 37.5%
83_c	132	62.9	37.9	15.88, 1.74	15.91, 6.26	11/132, 8.3%	25/132, 18.9%	38/132, 28.8%	44/132, 33.3%
161_c	11	100	18.2	15.45, 1.81	17.00, 5.20	0/11, 0%	2/11, 18.2%	2/11, 18.2%	2/11, 18.2%
80_c	41	70.7	26.8	17.05, 1.73	NA	1/41, 2.4%	7/41, 17.1%	10/41, 24.4%	13/41, 31.7%
2_c	21	66.7	42.9	15.62, 1.72	15.40, 5.04	1/21, 4.8%	2/21, 9.5%	4/21, 19%	5/21, 23.8%
84_c	37	81.1	29.7	16.00, 1.15	13.47, 6.61	8/37, 21.6%	11/37, 29.7%	14/37, 37.8%	15/37, 40.5%
160_c	25	84	20	15.04, 1.54	NA	4/25, 16.0%	6/25, 24%	8/25, 32%	8/25, 32%
79_c	22	81.8	27.3	13.36, 1.53	NA	2/22, 9.1%	2/22, 9.1%	3/22, 13.6%	4/22, 18.2%
3_c	39	92.3	7.7	17.33, .62	16.71, 5.72	4/39, 10.3%	10/39, 25.6%	15/39, 38.5%	15/39, 38.5%
85_c	22	90.9	27.3	15.68, 1.59	NA	2/22, 9.1%	3/22, 13.6%	5/22, 22.7%	6/22, 27.3%

159_c	26	80.8	19.2	15.65, 1.38	18.88, 5.67	1/26, 3.8%	3/26, 11.5%	4/26, 15.4%	8/26, 30.8%
78_c	10	80	70	14.80, 1.87	12.40, 4.74	1/10, 10%	2/10, 20%	3/10, 30%	3/10, 30%
4_c	25	76	88	16.08, 1.53	12.96, 5.91	1/25, 4%	5/25, 20%	7/25, 28%	7/25, 28%
86_c	25	76	88	16.08, 1.53	12.96, 5.91	1/25, 4%	4/25, 16%	4/25, 16%	4/25, 16%
158_c	12	91.7	75	16.25, 1.36	NA	1/12, 8.3%	1/12, 8.3%	2/12, 16.7%	2/12, 16.7%
77_c	115	88.7	69.6	NA	15.49, 5.40	12/115, 10.4%	23/115, 20%	31/115, 27%	39/115, 33.9%
5_c	8	62.5	50	15.38, 1.19	15.88, 6.18	1/8, 12.5%	1/8, 12.5%	1/8, 12.5%	1/8, 12.5%
87_c	11	81.8	9.1	15.91, 0.94	19.00, 6.03	0/11, 0%	1/11, 9.1%	2/11, 18.2%	2/11, 18.2%
157_c	12	83.3	50	15.00, 2.04	13.18, 4.17	2/12, 16.7%	5/12, 41.7%	5/12, 41.7%	7/12, 58.3%
76_c	14	64.3	21.4	18	12.00	1/14, 7.1%	3/14, 21.4%	5/14, 35.7%	6/13, 42.9%
6_c	25	100	12	16.04, 1.54	NA	8/25, 32%	10/25, 40%	15/25, 60%	19/25, 76%
88_c	31	77.4	19.4	16.10, 1.08	NA	8/31, 25.8%	10/31, 32.3%	12/31, 38.7%	13/31, 41.9%
156_c	27	81.5	22.2	16.93, 0.73	NA	4/27, 14.8%	9/27, 33.3%	15/37, 55.6%	17/27, 63.0%
75_c	17	82.4	35.3	15.53, 1.37	NA	4/17, 23.5%	6/17, 35.3%	6/17, 35.3%	7/17, 41.2%
7_c	17	82.4	35.3	15.53, 1.37	NA	4/17, 23.5%	4/17, 23.5%	4/17: 23.5%	5/17, 29.4%
89_c	37	64.9	10.8	15.38, 1.40	NA	4/37, 10.8%	9/37, 24.3%	11/37, 29.7%	11/37, 29.7%
155_c	15	66.7	13.3	15.73, 0.96	NA	2/15, 13.3%	3/15, 20%	3/15, 20%	3/15, 20%

