A Revalidation of the Standardized Program Evaluation Protocol in Pennsylvania

Final Report

Samuel W. Hawes

Edward P. Mulvey

Carol A. Schubert

2025-06-06

Acknowledgments: We would like to thank the staff at the Juvenile Court Judges Commission and EPIS at Penn State University for their careful attention to, and knowledge of, the data, as well as their collaborative spirit and efforts throughout the project.

Executive Summary

Purpose

This report presents findings from the 2024 revalidation of the Standardized Program Evaluation Protocol (SPEP™) across Pennsylvania. The evaluation aimed to determine whether SPEP™ and Program Optimization Percentage (POP) scores continue to predict reduced recidivism outcomes among youth involved in juvenile justice services. The study analyzed 113 service cohorts encompassing 3,459 youth who received services between 2018 and 2023.

Key Findings

- *Predictive Validity Maintained:* Higher SPEP™ and POP scores were associated with lower recidivism rates at 24 months, supporting prior validation results, though effect sizes were modestly smaller than in the original study.
- *Overall Quality Improvement:* The majority of programs now achieve 'high quality' ratings, narrowing differences between groups and suggesting system-wide service improvements.
- *Service Quality Over Type or Amount*: Service quality (adherence to best practices) was a more direct and consistent predictor of positive outcomes than amount of time in a program (duration), number of sessions (dosage), or service classification.
- *Program Change Linked to Outcome Gains:* Programs that improved their SPEP™ scores across reassessments saw greater recidivism reductions over time, underscoring the value of the SPEP™ system for ongoing Continuous Quality Improvement (CQI).
- *Youth Risk/Needs Matching:* Services were broadly beneficial across youth profiles, with particularly strong effects among higher-risk and higher-need youth.

Conclusions

Findings affirm that the SPEP™ remains a valuable framework for improving service quality and outcomes in Pennsylvania's juvenile justice system. The convergence of higher quality ratings across services suggests that broader system reforms — such as the Juvenile Justice System Enhancement Strategy (JJSES) — are achieving intended effects. Continued attention to aligning services with individualized youth risk/needs profiles represents a promising direction for further reducing recidivism and enhancing service effectiveness.

Recommendations

- *Enhance Individualized Service Matching:* Leverage available risk/needs assessment tools (e.g., YLS/CMI™) to better align youth with services tailored to their profiles. Expanded, focused research efforts on this initiative are needed.
- *Expand Periodic Reassessment:* Continue periodic SPEP™ reassessments to support dynamic improvement and reinforce the culture of continuous quality enhancement.
- *Sustain Infrastructure and Training:* Maintain investment in SPEP™ infrastructure, staff training, and technical assistance to preserve fidelity and to support emerging best practices.

Abstract

This report presents findings from the 2024 revalidation of the Standardized Program Evaluation Protocol (SPEP™) in Pennsylvania. Drawing on data from 113 service cohorts (3,459 youth) providing services between 2018 and 2023, it examines how SPEP™ and Program Optimization Percentage (POP) scores relate to 12- and 24-month recidivism outcomes and whether previous low-medium-high scoring cut-offs remain predictive. Results show that while higher SPEP™ and POP scores generally coincide with lower recidivism, these effects are more evident at 24 months. Service quality exhibits a more direct influence than either dosage or duration of services. Improvements in SPEP™ scores over time correlate with better-than-expected outcomes, suggesting the value of periodic re-evaluation and continued improvement. However, a restricted sample, potential ceiling effects, and pandemic disruptions limit certain findings. Overall, the revalidation affirms SPEP™ as a valuable evaluative tool for juvenile justice services and highlights the need for ongoing refinements and broader datasets to ensure its continued effectiveness as a tool for continuous quality improvement.

1 Introduction

This report summarizes findings from the 2024 revalidation of the Standardized Program Evaluation Protocol (SPEP™) in Pennsylvania, building on the initial validation study conducted by members of the same research team in 2019 (Mulvey, Schubert, Jones & Hawes, 2020). SPEP™ serves as a critical framework for aligning juvenile justice services with evidence-based practices. This study examines the continued effectiveness of SPEP™ in improving youth outcomes and ensuring that its scoring system remains valid.

The SPEP™ process evaluates juvenile justice programs across several dimensions (e.g., service type, quality, dosage, and duration of service) and their collective impact on reducing recidivism (Lipsey & Chapman, 2017). It assigns scores on a scale of 0 to 100 based on adherence to evidence-based standards, with higher scores reflecting stronger alignment with practices shown to reduce recidivism in the research literature (Lipsey, 2020). These scores are paired with the Program Optimization Percentage (POP), a metric contextualized by the program's optimal potential based on its service type. The POP score provides a comparison of how well the program is performing relative to both research-based standards and its inherent potential (Lieberman, 2017). The combined metrics provide a comprehensive view of how well a program performs relative to research-based standards and the program's chances of success given its approach.

Research in other states, such as Arizona and North Carolina, supports the validity of SPEP™ scores as predictors of positive outcomes. These studies demonstrate that higher SPEP™ scores are associated with reduced recidivism rates among justice-involved youth (Redpath & Brandner, 2017; Liberman, 2017). However, while these findings highlight the utility of SPEP™ as an evaluative tool, recent literature on its long-term effectiveness in system-wide applications remains sparse, particularly in terms of its capacity to sustain improvements over time (Lipsey, 2020).

This revalidation study aims to examine the continued effectiveness of SPEP™ in Pennsylvania, focusing on whether its scoring system remains a reliable tool for improving youth recidivism outcomes. By analyzing updated data and assessing SPEP™ scores for their relationship to program effectiveness, this study provides insights into both the strengths and limitations of the SPEP™ framework. In the sections that follow, we briefly review the findings from the initial validation before presenting the methodology, results, and implications of the 2024 revalidation effort.

1.1 Initial SPEP™ Validation Summary

The 2019 validation of SPEP™ in Pennsylvania provided a robust evaluation of its implementation, offering insights into the effects of its different components. The study integrated comprehensive data sources, including SPEP™-specific data from the Penn State EPIS (EPIS) and state-level juvenile justice records from the Juvenile Court Judges' Commission (JCJC), to address critical questions about its impact on juvenile recidivism and service quality.

There were several key findings from the initial SPEP™ assessment. These were:

- **Validity of scores**:
  The study confirmed that SPEP™ ratings effectively differentiated the performance of services based on their adherence to evidence-based practices. These ratings

were strongly associated with recidivism outcomes, validating their utility in juvenile justice settings. Programs with higher SPEP™ and Program Optimization Percentage (POP) scores consistently demonstrated better-than-expected recidivism outcomes. Conversely, lower scores correlated with worse outcomes, providing a clear performance gradient. Programs scoring within the middle range displayed weaker associations with recidivism reduction.

- **Useful score-based subgroups**:
  Data-driven subgroup analyses revealed three distinct performance levels (low: 23–43; middle: 44–77; high: 80–100 on the SPEP™ scoring scale) that correlated with recidivism benchmarks. These groupings offered actionable targets for improving service delivery and outcomes.

- **Effects of specific service dimensions**:
  The evaluation examined multiple dimensions of service (e.g., type, theoretical orientation, evidence-based or not, setting, as well as quality, dosage (number of contact hours) and duration (number of weeks in service). Service setting, quality, and dosage emerged as the most predictive of positive recidivism outcomes (duration was not significant in the models). Theoretical orientation and evidence-based programming showed more modest effects.

- **Reassessment validation of score changes:**
  Improvements in SPEP™ scores over time for the same service were linked to reductions in six-month recidivism rates, underscoring the protocol's potential value as a continuous quality improvement tool. These analyses, however, were limited in their generalizability as the sample of services receiving a second SPEP™ evaluation was constrained to a small subset of the larger validation sample (n = 42).

- **Methods for integrating available data sets:**
  The 2019 study addressed numerous challenges with consolidating existing data sets and creating relevant variables for consideration. This required extensive collaboration with EPIS and JCJC researchers to understand the meaning and soundness of specific variables in their respective data systems. This collaboration was essential to ensure the reliability and validity of findings. For example, the study highlighted the necessity of aligning service duration and dosage with SPEP™ benchmarks to maximize recidivism reduction potential.

1.2. Extension to the 2024 Revalidation Study

Building on these findings, the 2024 revalidation study delves deeper into the long-term effectiveness of SPEP™. Notably, it integrates lessons from prior analyses, such as ensuring adequate sample sizes for subgroup assessments and refining methodologies for calculating expected recidivism rates. The 2019 study and this subsequent revalidation provide a solid empirical basis demonstrating the SPEP™ protocol's validity and scalability. By situating SPEP™ within Pennsylvania's broader Juvenile Justice Systems Enhancement Strategy (JJSES), a compelling case can be made for this practice to serve as a cornerstone of evidence-based juvenile justice reform. The implications of the revalidation study findings will be elaborated later in this document.

1.3 Foci of the revalidation study

In consultation with the Juvenile Justice Systems Enhancement Strategy (JJSES) Leadership Team, SPEP™ Advisory Group, EPIS, and JCJC personnel, five primary questions were identified as focal points for this revalidation.