74_c	88	65.9	19.3	15.61, 1.99	13.51, 5.86	9/88, 10.2%	16/88, 18.2%	24/88, 27.3%	26/88, 29.5%
8_c	30	80	10	15.97, 1.63	14.07, 6.07	3/30, 10%	6/30, 20%	9/30, 30%	14/30, 46.7%
90_c	21	85.7	23.8	15.19, 1.69	NA	8/21, 38.1%	11/21, 52.4%	12/21, 57.1%	13/21, 61.9%
154_c	35	74.3	25.7	16.54, 1.09	NA	4/35, 11.4%	10/35, 28.6%	13/35, 37.1%	14/35, 40%
73_c	11	90.9	72.7	15.36, 1.03	16.27, 6.59	0/11, 0%	3/11, 27.3%	3/11, 27.3%	5/11, 45.5%
9_c	40	80	65	14.85, 1.53	14.20, 5.56	5/40, 12.5%	8/40, 20%	10/40, 25%	13/40, 32.5%
91_c	16	50	56.3	14.13, 1.26	13.31, 4.70	2/16, 12.5%	2/16, 12.5%	3/16, 18.8%	5/16, 31.3%
153_c	13	84.6	76.9	15.77, 1.17	20.08, 8.20	0/13, 0%	2/13, 15.4%	5/13, 38.5%	6/9, 66.7%
72_c	8	87.5	75.0	15.88, 2.26	9.63, 5.18	0/8, 0%	0/8, 0%	0/8, 0%	1/8, 12.5%
10_c	27	100	29.6	16.00, 1.21	17.44, 3.91	2/27, 7.4%	4/27, 14.8%	10/27, 37%	12/27, 44.4%
92_c	27	100	29.6	16.00, 1.21	17.44, 3.91	1/27, 3.7%	3/27, 11.1%	8/27, 29.6%	9/27, 33.3%
152_c	27	100	29.6	16.00, 1.21	17.44, 3.91	1/27, 3.7%	2/27, 7.4%	7/27, 25.9%	7/27, 25.9%
71_c	24	100	29.2	16.15, 1.03	17.58, 3.94	3/24, 12.5%	8/24, 33.3%	12/24, 50%	15/24, 62.5%
11_c	27	100	29.6	16.08, 1.10	17.44, 3.91	3/27, 11.1%	5/27, 18.5%	10/26, 38.5%	12/26, 46.2%
93_c	27	81.5	48.1	14.33, 1.88	14.11, 5.13	2/27, 7.4%	5/27, 18.5%	5/27, 18.5%	9/27, 33.3%
151_c	40	82.5	5	15.68, 1.05	14.97, 4.95	14/40, 35%	22/40, 55%	24/40, 60%	26/38, 68.4%
70_c	10	90	70	14.30, 1.83	NA	0/10, 0%	2/10, 20%	3/10, 30%	4/10, 40%

12_c	18	66.7	38.9	15.44, 1.62	15.61, 6.96	0/18, 0%	3/18, 16.7%	3/18, 16.7%	3/11, 27.3%
94_c	20	75	45	15.50, 1.43	16.7, 7.23	1/20, 5.0%	5/20, 25%	5/20, 25%	6/16, 37.5%
150_c	19	100	89.5	16.42, 1.01	16.61, 8.99	1/19, 5.3%	4/19, 21.1%	5/19, 26.3%	8/19, 42.1%
69_c	12	0	8.3	16.08, 0.90	17.92, 5.09	1/12, 8.3%	1/12, 8.3%	1/12, 8.3%	1/7, 14.3%
13_c	19	57.9	0	16.05, 2.84	21.16, 3.92	2/19, 10.5%	5/19, 26.3%	6/19, 31.6%	6/10, 60%
95_c	25	100	56	15.92, 1.41	13.52, 5.47	2/25, 8%	6/25, 24%	9/24, 37.5%	9/11, 81.8%
149_c	54	79.6	24.1	16.15, 1.46	17.56, 6.13	4/54, 7.4%	13/54, 24.1%	17/51, 33.3%	17/23, 73.9%
68_r	10	100	10	16.40, 1.51	14.75, 7.15	0/10, 0%	0/10, 0%	0/10, 0%	0/10, 0%
14_r	10	100	20	16.20, 1.48	14.75, 7.15	1/10, 10%	2/10, 20%	3/10, 30%	4/10, 40%
96_r	10	100	10	16.40, 1.51	14.75, 7.15	0/10, 0%	1/10, 10%	1/10, 10%	1/10, 10%
148_r	10	100	10	16.90, 2.08	14.75, 7.15	2/10, 20%	2/10, 20%	3/10, 30%	5/10, 50%
67_r	11	0	0	15.55, 1.13	19.09, 6.07	0/11, 0%	2/11, 18.2%	2/11, 18.2%	3/11, 27.3%
15_r	10	0	0	15.70, 1.42	21.00, 6.31	1/10, 10%	2/10, 20%	2/10, 20%	4/10, 40%
97_r	10	0	0	15.40, 1.17	18.90, 6.37	0/10, 0%	2/10, 20%	2/10, 20%	3/10, 30%
147_r	11	0	0	15.55, 1.13	18.75, 4.39	0/11, 0%	2/11, 18.2%	2/11, 18.2%	4/11, 36.4%
66_r	12	58.3	8.3	15.92, 1.08	18.75, 4.39	1/12, 8.3%	1/12, 8.3%	3/12, 25%	4/12, 33.3%
16_r	11	63.6	9.1	15.91, 1.14	19.27, 4.20	0/11, 0%	1/11, 9.8%	3/11, 27.3%	5/11, 45.5%

98_r	12	58.3	8.3	15.83, 1.03	18.75, 4.39	1/12, 8.3%	1/12, 8.3%	1/12, 8.3%	1/12, 8.3%
146_r	12	58.3	8.3	16.25, 1.54	18.75, 4.39	2/12, 16.7%	2/12, 16.7%	2/12, 16.7%	2/12, 16.7%
65_r	121	100	29.8	15.79, 1.48	17.56, 5.85	10/121, 8.3%	33/121, 27.3%	48/121, 39.7%	58/121, 47.9%
17_r	30	100	26.7	15.93, 1.34	16.88, 6.25	4/30, 13.3%	7/30, 23.3%	9/30, 30%	9/9, 100%
99_r	10	100	30	17.20, 0.92	NA	1/10, 10%	1/10, 10%	6/10, 60%	7/10, 70%
145_r	7	100	42.9	17.00, 1.15	NA	1/7, 14.3%	1/7, 14.3%	4/7, 57.1%	4/7, 57.1%
64_r	28	100	42.9	16.61, 1.71	NA	3/28, 10.7%	6/28, 21.4%	9/28, 32.1%	12/28, 42.9%
18_r	36	100	52.8	17.22, 1.46	19.69, 5.08	7/36, 19.4%	11/36, 30.6%	14/36, 38.9%	14/26, 53.8%
100_r	20	100	25	16.70, 0.87	NA	3/20, 15%	9/20, 45%	11/20, 55%	13/20, 65%
144_r	20	100	25	16.70, 0.86	NA	1/20, 5%	3/20, 15%	4/20, 20%	4/20, 20%
63_r	12	100	25	16.58, 1.08	NA	0/12, 0%	4/12, 33.3%	6/12, 50%	7/12, 58.3%
19_r	19	100	21.1	16.79, 0.92	NA	1/19, 5.3%	9/19, 47.4%	11/19, 57.9%	12/19, 63.2%
101_r	20	100	25	16.70, 0.86	NA	1/20, 5%	3/20, 15%	4/20, 20%	4/20, 20%
143_r	20	100	25	16.70, 0.86	NA	1/20, 5%	3/20, 15%	4/20, 20%	4/20, 20%
62_r	127	100	33.1	15.92, 1.27	NA	11/127, 8.7%	32/127, 25.2%	42/127, 33.1%	52/127, 40.9%
20_r	106	100	29.2	15.91, 1.19	NA	10/106, 9.4%	26/106, 24.5%	34/106, 32.1%	39/106, 36.8%
102_r	126	100	33.3	15.98, 1.53	17.78, 6.05	13/126, 10.3%	42/126, 33.3%	53/126, 42.1%	67/126, 53.2%