1. Is the SPEP™ score predictive of youth recidivism outcomes?

2. Can the low, medium, and high Basic/POP subgroup scoring ranges be revalidated as predictors of observed vs. expected recidivism? Are the ranges still valid? And are there optimal SPEP™ score ranges that predict outcomes above and beyond youth risk scores (YLS/CMI)?

3. Are certain SPEP™ components (e.g., service type, quality, dosage) more powerful than others in reducing recidivism, and how do they align with youth risk/needs profiles?

4. How are changes in SPEP™ scores from baseline to reassessment related to recidivism outcomes?

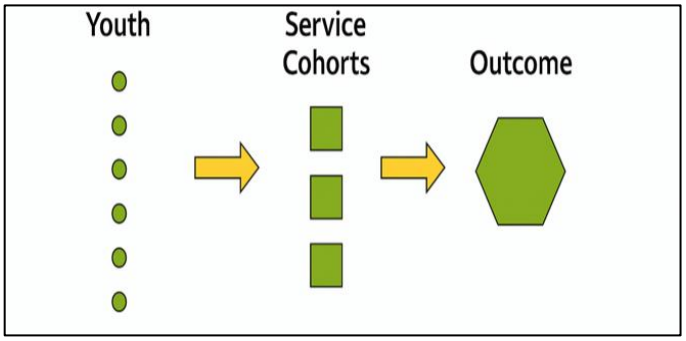
5. What role does matching youth criminogenic needs to appropriate services play in recidivism outcomes?

These questions guide the analyses presented here.

1.4 The basic research design.

The basic analytic design for the revalidation mirrored that of the initial validation. It is shown below in Figure 1. The primary unit of analysis for the SPEP™ process is a single service. Each service is evaluated during a specified period, with scores assigned to that service based on established guidelines and include separate scores for aspects of service provision, such as quality and amount (Lipsey & Chapman, 2017). The scores assigned to each service represent the characteristics of that service received by the cohort of individual youth involved with the service concurrent to the SPEP™ ratings.

Figure 1: Basic Design



The impact of the SPEP™-evaluated service on the outcomes for each youth serves as the foundation for assessing the overall outcomes of the service cohort. If a high proportion of the adolescents in a particular cohort recidivate, then that cohort will have a higher calculated recidivism rate.

The revalidation analyzed data from 113 service cohorts. For each one, there are SPEP™ scores from the assessment of that service and outcomes for that service determined by the aggregation of the individual outcomes of the adolescents comprising that service cohort. This approach necessitates merging data from multiple sources. SPEP™ service scores and youth attendance data were obtained from EPIS. Youth risk levels, background characteristics, and outcomes were drawn from JCJC records.

1.5 Differences Between Initial Validation and Revalidation

The revalidation study has several features that differ from the initial validation, some of which enhance its scope. These differences reflect changes in the operational environment, the availability of data, and adjustments in the study design:

1. **Time Frame**: The revalidation examines services delivered between 2018 and 2023, compared to the 2010–2018 period analyzed in the initial validation. This expanded time frame captures more recent program implementation trends and reflects the impact of continued juvenile justice enhancement strategies. However, external factors such as the COVID-19 pandemic and changes in EPIS staff overseeing SPEP™ evaluations may have introduced unmeasured differences in service delivery or data quality.

2. **Sample Size and Demographic Characteristics**: The revalidation includes fewer youth and service cohorts than the initial validation, which had a broader dataset. Despite this reduction, the current study benefits from improved data availability, particularly the inclusion of Youth Level of Service/Case Management Inventory (YLS/CMI) scores. These scores provide a consistent measure of risk levels across the sample, decreasing the need for proxy calculations based solely on background characteristics, as required in the initial validation. The youth included in the initial and revalidation samples are nearly identical in age and gender. The revalidation sample has a more diverse race/ethnicity composition, with 56% considered minorities as compared to 65% of the initial validation sample.

3. **Outcome Period**: The primary outcome period in the revalidation focuses on 24-month recidivism outcomes, aligning with reporting standards used by the Juvenile Court Judges' Commission (JCJC). A secondary analysis was conducted for 12-month outcomes to address cases where 24-month data were unavailable. This dual approach balances methodological rigor with data completeness, while recognizing that recidivism outcomes may change over time due to intervening events in youths' lives.

4. **Reassessments**: The revalidation includes a significantly larger number of services with reassessment data compared to the initial validation, providing a more solid basis for analyzing changes in SPEP™ scores over time. These reassessments allow for a deeper exploration of the relationship between score improvements and recidivism outcomes.

In summary, this revalidation study builds upon the strengths of the initial evaluation by integrating new data and extending the study design. These enhancements reaffirm the initial findings, but also the durability of the SPEP™ process in improving youth outcomes. By extending the outcome evaluation period, leveraging higher-quality data, and incorporating a larger number of reassessments, the study provides a more robust evidence base regarding SPEP™ implementation in Pennsylvania.

2 Methodology

This section outlines the data sources, sample description, analytic techniques, and limitations of the revalidation study.

2.1 Data Sources

The revalidation integrates data from two primary sources:

1. **EPIS Data**: Managed by the Evidence-based Prevention and Intervention Support Center (EPIS) at Pennsylvania State University, this dataset includes detailed SPEP™ service scores and youth attendance information for each service cohort. The EPIS played a pivotal role not only in providing well-organized datasets but also in coordinating SPEP™ evaluations across Pennsylvania, offering technical assistance, and standardizing the data collection processes for program assessments.

2. **JCJC Data**: Demographic, background, and outcome data, including Youth Level of Service/Case Management Inventory (YLS/CMI™) scores and recidivism outcomes were provided by the Juvenile Court Judges' Commission (JCJC). These datasets were derived from the Juvenile Case Management System (JCMS), which offers a rich source of administrative records for juvenile cases. The collaboration with JCJC ensured the availability of critical variables such as prior offenses, referral details, and placement histories.

These data sources were merged to create a comprehensive dataset for analyses, ensuring consistency in assessing both service-level and youth-level outcomes. Integrating this data was a complex and collaborative effort that involved multiple stages of data cleaning, reorganization, and evaluation. This included resolving discrepancies, creating consistent identifiers, and generating new summary variables for analysis. Both the EPIS and JCJC worked closely with the study team to ensure that the integrated dataset was as robust and accurate as possible.

The efforts of both organizations were integral to the success of the revalidation study through their strong commitment to data integrity and collaborative support throughout the process. Their shared dedication to advancing juvenile justice practices represents the standard for sustaining and enhancing evidence-based evaluation systems.

2.2 Sample Description

The study focuses on 113 distinct services that were "SPEP'd" between 2018 and 2023; those services involved 3,459 youth. Of these, 1,708 were unique individuals, as some youths participated in multiple services. Table 1 summarizes the demographic characteristics of the unique cases.

Table 1: Characteristics Summary

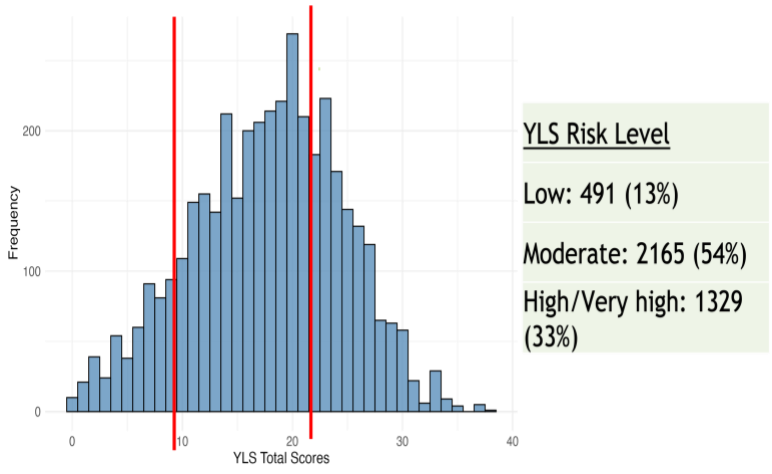
Characteristic	Number (%) or Mean
Age at SPEP™ service start date	16.6 years
Gender: Male	1,407 (82%)
Gender: Female	302 (18%)

The primary unit of analysis in this study is the service cohort, which represents a distinct service offered during a specified period, with scores assigned based on established SPEP™ guidelines (Lipsey & Chapman, 2017). To evaluate outcomes effectively, youth outcomes—such as recidivism rates—were aggregated at the cohort level. The definition of recidivism in this study was *a subsequent adjudication of delinquency or conviction in criminal court for a misdemeanor or felony offense within two years of case closure*. This definition is a regular part of statewide and county-specific juvenile justice practices (Juvenile Justice and Delinquency Prevention Committee on behalf of the Pennsylvania Commission on Crime and Delinquency, 2021).