142_r	127	100	33.1	15.87, 1.34	17.70, 6.03	12/127, 9.4%	40/127, 31.5%	51/127, 40.2%	61/127, 48%
61_r	11	100	9.1	16.45, 0.82	NA	0/11, 0%	3/11, 27.3%	4/11, 36.4%	8/11, 72.7%
21_r	50	100	28	15.54, 1.57	17.82, 5.88	6/50, 12%	13/50, 26%	18/50, 36%	25/50, 50%
103_r	72	100	37.5	15.57, 1.52	18.32, 5.28	6/72, 8.3%	18/72, 25%	21/72, 29.2%	27/72, 37.5%
141_r	50	100	28	15.56, 1.49	18.16, 5.90	7/50, 14%	16/50, 32%	21/50, 42%	26/50, 52%
60_r	22	45.5	63.6	15.86, 1.55	16.50, 4.88	3/22, 13.6%	6/22, 27.3%	6/22, 27.3%	8/22, 36.4%
22_r	21	47.6	61.9	15.86, 1.59	16.86, 4.69	3/21, 14.3%	5/21, 23.8%	5/21, 23.8%	5/21, 23.8%
104_r	11	9.1	63.6	16.36, 1.43	15.09, 3.48	0/11, 0%	0/11, 0%	0/11, 0%	0/11, 0%
140_r	36	100	22.2	15.61, 1.98	17.27, 6.28	4/36, 11.1%	10/36, 27.8%	12/36, 33.3%	12/12, 100%
59_r	27	100	40.7	16.07, 1.54	19.09, 8.57	2/27, 7.4%	4/27, 14.8%	6/18, 33%	6/6, 100%
23_r	4	75	0	16.75, 1.71	NA	0/4, 0%	1/4, 25%	2/4, 50%	2/4, 50%
105_r	17	52.9	29.4	16.41, 1.28	22.08, 4.61	4/17, 23.5%	5/17, 29.4%	5/17, 29.4%	6/17, 35.3%
139_r	17	47.1	29.4	16.53, 1.33	21.77, 4.68	1/17, 5.9%	2/17, 11.8%	2/17, 11.8%	4/17, 23.5%
58_r	17	47.1	29.4	16.42, 1.28	22.08, 4.61	2/17, 11.8%	3/17, 17.6%	4/17, 23.5%	5/17, 29.4%
24_r	17	47.1	29.4	16.53, 1.33	21.77, 4.68	0/17, 0%	1/17, 5.9%	2/17, 11.8%	4/17, 23.5%
106_r	38	100	31.6	16.37, 1.87	16.84, 5.86	1/38, 2.6%	6/38, 15.8%	11/38, 28.9%	14/38, 36.8%
138_r	22	100	36.4	16.86, 0.94	15.14, 8.20	2/22, 9.1%	5/22, 22.7%	8/22, 36.4%	8/22, 36.4%

57_r	30	100	40	16.87, 1.25	17.00, 5.61	4/30, 13.3%	7/30, 23.3%	12/30, 40%	14/30, 46.7%
25_r	28	0	78.6	16.89, 1.10	18.63, 7.34	1/28, 3.6%	1/28, 3.6%	3/28, 10.7%	3/28, 10.7%
107_r	12	100	25	16.08, 1.73	16.33, 9.25	2/12, 16.7%	5/12, 41.7%	7/12, 58.3%	7/12, 58.3%
137_r	120	77.5	67.5	16.75, 1.15	19.19, 6.26	14/120, 11.7%	31/120, 25.8%	42/120, 35%	48/120, 40%
56_r	53	100	35.8	16.47, 1.26	17.33, 6.68	8/53, 15.1%	18/53, 34%	25/53, 47.2%	27/53, 50.9%
26_r	22	100	22.7	17.91, 1.48	20.18, 4.33	2/22, 9.1%	9/22, 40.9%	11/22, 50%	13/22, 59.1%
108_r	22	100	22.7	17.91, 1.48	20.18, 4.33	2/22, 9.1%	8/22, 36.4%	11/22, 50%	12/22, 54.5%
136_r	24	0	8.3	16.92, 1.56	19.63, 6.49	1/24, 4.2%	3/24, 12.5%	4/24, 16.7%	6/24, 25%
55_r	22	0	9.1	16.86, 1.25	19.77, 6.31	0/22, 0%	2/22, 9.1%	4/22, 18.2%	5/22, 22.7%
27_r	26	0	7.7	16.65, 1.38	19.23, 6.42	0/26, 0%	3/26, 11.5%	3/26, 11.5%	5/26, 19.2%
109_r	26	0	7.7	16.63, 1.44	19.71, 6.38	1/26, 3.8%	5/26, 19.2%	5/26, 19.2%	7/26, 26.9%
135_r	13	100	7.7	17.54, 1.81	22.92, 5.28	3/13, 23.1%	5/13, 38.5%	6/13, 46.2%	6/13, 46.2%
54_r	46	100	15.2	17.96, 1.67	20.85, 4.56	6/46, 13%	16/46, 34.8%	22/46, 47.8%	22/46, 47.8%
28_r	26	0	7.7	16.69, 1.35	19.23, 6.42	0/26, 0%	3/26, 11.5%	3/26, 11.5%	5/26, 19.2%
110_r	10	100	90	16.60, 1.174	19.50, 7.09	0/10, 0%	0/10, 0%	1/10, 10%	1/10, 10%
134_r	16	100	6.3	16.69, 0.87	23.25, 4.11	5/16, 31.3%	7/16, 43.8%	9/16, 56.3%	10/14, 71.4%
53_r	12	100	41.7	17.25, 0.87	22.50, 4.66	1/12, 8.3%	6/12, 50%	6/12, 50%	6/12, 50%

29_r	14	100	7.1	17.50, 1.79	22.57, 4.64	1/14, 7.1%	6/14, 42.9%	8/14, 57.1%	9/12, 75%
111_r	16	100	6.3	17.88, 0.89	20.00, 3.61	3/16, 18.8%	5/16, 31.3%	7/16, 43.8%	8/15, 53.3%
133_r	12	100	16.7	17.58, 1.62	18.50, 5.49	1/12, 8.3%	3/12, 25%	4/12, 33.3%	4/10, 40%
52_r	11	100	18.2	15.73, 1.42	23.73, 4.92	2/11, 18.2%	4/11, 36.4%	5/11, 45.5%	6/11, 54.5%
30_r	14	100	28.6	14.43, 0.76	22.50, 4.75	2/14, 14.3%	8/14, 57.1%	9/14, 64.3%	9/14, 64.3%
112_r	10	100	90	16.60, 1.17	20.70, 6.68	0/10, 0%	2/10, 20%	4/10, 40%	4/10, 40%
132_r	13	100	7.7	17.54, 1.81	22.92, 5.28	0/13, 0%	2/13, 15.4%	3/13, 23.1%	3/13, 23.1%
51_r	35	100	17.1	17.71, 1.564	21.20, 4.82	4/35, 11.4%	12/35, 34.3%	15/35, 42.9%	16/35, 45.7%
31_r	21	100	23.8	17.71, 1.42	20.57, 4.02	2/21, 9.5%	6/21, 28.6%	7/21, 33.3%	7/21, 33.3%
113_r	14	71.4	50	15.50, 2.03	17.50, 5.45	0/14, 0%	1/14, 7.1%	3/14, 21.4%	4/11, 36.4%
131_r	12	91.7	66.7	16.67, 1.87	21.33, 6.05	1/12, 8.3%	4/12, 33.3%	5/12, 41.7%	5/7, 71.4%
50_r	12	66.7	83.3	15.08, 1.38	11.27, 4.96	0/12, 0%	1/12, 8.3%	1/11, 9.1%	3/8, 37.5%
32_r	14	71.4	35.7	16.64, 0.84	17.43, 6.76	1/14, 7.1%	2/12, 16.7%	2/4, 50%	2/2, 100%
114_r	34	100	29.4	16.03, 1.45	18.74, 6.94	1/34, 2.9%	7/34, 20.6%	12/34, 35.3%	12/14, 85.7%
130_r	35	100	31.4	15.89, 1.45	17.80, 6.19	2/35, 5.7%	8/33, 24.2%	15/33, 45.5%	15/15, 100%
49_r	10	100	80	16.30, 1.49	18.10, 7.48	0/10, 0%	1/10, 10%	1/4, 25%	1/1, 100%
33_r	14	100	28.6	15.57, 1.50	15.57, 4.77	0/14, 0%	3/14, 21.4%	4/11, 36.4%	4/4, 100%

115_r	40	100	40	16.10, 1.22	18.23, 5.82	3/40, 7.5%	7/34, 20.6%	7/8, 87.5%	7/7, 100%
129_r	29	100	34.5	16.45, 1.02	18.21, 4.84	2/29, 6.9%	5/29, 17.2%	11/29, 37.9%	12/29, 41.4%
48_r	29	100	34.5	16.38, 0.94	18.21, 4.84	2/29, 6.9%	7/29, 24.1%	10/29, 34.5%	13/29 44.8%
34_r	65	98.5	24.6	16.51, 1.13	17.65, 5.29	9/65, 13.8%	22/65, 33.8%	29/65, 44.6%	31/61, 50.8%
116_r	42	100	28.6	16.40, 0.96	18.83, 4.88	4/42, 9.5%	9/42, 21.4%	16/42, 38.1%	19/42, 45.2%
128_r	42	100	28.6	16.40, 0.96	18.83, 4.88	4/42, 9.5%	11/42, 26.2%	16/42, 38.1%	18/42, 42.9%
47_r	20	75	50	15.45, 1.43	16.70, 7.23	0/20, 0%	2/20, 10%	2/20, 10%	2/16, 12.5%
35_r	20	75	50	15.45, 1.43	16.70, 7.23	0/20, 0%	2/20, 10%	2/20, 10%	2/16, 12.5%
117_r	25	48	56	16.80, 1.26	17.32, 6.20	2/25, 8%	3/25, 12%	3/21, 14.3%	4/7, 57.1%
127_r	44	65.9	56.8	16.91, 1.64	19.23, 6.12	2/44, 4.5%	5/42, 11.9%	7/27, 25.9%	7/16, 43.8%
46_r	13	84.6	69.2	17.15, 1.28	14.62, 5.85	1/13, 7.7%	3/12, 25%	3/3, 100%	3/3, 100%
36_r	42	69	57.1	17.00, 1.50	18.79, 6.22	1/42, 2.4%	3/40, 7.5%	4/24, 16.7%	4/13, 30.8%
118_r	44	68.2	56.8	17.00, 1.52	18.79, 6.21	1/44, 2.3%	4/42, 9.5%	5/27, 18.5%	5/15, 33.3
126_r	91	100	9.9	16.23, 1.10	16.52, 7.24	10/91, 11%	22/91, 24.2%	26/78, 33.3%	26/37, 70.3%
45_r	90	100	8.9	16.24, 1.10	16.45, 7.24	7/90, 7.8%	14/90, 15.6%	24/90, 26.7%	27/81, 33.3%
37_r	146	100	8.9	16.62, 0.98	16.89, 6.96	11/146, 7.5%	29/146, 19.9%	39/143, 27.3%	44/54, 81.5%
119_r	90	100	8.9	16.23, 1.09	16.51, 7.23	8/90, 8.9%	17/90, 18.9%	27/90, 30%	NA