Any analysis of recidivism outcomes, however, must account for individual characteristics that influence the likelihood of reoffending. To address this, the Youth Level of Service/Case Management Inventory (YLS/CMI™; Andrews, Bonta, & Wormith, 2004), a validated tool designed to assess the risk of reconviction over a 12-month period, was used with other measures. Regularly employed in Pennsylvania as part of JJSES, the YLS/CMI™ provided a consistent scale for assessing risk for this study. The Juvenile Court Judges' Commission (JCJC) supplied YLS/CMI™ scores closest to the start of each SPEP'd service for each youth. The YLS/CMI™ score at time of service was then combined with demographic variables to determine a youth's overall likelihood of recidivism.

The distribution of these YLS/CMI™ scores, which demonstrates relative normality, and the associated risk level percentages (as defined for use with this instrument) are depicted in Figure 2. This figure shows a broad distribution of assessed risk and a sufficient number of adolescents at each risk level to form a generally representative group of adolescents involved in the juvenile justice system.

Figure 2: Youth YLS Risk Score & Risk Level



2.3 Characterizing Recidivism

The Standardized Program Evaluation Protocol (SPEP™) is grounded in research indicating that specific aspects of service provision—such as type, quality, dosage, and the risk level of recipients—are associated with a reduced likelihood of reoffending among juvenile offenders (Lipsey, 2020). Services that adhere to these evidence-based standards achieve higher SPEP™ scores and are expected to exhibit a corresponding reduction in recidivism rates

among youth who complete the service. Consequently, recidivism serves as a central construct in evaluating the validity and effectiveness of SPEP™.

Assessing the impact of higher SPEP™ scores on recidivism outcomes is inherently complex due to variability in the characteristics and circumstances of the youth served. Differences in reoffending risk, as well as factors influencing service placement (e.g., court mandates, resource availability, or program goals), create disparities across service populations. Therefore, comparing observed recidivism rates without accounting for these effects would lead to inadequately tested, and therefore potentially misleading, conclusions. To address this, the evaluation examines the extent to which an adolescent's observed recidivism (0 or 1) differs from their expected rate, with the latter (the expected rate) reflecting what would reasonably be anticipated given the YLS/CMI™ score and background characteristics of the youth receiving the service.

2.4 Individual observed recidivism

In this study, "observed recidivism" is defined as a new adjudication or conviction occurring within a specified outcome period. This definition aligns with practices established by the Juvenile Court Judges' Commission (JCJC). The calculation involves several steps:

1. **Defining the Outcome Period**: The outcome period begins at the end date of the "SPEP'd" service and spans either 12 months (365 days) or 24 months (730 days).

2. **Adjusting for Time at Risk**: Days spent in placement during the outcome period are excluded to accurately represent the time the youth was at risk of reoffending.

3. **Classification**: Each youth is categorized as follows:

    - **Recidivist**: If a new adjudication or conviction occurred during the outcome period.
    - **Non-recidivist**: If no adjudication or conviction occurred during the entirety of the outcome period.
    - **Unknown**: Cases where the outcome period was incomplete (e.g., fewer than 365 or 730 days) and no adjudication or conviction was recorded. These cases are excluded from the analysis to ensure the integrity of the findings.

The exclusion of the "Unknown" cases, which lack sufficient follow-up observations, helps to eliminate what might be considered "false negatives." By focusing on cases with complete and observable outcome periods, the analysis provides a more accurate representation of recidivism rates within the defined periods. Cases with complete follow-up data have been observed for a sufficient time period to be confident that they did or did not recidivate. Cases with incomplete data may have recidivated during the unobserved time period.

In this revalidation study, 18% of youth (n = 617) were identified as recidivists within the 12-month outcome period, and 30% (n = 793) recidivated within the 24-month outcome period. Due to incomplete data, 36% of cases were excluded from the 24-month analyses, whereas only 16% were excluded from the 12-month analyses.

2.5 Individual adolescent's expected recidivism

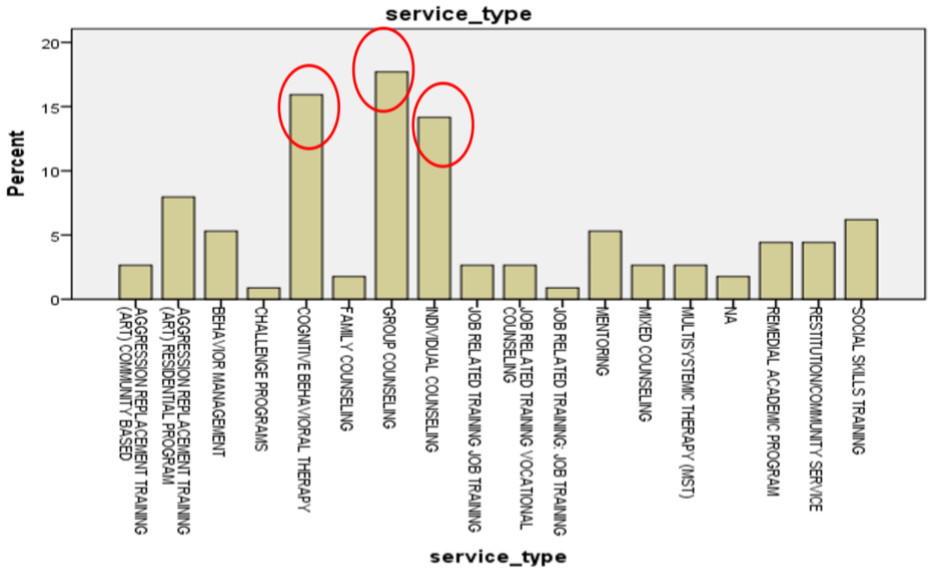
To evaluate expected recidivism, a score was calculated for each youth to estimate their likelihood of reoffending during the follow-up period. This score is derived from a regression model incorporating the following variables: - Gender - Race/Ethnicity - Age at the start of the SPEP'd service - YLS/CMI™ Total Score (score closest to service entry). Combining the individual adolescent's values for these variables generates a quantitative measure of the inherent reoffending risk for each youth at the time-of-service entry. A higher score (between 0 and 1) indicates a greater likelihood of recidivating.

2.6 Service Cohort Description

While individual-level data provides valuable insights, SPEP™ is fundamentally designed to evaluate services as a whole, not individual participants. Consequently, creating service cohort-level indicators is a critical component of this validation process. Each service cohort is comprised of the individual youth who were involved in a service for the concurrent time period covered by the SPEP ratings.

The EPIS provided data for 113 service cohorts delivered between 2018 and 2023, with each cohort representing a distinct service defined by a specific start and end date. These services were predominantly locally developed (81%) and residential (72%), encompassing a range of service types. The most frequently represented types were cognitive-behavioral therapy (CBT), group counseling, and individual counseling, as shown in Figure 3.

Figure 3: Types of Services Represented in the Data



- **Primarily residential settings (72%)**
- **Primarily locally developed programs (81%)**

On average, service cohorts included 35 youth, with a range of 9 to 150 participants. Youth in these cohorts had a mean age of 16.6 years (s.d. = 0.72) at the start of the service and typically participated for an average of 21 weeks (s.d. = 11; range = 2–58 weeks). The average YLS/CMI™ score for each cohort was calculated by simply taking the average of individual scores across that cohort's participants; the average of these averages across the 113 cohorts was within the "moderate" range (mean = 17.12, s.d. = 2.9; range = 6.9–21.3).

2.7 SPEP™ Scores

Two primary scores generated by the SPEP™ process are central to this revalidation study. While this report provides an overview, readers are encouraged to consult the initial validation report or the SPEP™ scoring guide (Lipsey & Chapman, 2017) for more detailed information.

1. Total SPEP™ Score: This score represents points awarded across five dimensions:

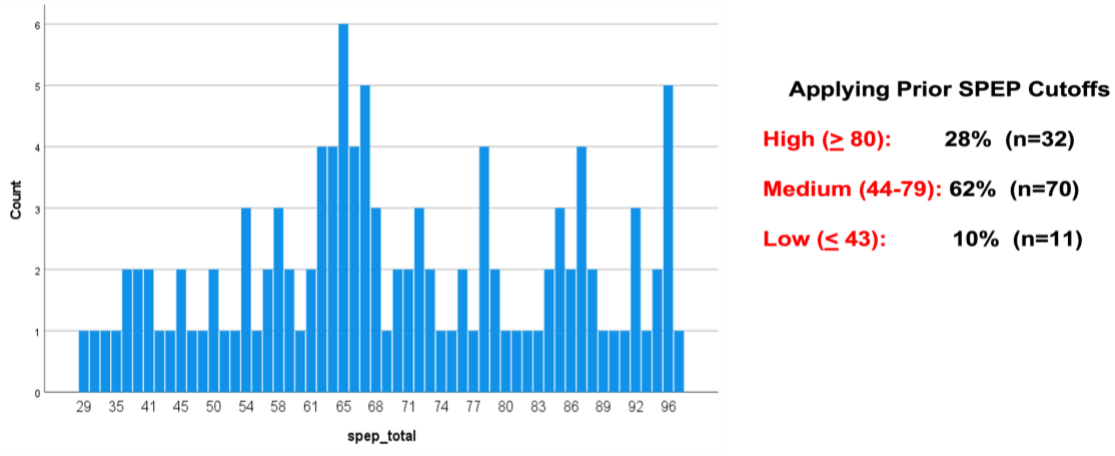
    - Service type
    - Supplementary services
    - Quality of service delivery
    - Amount of service (duration and dosage)[1]
    - Youth risk level

    The Total SPEP™ Score reflects the cumulative performance of a service based on these dimensions.