125_r	88	100	11.4	16.14, 1.11	16.88, 7.22	5/88, 5.7%	16/88, 18.2%	23/86, 26.7%	27/75, 36%
44_r	128	100	14.8	16.31, 1.18	17.56, 7.22	25/128, 19.5%	42/128, 32.8%	48/110, 43.6%	48/52, 92.3%
38_r	11	100	9.1	16.91, 1.04	14.09, 7.58	2/11, 18.2%	2/11, 18.2%	2/6, 33.3%	2/2, 100%
120_r	14	100	28.6	16.57, 1.16	19.62, 7.03	3/14, 21.4%	5/14, 35.7%	5/10, 50%	5/5, 100%
124_r	18	100	0	16.61, 1.04	16.67, 6.70	3/18, 16.7%	4/18, 22.2%	5/11, 45.5%	5/5, 100%
43_r	28	67.9	42.9	15.61, 1.57	15.93, 7.54	0/28, 0%	3/28, 10.7%	4/14, 28.6%	5/5, 100%
39_r	24	4.2	20.8	15.63, 1.50	18.29, 6.30	0/24, 0%	2/24, 8.3%	2/6, 33.3%	2/2, 100%
121_r	24	4.2	20.8	15.63, 1.50	18.29, 6.30	0/24, 0%	1/24, 4.2%	1/6, 16.7%	1/1, 100%
123_r	25	4.0	24	15.56, 1.50	18.41, 6.18	0/25, 0%	1/25, 4%	1/6, 16.7%	1/1, 100%
42_r	25	4	24	15.56, 1.50	18.41, 6.18	0/25, 0%	1/25, 4%	1/6, 16.7%	1/1, 100%
40_r	24	4.2	20.8	15.63, 1.50	18.29, 6.30	0/24, 0%	1/24, 4.2%	1/6, 16.7%	1/1, 100%
122_r	22	100	27.3	18.00, 0.93	22.27, 7.86	2/22, 9.1%	4/22, 18.2%	4/12, 33.3%	4/4, 100%
41_r	54	100	24.1	17.52, 1.51	22.39, 7.82	5/54, 9.3%	9/36, 25%	9/19, 47.4%	9/9, 100%

APPENDIX D

SPEP Scores by Cohort

Pitt Cohort ID	Type	SPEP Scores						
	1=Individual counseling 2=job related training 3=remedial academic 4=rest,comm service 5=challenge 6=family counseling 7=mediation 8=mixed counseling 9=social skills 10=behavioral, contingency management 11=group counseling 12=mentoring 13=CBT	Primary Service Type points	Supplemental Service points	Service Quality points	Service Duration points	Service Dosage points	Service Risk points	Total SPEP Raw score
82_c	13	30	5	10	2	0	20	67
162_c	4	10	5	20	4	0	5	44
81_c	13	30	5	10	6	8	13	72
1_c	13	30	5	20	6	6	17	84
83_c	6	15	5	20	4	4	10	58
161_c	6	15	5	20	2	2	15	59
80_c	2	5	0	10	2	0	15	32
2_c	13	30	5	10	6	4	10	65
84_c	1	5	5	10	0	0	5	25
160_c	9	15	5	10	2	4	13	49
79_c	12	25	5	5	2	0	2	39
3_c	9	15	5	10	0	6	15	51
85_c	8	15	0	20	2	6	25	68
159_c	6	15	5	20	6	6	20	72