2. Program Optimization Percentage (POP) Score: The POP Score is a standardized metric that compares the Total SPEP™ Score to the maximum achievable score for a given service type and risk level. Expressed as a percentage, the POP Score reflects how well a service performs relative to its potential, effectively "correcting" for inherent differences in how particular types of services are designed and delivered. This adjustment ensures that comparisons across service types account for structural variations in service delivery, providing a fair and consistent benchmark for evaluation. Throughout this report, the POP Score is utilized as a key indicator of service dimensions and alignment with evidence-based practices.

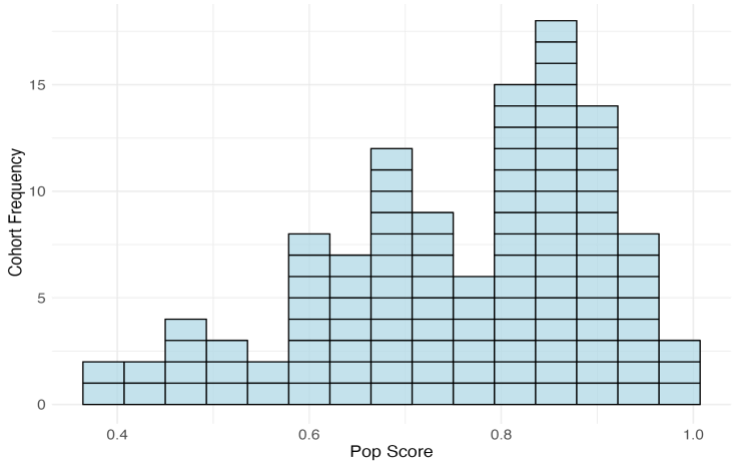
Across the 113 cohorts included in this study, the average Total SPEP™ Score was 68.5 (sd=17), with a range of 29–98 (Figure 4). Similarly, the POP Score averaged .76 (sd=.15), with a range of .98–.76 (Figure 5).

Figure 4: Cohort SPEP Scores



[1] Within the SPEP™ framework, dosage (total contact hours) and duration (weeks in service) are scored as distinct components. However, this report occasionally uses "amount of service" to refer to these two dimensions collectively when describing overall service exposure. Findings specifically independently associated with dosage or duration are specified as such.

Figure 5: Cohort POP Scores



Applying Prior POP Cutoffs

High (≥ .79): 54% (n=61)

Medium (.50-.78): 39% (n=44)

Low (≤ 49): 7% (n=8)

SPEP™ Total and POP scores can be analyzed by service category, service type, and primary service group. The tables below provide a breakdown of average scores for each of these classifications, offering a nuanced view of performance metrics across different service configurations.

2.8 SPEP™ and POP mean scores by *service category*

The revalidation study encompasses a diverse range of services used regularly in juvenile justice interventions, each tailored to address specific behavioral and developmental needs. Service category and type are determined as part of the SPEP™ process. Trained SPEP™ staff conduct a review of materials and a thorough interview with the provider, a probation officer, and a SPEP™ consultant. Once the service category and type are finalized, the service is assigned points in proportion to the average overall magnitude of recidivism effects found in the research for that service type (see Lipsey & Chapman, 2017 for additional elaboration of the service types and associated scoring). The primary service categories and service type developed by the process include: Restorative Justice Services, which emphasize repairing harm and promoting accountability through victim-offender mediation or community service (i.e., mediation, restitution/community service); Skill-Building Services, designed to enhance interpersonal, vocational, or life skills (i.e., behavior management, cognitive-behavioral therapy, skills training, challenge programs, remedial academic programs and job-related training); and Counseling services, designed to assist individuals and families with personal, social or psychological difficulties (i.e., individual counseling, mentoring, family counseling, family crisis counseling, group counseling, mixed counseling).

Each service category reflects distinct goals and methodologies, contributing uniquely to youth development and reducing recidivism. The following table summarizes the mean SPEP™ Total and POP scores across these three service categories. These scores provide a high-level view of how services with distinct theoretical orientations perform relative to the SPEP™ standards.

Table 2: SPEP & POP Mean Scores by Service Category

Service Category	N of Cohorts	SPEP Score (average)	POP Score (average)
Restorative	5	51	.64
Counseling	50	61	.70
Skill-building	56	76	.82

Notes. Significantly different (p < .000) SPEP and POP scores across categories. Table totals (N = 111) reflect two cohorts without assigned service categories; mean SPEP™ and POP scores are unaffected.

2.9 SPEP™ and POP mean scores by *service type*

Service-level analysis breaks down the SPEP™ Total and POP scores by specific service types. This detailed view highlights the variability in performance among different service types.

Table 3: Average SPEP and POP Scores by Service

Service Type*	N of Cohorts	SPEP score	POP score
Restitution/Community service	5	51	.64
Individual Counseling	16	51	.67
Mentoring	6	65	.68
Family Counseling	1	72	.85
Group Counseling	20	67	.71
Mixed Counseling	3	57	.67
Behavioral Management	6	86	.90
Cognitive behavioral management	17	84	.84
Multisystemic Therapy (MST)	4	71	.84
Aggression Replacement Therapy (ART)	12	90	.90
Social Skills training	7	56	.66
Challenge Programs	1	73	.86
Remedial Academic Program	5	65	.81
Job related training	7	54	.72

*Notes. *2 cohorts were not assigned a type*

2.10 SPEP™ and POP Mean Scores by *Service Group*

Lipsey and Chapman (2017) also note that, as part of the SPEP™ process, service types covered in the meta-analysis database were assigned to one of five groups according to their average level of effectiveness for reducing recidivism. In Table 4, we display SPEP™ Total and POP scores within this grouping framework, offering insights into performance trends within defined clusters of service offerings. These differences are not particularly surprising, since the categories were developed by Lipsey and his colleagues (Lipsey and Chapman, 2017), page 19) based on their progressively positive values seen in their research. These obtained values support that original work and indicate that the sample of cohorts in this study reflect values seen in other samples.

Table 4: SPEP & POP Mean Scores by Service Group

Service Group	N of Cohorts	SPEP Score	POP Score
Group 1: *Individual counseling; Job-related training*	23	39.2	.68
Group 2: *Restitution; community service; Remedial academic program*	10	57.7	.72
Group 3: *Family counseling; Family crisis counseling; Mixed counseling; Social skills training; Challenge programs; Mediation*	16	62.1	.73
Group 4: *Group counseling; Mentoring; Behavioral contracting; Contingency management*	34	70.5	.74
Group 5: *Cognitive-behavioral therapy*	30	86.2	.86

2.11 SPEP™ cut-off scores

In the initial validation study, data-driven subgroups of SPEP™ Total and POP scores were identified to establish low, medium, and high-performance benchmarks. These cut-offs aligned strongly with recidivism outcomes in the initial sample, and as a result, were integrated into the messaging for SPEP™ staff in setting goals for programs; program staff were encouraged to strive for a SPEP™ score at or above the mean score for the "high" group. Table 5 shows the distribution of cohorts across these three designations in the initial validation study.

Table 5: Cohort Scores in Benchmark Groups (from Initial Validation Study)

Group	% of Cohorts	N (Cohorts)	Min Value	Max 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

For the present revalidation, we retained the same cut-off thresholds (e.g., below 50 for "low," 50–80 for "medium," and above 80 for "high") to facilitate direct comparison with prior findings. Each service's Total and POP scores were evaluated against these benchmarks to determine whether the previously established ranges still offer a meaningful way to predict outcomes. This approach allows us to examine how shifts in program performance—as indicated by changes in their SPEP™ or POP scores—might continue to relate to recidivism. A more detailed presentation of how these categories were applied in the new sample and the associated findings is provided in Section 3.2

2.12 Cohort recidivism rates

To assess service-level impacts, recidivism metrics were calculated at the cohort level. These metrics help evaluate the effectiveness of services assessed using the SPEP™ protocol. The calculation of individual level recidivism values has been discussed in Section 2.3 above. This section outlines the process of transforming these individual rates into cohort recidivism rates.

2.12.1 Cohort observed recidivism rate

The cohort observed recidivism rate represents the proportion of youth within a service cohort classified as recidivists. This rate is calculated by dividing the number of recidivists by the total number of youths with valid recidivist or non-recidivist classifications in that cohort, excluding cases marked as “unknown.”

For example, consider a cohort of 10 individuals, 8 of which have valid individual observed recidivism, and 5 of the 8 with positive recidivism data (they were arrested). The calculation for the observed recidivism figure for this cohort would be 5/8, or .63.

2.12.2 Cohort expected recidivism rate

This section describes the process for calculating the expected recidivism rate for individuals. The cohort expected recidivism rate is the average of individual expected recidivism scores for youth within the cohort. These scores reflect the likelihood of recidivism based on youth characteristics upon service entry.

For example, consider again our fabricated cohort of 10 individuals from above. Let us posit that their expected recidivism scores are 0.34, 0.40, 0.23, 0.19, 0.33, 0.38, 0.18, 0.23, 0.26, and 0.39. The sum of these scores is 2.93, which, when divided by 10 (the cohort size), yields an average expected recidivism score of 0.29. This aggregated measure indicates the cohort’s inherent risk of recidivism.

It is important to note that a single youth may participate in multiple service cohorts with different start and end dates. Each unique youth-service combination generates an independent outcome period. For example, if a youth begins Service A upon entry and starts Service B at a later date, recidivism is calculated separately for each service involvement. This ensures that outcomes reflect the distinct contexts of each service experience.

2.13 Cohort recidivism difference score

A key question in this revalidation is whether higher-rated SPEP’d services demonstrate significantly lower observed recidivism rates compared to their expected rates. This is evaluated by calculating the recidivism difference score for each cohort. The recidivism difference score for each cohort is computed as the *observed recidivism rate minus the expected recidivism rate for that cohort*. Positive scores indicate worse-than-expected outcomes (observed rate exceeds expected rate), while negative scores indicate better-than-expected outcomes (observed rate is lower than expected rate).

For instance, in the hypothetical cohort described earlier, if the observed recidivism rate is 0.63 and the expected rate is 0.29, the recidivism difference score would be 0.63 - 0.29 = 0.34. This positive value suggests that the cohort performed worse than anticipated based on the youths’ risk profiles.

3 Results

3.1 Relationship between SPEP and POP scores and program effectiveness (Aim 1)

This first aim investigated how Standardized Program Evaluation Protocol (SPEP™) and Program Optimization Protocol (POP) scores relate to service effectiveness, as measured by

reductions in recidivism. Results showed no statistically significant relationship between these scores and recidivism at the 12-month follow-up. However, at 24 months, there was evidence of an association for both SPEP™ ($p < .10$) and POP ($p < .04$) scores, with higher-rated programs yielding better recidivism outcomes.

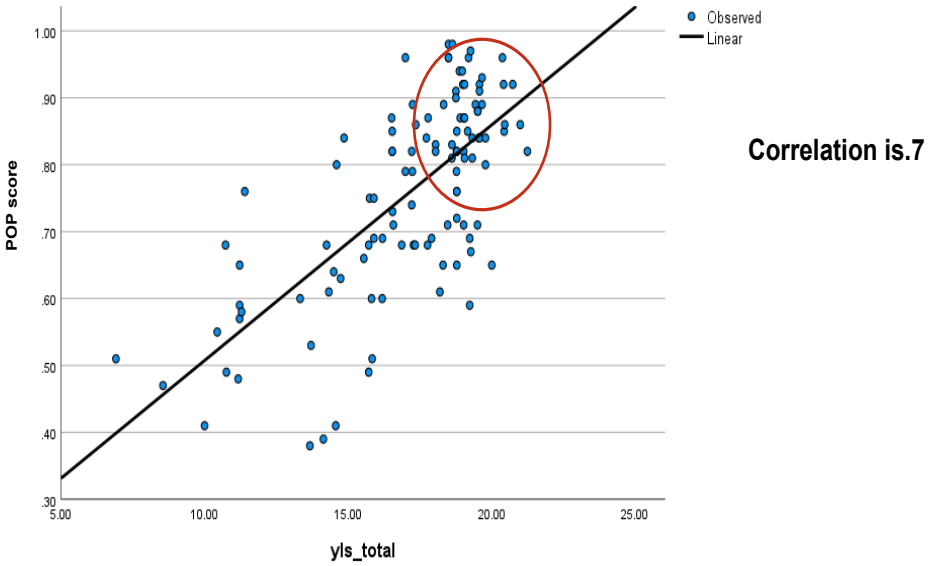
The relationship between the POP scores and recidivism differences at 24 months is shown in Figure 6 below. Each dots indicate a cohort and its corresponding POP value (on the x-axis) and its recidivism difference score value (on the Y-axis). There is a distinct and statistically significant linear relationship between the POP scores and the recidivism differences at 24 months, as indicated by a dark line heading downward in this figure. In general, services with higher POP scores tended to achieve more favorable than expected recidivism rates, in line with previous validation studies. This observed relationship, however, is not as strong as the effect observed in the original validation study.

Figure 6: 24-Month POP Score Differences

Several factors may account for this smaller (but still statistically significant) difference in the impact of the overall service scores[2]. First, high-risk youth might be funneled into the highest-rated programs, which could limit their effectiveness; higher-risk youth might be more difficult to affect positively. Figure 7 below shows the relationship between the cohort POP score (on the Y-axis) and the average YLS score for that cohort (on the X-axis). Given that the correlation between these two variables is .7, this selection of high risk/need cases to better-rated services seems to be occurring in this sample.

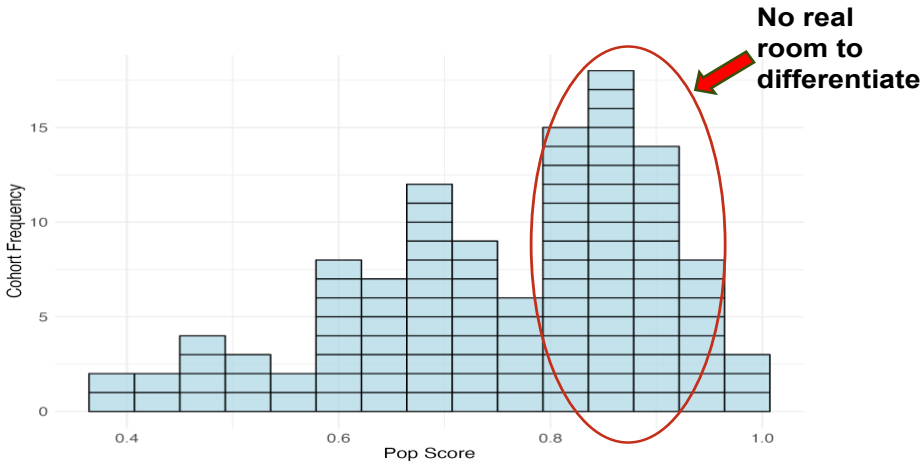
[2] An analysis of cases that did not follow this general pattern of the relationship between POP scores and recidivism difference was conducted ("outliers analysis"). This analysis in summarized in Appendix A. These analyses had small samples but generally support the importance of particular dimensions of the SPEP™ scoring system for reducing recidivism reduction.

Figure 7: Cohort POP Score Services and Average YLS



Second, a "ceiling effect" could be limiting the ability to see differences among cohorts at the top of the scale. If a high number of programs are performing near the top of the scale, there is little room to score higher on the measure of performance. It is difficult to distinguish among the services at the high end because there is no way to achieve more points for performance. An overall test of the scale's performance can therefore underestimate it's association since the amount of variability observed at the top of the scale is constricted. This seems to be happening in this revalidation sample, as indicated by the disproportionate number of cohorts at the top of the scoring distribution shown below in Figure 8.

Figure 8: Cohort Frequency & POP Scores



These two phenomena will limit the strength of the observed association between the POP score and recidivism differences. In some sense, then, the the recidivism assessment of the SPEP™ initiative could be thought of as "a victim of its own success" in terms of achieving it's operational goals. More youth with higher risk are referred to better services, and the services have shown demonstrably better (but possibly underestimated) performance overall.

In sum, while higher SPEP™ and POP scores showed promising associations with improved recidivism outcomes at the 24-month mark, these effects were not firmly established at 12 months (although the effect moved in the same positive outcome direction). Such findings underscore the need for ongoing refinements to SPEP™ and POP metrics to strengthen their utility as predictors of program effectiveness.

3.2 Validity of prior scoring ranges (Aim 2)

The second aim of the revalidation study evaluated whether the previously established cut-off thresholds for SPEP™ Total and POP scores remain useful in distinguishing program effectiveness. Specifically, the study re-applied the original three-tier classification system—low, medium, and high scoring groups—to the revalidation sample to determine whether these ranges continue to meaningfully differentiate observed versus expected recidivism outcomes. The application of these cutoffs to the original validation sample and the revalidation sample is shown below in Table 6.

Table 6: Prior Validation Cut-offs Applied to Current Sample

SPEP Score Level	Prior cut-offs	Prior Sample N (%)	Current sample N (%)
High	≥ 80	35 (22%)	32 (28%)
Medium	44–71	97 (61%)	70 (62%)
Low	≤ 43	26 (16.5%)	11 (10%)

POP Score Level	Prior cut-offs	Prior Sample N (%)	Current sample N (%)
High	≥ .79	56 (35%)	61 (54%)
Medium	.50–.78	89 (56%)	44 (39%)
Low	≤ .49	13 (8%)	8 (7%)

Overall, the findings aligned with earlier trends—higher SPEP™ scores generally correlated with better outcomes—but no statistically significant differences emerged among the three scoring tiers at 12 or 24 months. For instance, although higher-scoring programs tended to yield lower recidivism, the differences between low-, medium-, and high-scoring groups were not statistically meaningful. Similarly, POP score thresholds did not generate notable distinctions in observed versus expected recidivism at either 12 or 24 months. This differs from previous validation efforts, which found that cohorts in the lower SPEP™ tier performed worse than expected, especially at the 12-month interval.