78_c	12	25	0	10	6	0	7	48
4_c	6	15	5	10	6	0	5	41
86_c	12	25	0	10	2	0	5	42
158_c	11	25	5	20	2	6	10	68
77_c	9	15	5	20	2	8	10	60
5_c	6	15	5	20	8	6	12	66
87_c	9	15	5	5	4	2	18	49
157_c	8	15	5	20	0	4	7	51
76_c	6	15	5	20	6	6	10	62
6_c	12	25	0	20	0	0	10	55
88_c	13	30	5	20	0	0	10	65
156_c	3	10	5	10	0	2	10	37
75_c	6	15	5	20	2	0	7	49
7_c	1	5	5	20	0	4	7	41
89_c	12	25	0	10	4	2	18	59
155_c	6	15	5	10	2	2	22	56
74_c	13	30	5	20	6	0	7	68
8_c	13	30	5	20	6	6	10	77
90_c	6	15	5	20	4	4	15	63
154_c	12	25	0	20	0	0	10	55
73_c	6	15	5	20	6	4	10	60
9_c	6	15	5	20	8	6	7	61
91_c	6	15	5	20	8	8	10	66
153_c	2	5	0	20	2	6	23	56
72_c	8	15	0	20	2	0	5	42
10_c	10	25	5	20	4	6	15	75
92_c	1	5	5	10	2	0	15	37,0
152_c	4	10	5	10	6	0	15	46
71_c	13	30	5	10	2	0	15	62
11_c	13	30	5	10	6	4	15	70
93_c	6	15	5	10	4	6	10	50

151_c	3	10	5	20	0	8	10	53
70_c	6	15	5	20	8	8	18	74
12_c	8	15	0	20	6	8	15	64
94_c	13	30	5	20	10	8	13	86
150_c	4	10	5	10	0	8	12	45
69_c	9	15	5	20	2	8	13	63
13_c	13	30	5	20	6	6	25	92
95_c	12	25	0	5	0	0	7	37
149_c	12	25	0	20	2	0	20	67
68_r	1	5	5	20	4	2	23	59
14_r	13	30	5	20	6	6	22	89
96_r	13	30	5	20	4	6	23	88
148_r	10	25	0	20	2	4	23	74
67_r	1	5	5	20	2	2	20	54
15_r	13	30	5	20	8	8	25	96
97_r	13	30	5	20	6	8	22	91
147_r	10	25	0	20	2	8	20	75
66_r	1	5	5	20	6	4	15	55
16_r	13	30	5	20	10	10	20	95
98_r	13	30	5	20	10	10	15	90,0
146_r	10	25	0	20	6	10	15,0	76
65_r	13	30	5	20	6	6	20	87
17_r	11	25	5	20	0	0	15	65
99_r	2	5	5	20	2	6	25	63
145_r	13	30	5	20	10	10	25	100
64_r	3	10	5	20	0	8	25	68
18_r	13	30	5	20	8	8	25	96
100_r	10	25	5	5	6	8	23	72
144_r	3	10	5	20	2	10	23	70
63_r	13	30	5	20	10	10	25	100
19_r	7	15	0	10	10	2	23	60

101_r	11	25	5	10	6	0	23	69
143_r	1	5	5	5	4	2	23	44
62_r	1	5	5	20	0	0	15	45
20_r	13	30	5	20	6	4	15	80
102_r	3	10	5	20	0	8	15	58,0
142_r	11	25	5	20	0	0	15	65
61_r	13	30	5	20	10	10	25	100
21_r	10	25	5	20	0	10	17	77
103_r	5	15	5	20	10	8	15	73
141_r	11	25	5	10	0	4	17	61
60_r	1	5	5	20	4	2	15	51
22_r	11	25	5	20	4	0	15	69
104_r	11	25	5	20	4	0	12	66
140_r	1	5	5	10	4	2	17	43
59_r	13	30	5	10	6	4	25	80
23_r	9	15	5	10	6	10	10	56
105_r	13	30	5	20	10	10	25	100
139_r	13	30	5	20	10	10	25	100
58_r	1	5	5	20	6	4	25	65
24_r	10	25	5	20	6	10	25	91
106_r	13	30	5	10	6	6	15	72
138_r	5	15	5	10	4	0	10	44
57_r	2	5	0	10	8	0	13	36
25_r	11	25	5	5	0	0	15	50
107_r	13	30	5	10	8	0	20	73
137_r	11	25	5	5	2	0	18	55
56_r	11	25	5	5	2	0	18	55
26_r	9	15	5	20	6	2	22	70
108_r	13	30	5	20	6	6	22	89
136_r	13	30	5	20	6	4	15	80
55_r	11	25	5	10	0	0	15	55