Notably, the revalidation sample included a much higher percentage of cohorts (54%) in the “high” scoring group, compared to only 22% in the original validation study. This shift suggests that, as a whole, programs in Pennsylvania have made positive improvements in the areas assessed by the SPEP™ process. These improvements do not appear to be the result of shifts in the demographic characteristics of the youth (since the initial and revalidation samples were very similar in that regard). Rather, it may be that the implementation of the SPEP™ process has helped raise the overall quality of programs and services. In particular, the process may have informed service providers about which aspects of service implementation should be monitored and targeted for improvement.

Despite the lack of statistical significance in this revalidation, general patterns suggest that the original SPEP™ cut-offs (e.g., above 80 for high performance, below 50 for low performance) remain useful. Programs classified as “high-performing” continued to be associated with reduced recidivism, reinforcing the practical merit of these benchmarks. However, the medium-scoring range offered less clarity, as outcomes varied widely across services. These observations highlight a continued need to test and refine the score thresholds, ensuring they remain applicable to diverse program contexts and youth populations.

A few challenges affected these results. First, the revalidation sample had fewer programs at the low-performing end, limiting the ability to evaluate cut-offs across the entire scoring spectrum. Second, high-performing cohorts were overrepresented, potentially making it harder to detect group-level differences. Future research should incorporate broader, more balanced datasets to fully assess and update these scoring categories.

3.3 Impact of Service Components (Aim 3)
Aim 3 of the revalidation study examined how the core components of the SPEP™—namely Amount of Service, Service Quality, and Service Type—influence recidivism outcomes at 12 and 24 months.

Amount of Service The analysis evaluated both total points for service duration (defined as the number of weeks between the start and end date of the cohort-specific service) and dosage (number of contact hours) for their effect on recidivism reduction. Neither measure showed a significant relationship to recidivism at 12 or 24 months. While these two dimensions of service amount might well bolster other positive outcomes (e.g., improved employment or peer relationships), it does not appear to exert a direct effect on reducing recidivism.

Service Quality Two quality indicators were investigated: the total count of quality measures adopted by a program and the points accrued from those measures. These indicators were strongly correlated ($r = .9$), and taken together, they had a marginally significant link ($p < .08$) to 12-month recidivism outcomes. When analyzed separately, each indicator showed a significant effect ($p < .03$) on 12-month recidivism. However, these effects did not persist at 24 months.

Service Type Cohorts were initially categorized into five groups in line with Lipsey’s established protocol and coinciding with SPEP™ and POP scores:

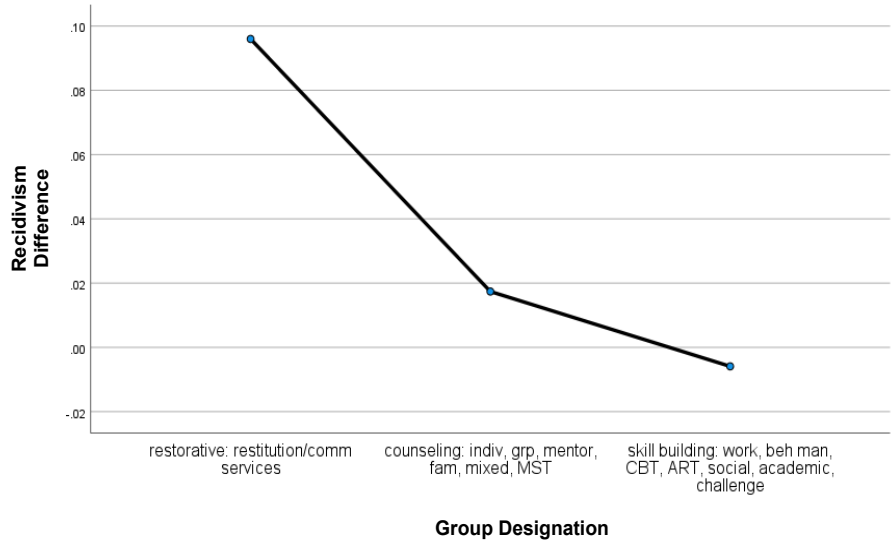
- Group 1 (individual counseling, job training; n=23) – lowest average scores (SPEP 39.2, POP 0.68)
- Group 2 (restitution, community service, remedial academics; n=10) – moderate increase (SPEP 57.7, POP 0.72)
- Group 3 (family counseling, social skills, challenge, mediation; n=16) – higher average (SPEP 62.1, POP 0.73)
- Group 4 (group counseling, mentoring, behavioral contracting; n=34) – further improvement (SPEP 70.5, POP 0.74)
- Group 5 (cognitive-behavioral therapy; n=30) – highest overall scores (SPEP 86.2, POP 0.86)

Testing recidivism differences across these five categories revealed a marginally significant ($p < .16$) effect at 12 months and none at 24 months. However, average difference scores generally aligned with the SPEP™ and POP trends, with Group 5 consistently showing the most favorable

outcomes. Further analysis found that Group 5 significantly outperformed Groups 1 and 2 at 12 months ($p < .05$), though the overall effect across all five groups was modest.

Two other comparisons of cohort grouping strategies were also done. A comparison of evidence-based vs. locally developed programs showed no significant differences at either follow-up period. However, reclassifying these cohorts into three service categories (restorative, counseling, and skill-building) provided clearer distinctions, yielding statistically significant differences in 12-month ($p < .02$) and 24-month ($p < .03$) recidivism difference scores. These values are seen in Figure 9, below. This suggests that more aggregated groupings based on theoretical orientation may better reveal meaningful differences in outcomes.

Figure 9: Service Type (3 groups)



Overall, Aim 3 highlights the nuanced roles of SPEP™ components in reducing recidivism. Although amount of service did not predict recidivism directly, service quality demonstrated short-term effectiveness, and service type emerged as an important factor over both timeframes[3]. These insights underscore the need to customize evaluations based on quality and type, while acknowledging that increased service amount may benefit areas outside of recidivism. Additional research is recommended to refine the alignment of these SPEP™ components with the specific needs and goals of diverse juvenile justice programs.

3.4 Changes in SPEP™ Scores Over Time (Aim 4)

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 reviews by the SPEP™ team are conducted to determine if the service has improved.

[3] There was no evidence of significant interaction effects between these predictors.

An examination of SPEP™ scores over time provides an additional, more stringent test of the validity of the SPEP™ process. 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 in the sme agency being measured twice (once before SPEP™ involvement and once after the performance improvement plan is implemented).

Aim 4 examined whether shifts in SPEP™ and POP scores between initial evaluations and later reassessments are related to recidivism difference outcomes using a larger sample of services than was available at the time of the initial validation. A service can improve, stay the same, or reduce its scores from the initial assessment to the reassessment. The test here was whether changes in the same service ratings over time (moving up or down in SPEP™ or POP score) significantly affect recidivism difference scores.

Figure 10 below shows the direction of SPEP™ and POP score changes for this sample of reassessed services. As can be seen in the bar charts below, slightly more services moved up in their ratings (above zero on the x-axis) than moved downward (below zero on the X-axis). There were still large enough groups moving in each direction, however, to assess the relationship between recidivism difference scores and shifts in the scores for the service.