27_r	1	5	5	10	8	6	17	51
109_r	4	10	5	10	8	6	17	56
135_r	13	30	5	20	8	4	25	92
54_r	3	10	5	20	4	8	25	72,0
28_r	3	10	5	20	6	8	17	66
110_r	13	30	5	20	10	6	25	96
134_r	13	30	5	10	10	10	25	90
53_r	13	30	5	10	10	6	25	86
29_r	13	30	5	10	10	6	25	86
111_r	13	30	5	10	10	0	22	77
133_r	13	30	5	20	10	0	20	85
52_r	2	5	5	10	8	8	25	61
30_r	13	30	5	20	10	10	25	100
112_r	13	30	5	20	10	10	25	100
132_r	13	30	5	20	8	4	25	92
51_r	1	5	5	20	6	4	25	65
31_r	2	5	5	10	8	0	22	50
113_r	13	30	5	20	6	10	17	88
131_r	13	15	5	20	10	6	25	81
50_r	4	10	5	20	6	0	15	56
32_r	13	30	5	20	6	6	15	82
114_r	4	10	5	5	4	0	23	47
130_r	11	25	5	10	0	0	20	60
49_r	13	30	5	5	8	0	15	63
33_r	3	10	0	20	2	8	10	50
115_r	5	15	5	20	10	10	15	75
129_r	13	30	5	20	10	10	15	90
48_r	13	30	5	10	8	0	17	70
34_r	3	10	0	20	6	8	17	61
116_r	11	25	5	10	0	0	20	60
128_r	11	25	5	10	0	2	20	62

47_r	9	15	5	20	8	6	13	67
35_r	1	5	5	10	6	4	13	43
117_r	3	10	0	20	0	0	20	50
127_r	1	5	5	10	2	0	22	44
46_r	11	25	5	20	0	0	10	60
36_r	2	5	0	10	2	0	18	35
118_r	2	10	5	5	8	2	20	50
126_r	13	30	5	10	8	0	15	68
45_r	13	30	5	5	2	0	15	57
37_r	9	15	5	10	0	0	15	45
119_r	11	25	5	10	0	0	15	55
125_r	9	15	5	10	6	4	15	55
44_r	4	10	5	10	8	4	18	55
38_r	2	5	0	10	0	0	8	23
120_r	2	5	0	5	4	0	20	34
124_r	2	5	0	10	4	0	10	29
43_r	11	10	5	5	2	0	12	34
39_r	1	25	5	5	2	0	23	60
121_r	4	5	5	20	2	0	23	55
123_r	4	10	5	10	8	0	23	56
42_r	5	15	5	20	10	2	23	75
40_r	13	30	5	5	4	0	23	67
122_r	13	30	5	10	10	10	25	90
41_r	9	15	5	20	8	8	25	81

APPENDIX E

Initial and Reassessment SPEP™ total and POP score

	Pitt Cohort ID	Initial SPEP Total	Reassessment SPEP Total	SPEP Total Score Change	Initial POP Score	Reassessment POP Score	POP Score Change
1	82_c	67	86	19	67	86	19
2	81_c	72	72	0	72	72	0
3	1_c	84	78	-6	84	78	-6
4	83_c	58	58	0	69	69	0
5	2_c	65	79	14	65	79	14
6	159_c	72	74	2	85	88	3
7	78_c	48	68	20	61	72	11
8	4_c	41	50	9	49	59	10
9	86_c	42	57	15	45	60	15
10	158_c	68	71	3	72	75	3
11	77_c	60	69	9	71	82	11
12	5_c	66	68	2	78	80	2
13	87_c	49	59	10	58	70	12
14	76_c	62	57	-5	73	68	-5
15	6_c	55	72	17	58	76	18
16	156_c	37	50	13	47	63	16
17	74_c	68	79	11	68	79	11
18	8_c	77	80	3	77	80	3
19	90_c	63	67	4	75	79	4
20	68_r	59	56	-3	79	75	-4
21	14_r	89	89	0	89	89	0
22	96_r	88	85	-3	88	85	-3
23	148_r	74	80	6	78	85	7
24	67_r	54	55	1	72	74	2
25	15_r	96	84	-12	96	84	-12
26	97_r	91	86	-5	91	86	-5
27	147_r	75	79	4	79	84	5
28	66_r	55	62	7	74	83	9
29	16_r	95	93	-2	95	93	-2
30	98_r	90	89	-1	90	89	-1
31	146_r	76	84	8	80	89	9
32	65_r	87	85	-2	87	85	-2
33	99_r	63	54	-9	84	72	-12
34	145_r	100	91	-9	100	91	-9

35	64_r	68	67	-1	85	84	-1
36	21_r	77	72	-5	82	76	-6
37	103_r	73	70	-3	86	83	-3
38	141_r	61	70	9	65	74	9