Figure 10. Change in SPEP & POP Scores between B01 and R01

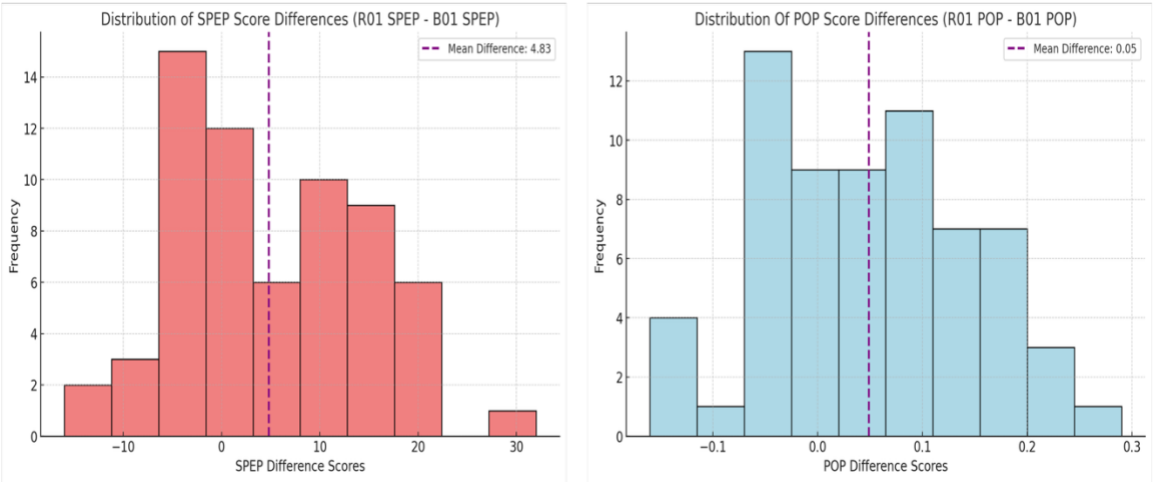
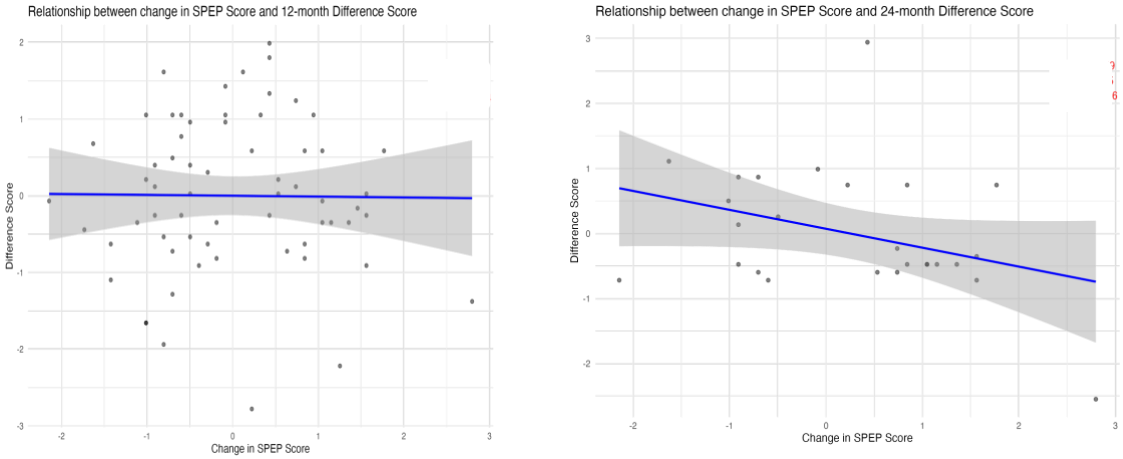


Figure 11 shows scatterplots of change in SPEP™ score (X-axis) versus recidivism difference score (Y-axis) at 12- and 24-month follow-ups. Each point represents one service case that received a reassessment. At 24 months (n = 26), greater improvements in SPEP™ were significantly associated with larger reductions in recidivism difference scores ($\beta = -0.29$, $p < 0.05$), meaning each one-point increase in SPEP™ change corresponded to a 0.29-point drop in recidivism relative to baseline. No relationship was evident at 12 months (n = 64; $\beta = -0.01$, $p = 0.93$).

These results suggest that SPEP™ gains may predict longer-term recidivism reductions, highlighting the value of repeated assessments for monitoring program effectiveness and guiding targeted improvements. However, these findings were constrained by the small sample

sizes—and the fact that the 12- and 24-month cohorts only partially overlap—underscoring the need for larger, matched-sample studies to confirm the effect.

Figure 11: Change in SPEP Score for 12-month & 24-month Difference Scores



Additional analyses indicated that programs initially rated in the medium-performance range exhibited the greatest changes over time, moving either upwards or downwards between assessments. In contrast, high- or low-performance cohorts tended to remain relatively stable. This pattern suggests that medium-performing services may be especially responsive to interventions aimed at strengthening evidence-based practices. Although periodic reassessment of SPEP™ and POP scores offers a dynamic view of how programs evolve, additional research is essential to clarify how these score changes relate to long-term recidivism outcomes.

Overall, the results reinforce the value of ongoing reassessments as a mechanism for continuous quality improvement (CQI). By monitoring fluctuations in SPEP™ and POP scores, stakeholders gain valuable insights into both the effectiveness of interventions and the progress of service enhancements—ultimately helping programs reduce recidivism and achieve improved outcomes for youth.

In particular, the observed association between improvements in SPEP™ scores and reduced recidivism at 24 months highlights the practical utility of the scoring system—not just as a static benchmark, but as a dynamic tool tied to meaningful youth outcomes. This finding supports institutionalizing periodic SPEP™ reassessment as part of routine quality monitoring and suggests that mid-range programs, which showed the greatest movement, may be especially well-positioned to benefit from targeted technical assistance and CQI support.

3.5 Matching youth criminogenic risk/need to appropriate services (Aim 5)

The risk-need-responsivity (RNR) model (Andrews & Bonta, 2010a & b; Andrews, Bonta, & Wormith, 2006) is widely regarded as a cornerstone of effective intervention within the justice system. This model promotes an evaluation of an individual's risks and needs for referring individuals to effective interventions to reduce recidivism. In Pennsylvania, the systematic use of the YLS/CMI™ is the preferred method within the juvenile justice system to identify risk/needs. The YLS/CMI ™ has strong empirical support for its reliability and construct validity, with more limited empirical support for its ability to predict recidivism (Onifade, et al., 2008).

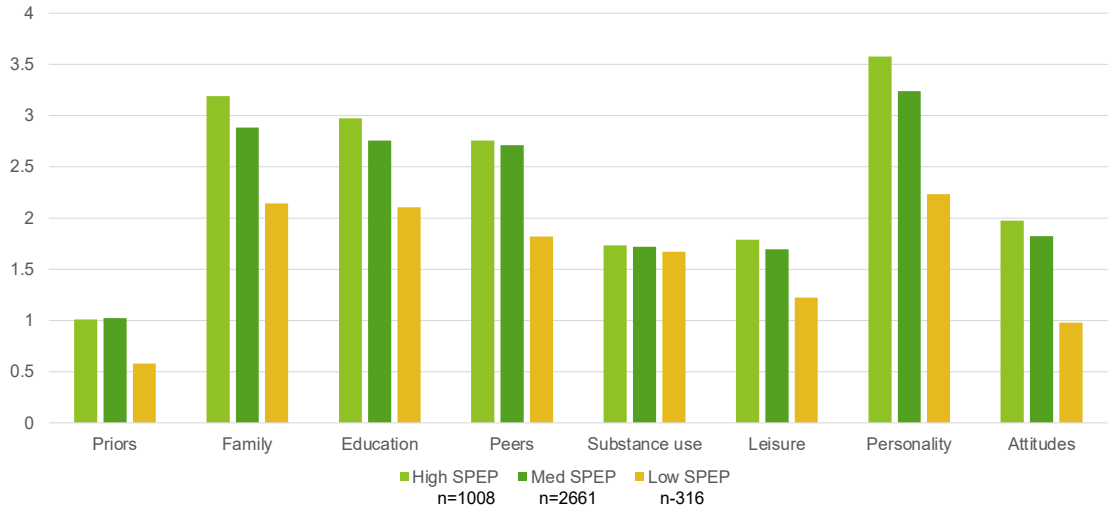
Work done by the originators of the SPEP™ process highlighted that the risk of the youth (in terms of prior history) and the needs of the youth should be considered together in the assessment of a service's projected recidivism reduction. This work considered the general risk/need level of youth within services cohorts by including the percentage of the cohort with a low, medium, or high YLS total score into the SPEP™ scoring algorithm. As noted in the SPEP™ Guide (Lipsey & Chapman, 2017, page 15), "research on delinquency intervention programs on which the SPEP™ is based has shown that, on average, there are larger positive effects on recidivism with higher risk juveniles than with their lower risk counterparts. As a result, juveniles' risk scores must be included in the SPEP™ scoring scheme along with the other elements found to be significantly related to program effects on recidivism." Although the SPEP process makes no assumptions regarding the match between the specific risk/need profile of youth and specific service recommendations, this prior work indicates that the examination of prior needs alone (without prior criminal history) may show limited relations.

In this context, the goal of this particular revalidation study aim is to examine the specific risk/need "profiles" for youth in the SPEP'd services; a retrospective examination of this sort might uncover patterns in the risk/need-service dyad that could explain some of the previously discussed findings regarding recidivism outcomes for youth in these SPEP's services. As a reminder, there are eight risk/needs assessed (i.e., prior and current offending, family circumstances & parenting, education & employment, peer relations, substance use, use of leisure time & recreation, personality/behavior, and attitudes/orientation). The YLS/CMI™ is completed by juvenile justice staff (usually probation officers) who use a combination of file information, youth self-report, and interviews to rate each risk/need of that adolescent. Consideration of these ratings is meant to inform the probation officer about the most reasonable service(s) needed by the adolescent.

3.5.1 *Comparing the risk/needs ratings to SPEP™ scores.*

Our initial step in this examination involved plotting the average risk/needs scores from several vantage points. Specifically, we looked at the average risk/need score by service type, by residential vs. community settings for each service type and by services with high, medium or low SPEP scores (this latter view is shown in Figure 12).

Figure 12: Average Risk/Needs Scores for Individuals Within Cohorts at Various SPEP Score Levels



*Significant between-group differences in all domains except Substance Use

This figure illustrates two points. First is the relatively equal level of average needs scores across all the types of needs; family, education, peers, and personality need groups have similar average values (about 2.5 to 3). Priors, substance use, leisure, and attitudes have somewhat lower average needs scores (1 to 1.5). Judging from this graph, however, the difference in the average overall needs scores for each level of SPEP™ score is not great enough to have much practical utility.

Second, the overall pattern of the distribution of needs scores in relation to the SPEP™ scores is very similar across almost all of the types of needs. For six of the eight needs groups, the average score of needs is higher in the cohorts with higher SPEP™ scores. Within each need, the cohorts with higher SPEP™ scores are getting adolescents with higher needs. The only exception to this pattern is the need of substance use services. For the ratings of substance abuse needs, each level of SPEP™ score is receiving youth with roughly the same level of substance use needs. In general, this graph illustrates that the previously reported assignment of higher risk/needs cases to better SPEP™-scoring services is occurring evenly across all identified needs.

In general, it appears that adolescents with higher needs are going to services with higher SPEP™ scores. This expands the earlier finding of adolescents with higher YLS/CMI™ scores being referred to services with higher SPEP™ or POP scores, but it illustrates that this phenomenon is occurring across just about all service types since a mix of service types are included in the services which yield a high, medium or low SPEP™ score. Within each individual risk/need type, better scored services are receiving adolescents with higher level of needs. This similarity of mean values for the majority of needs at each level of the SPEP™ score makes it unlikely that there are distinct sets of needs that might prove worthwhile for focused assessment and consideration.

3.5.2 *The relationship between risk/needs scores and service type*

Our second approach was to see if there were distinct risk/needs that were significantly related to receiving a service of “x” type. Logistic regression models used service type as the outcome variable and the risk/need scores as predictors. These models were run singularly for each of the service types thus testing what needs were significantly related to assignment to that service type (Table 7)

This table shows the p-values (indicating statistical significance or close to statistical significance) for each of the risk/need scores (in the first column) to predict involvement with a particular service type. The first row of the table indicates that, taken together, the full group of all the risk/need scores predicted referral to each service type. The other cells of the table show which particular risk/need factors were statistically significant (or close to statistically significant) in producing this overall effect.

These models showed no clearly discernible patterns. Each of the risk/needs ratings are significantly related to multiple service types, with each risk/need rating significantly related to between two and eight service types. In addition, each service type is predicted by more than one or two risk/need factors. Finally, risk/need ratings do not have a clean association with service type, and the profiles of each service type in terms of risk/need do not follow a consistently logical pattern. For example, higher educational needs are related to receiving family, group, individual, and mixed counseling and mentoring, but not remedial education or social skills training.

Table 7: P-values for Risk/Need Scores Predicting Service Type

	ART	Behav Mgt	Challenge	CBT	Fam Counsel	Grp Counsel	Ind Counsel	Job Rel	Mentor	Mixed Counsel	MST	Rem Acad	Resti/ Comm	Social skills
Overall Model	.000	.000	.010	.006	.001	.000	.003	.000	.000	.000	.000	.000	.000	.015
Priors	.003	.015		.022	.009	.092				.005	.059		.006	
Family	.088	.028								.068	.022			
Education					.018	.018	.061		.020	.086				
Peers			.037				.085				.033	.023		.052
Sub Use	.085	.063					.012	.073		.059	.018		.065	.030
Leisure													.059	.035
Personality	.000	.002	.043					.003		.011		.004	.002	
Attitude				.023	.012	.012		.018	.000			.004		

Notes. notes: ART = Aggression Replacement Training; Behav Mgt = Behavior Management; Challenge = Challenge Program; CBT = Cognitive Behavioral Therapy; Family Counsel = Family Counseling; Grp Couns = Group Counseling; Ind Counsel = Individual Counseling; Job Rel = Job-Related Services; Mentor = Mentoring Program; Mixed Counsel = Mixed Counseling (e.g., combination of individual/group/family); MST = Multisystemic Therapy; Rem Acad = Remedial Academic Instruction; Resti/Comm = Restitution and/or Community Service; Social skills = Social Skills Training.

3.5.3 *Factor analysis of risk/need scores*

One possible explanation for the lack of clear patterns in the risk/need-service type analyses described above is that each of the individual scores of risk/needs may be too specific to yield clear findings regarding service types. It might be that a single need is not salient enough to determine the decision-making regarding the type of service recommended. We therefore looked to whether there is a way to combine the specific risk/needs scores of the youths into fewer, agglomerative scores reflecting more general case characteristics.

We used factor analysis to achieve the goal of reducing a number of variables or items collected on a sample to a few calculated scores ("factor scores") that simplifies the interpretation of the set of variables. Factor analysis allows for a case to be characterized by several summary scores that derive from weighting the raw values (in this case, particular risk/need scores) into a single factor or set of factors. In other words, this statistical approach tests whether certain specific individual scores "hang together" closely. For instance, it might be the case that the risk/need scores for personality, attitude, and peers could be combined into a score for something like a "tough attitude" score that might be related to a service type.

This factor analysis approach starts by examining the correlation matrix of the risk/needs scores. The correlation matrix is simply a table of calculations indicating how strongly each individual risk/needs score is related to each other risk/needs score. The correlation of the risk/needs scores is presented in Table 8 below.

Table 8: Correlations of Risk/Needs Scores

	Prior Score	Family Score	Education Score	Peer Score	Substance Score	Leisure Score	Personality Score
Family	.093**						
Education	-.029	.320**					
Peer	.171**	.345**	.227**				
Substance	-.013	.221**	.104**	.246**			
Leisure	.169**	.287**	.208**	.369**	.167**		
Personality	.015	.422**	.426**	.216**	.088**	.169**	
Attitude	.197***	.442**	.338**	.389**	.156**	.333**	.492**

This table indicates that the Pearson correlations between the needs scores vary between slightly below .15 to slightly above .40. This level of association among rating variables is neither exceptionally low or high; it is not indicative of particularly strong relationships among readily identifiable pairs of risk/need scores. All of the individual correlations between needs scores are highly statistically significant ($p<.001$), with the only nonsignificant correlations occurring with some particular needs and the prior history rating (as would be predicted by the Lipsey et. al research cited above). These consistent estimates of statistical significance among the needs scores, however, have no real-world significance; they are undoubtedly driven by the very large sample size. Larger sample sizes make smaller associations between scores much more likely to be statistically significant.

Using the correlation matrix, the factor analysis then identifies the risk/needs scores that are most correlated to each other and can be logically weighted and combined into a set of more general scores. The procedure then sees if this set of distinct, more general scores is strong enough to provide an adequately accurate account of relationships seen across the whole correlation matrix. These more general scores are called "factors." If the identified factors are accurate representations of the correlation matrix and make logical sense, we could then score each individual on each "factor" and see if these are related to other constructs or variables, like service type or recidivism outcomes.

The factor analysis of the individual risk/needs scores yielded three factors. These three factors accounted for 62% of the variance seen in the initial correlation matrix, an acceptable level of performance from such an analysis. With a weighted scoring of the risk/needs, there appear to be three factors in the existing rating.

The potential risk/needs scores of the youths in the SPEP'd services can be characterized as:

> ***Factor 1****. Multiple problems across needs.* This factor score is an indication of the case having a rather consistent scoring of needs across all the rating scales.
>
> **Factor 2**. *Relative importance of substance use needs*. This factor score indicates substance abuse needs are higher than other needs scores.
>
> **Factor 3**. *Higher priors; antisocial peers; limited leisure activities*. This factor score indicates the extent to which the case had this particular combination of needs.

These factor scores are inconsistently related to recidivism at 24 months. Taken together, these factor scores are significantly associated ($p<.001$) with whether or not an adolescent is a recidivist at this follow-up point. Of the three factor scores, however, only the scores for Factor 1 (multiple problems across needs) and Factor 3 (higher priors; antisocial peers; limited leisure activities) contributed significantly ($p<.001$) to the observed association with this recidivism measure. The score for Factor 2 (Relative importance of substance use needs) does not contribute significantly to the association with recidivism when the other two factor scores are included as independent variables. In addition, this Factor 2 (Substance use needs) is not a significant predictor, being unrelated to recidivism at 24 months when tested alone ($\beta=.02$; $p<.66$)[4].

[4] Factor analyses on the full adolescent revalidation sample (n = 3,985) and on unique adolescents (one YLS score each; n = 1,706) yielded two factors with eigenvalues > 1.0, explaining 52 % of variance. These corresponded to Factors 1 and 3, with no substance-use-needs factor.

The above analyses lead to several conclusions. First, there are statistically significant correlations among the risk/needs ratings, but these associations are not very powerful, and the statistical significance values seen in the above table are largely attributable to the large sample size. Second, there are some ratings that do "hang together" enough to create three factor scores that adequately reflect the associations among these ratings (multiple problems across needs, relative importance of substance use needs, and higher ratings of prior arrests, antisocial peers, and limited leisure activities). Higher scores on these three factors together would indicate a higher likelihood of recidivism at 24 months in the SPEP'd sample.

The conceptual coherence and practical utility of this factor solution, however, seem rather limited. Importantly, the factors identified with this sample of youth in SPEP'd service may or may not hold if the sample included a different group of justice-involved youth. Second, even if these factors were determined to be valid in a different sample, applying an algorithm to the risk/needs ratings could produce factor scores for each adolescent, but these scores would seem to contribute little to the formulation of a service plan. The first factor would be based on an adolescent having generally a higher number of needs across multiple ratings, something that can be determined by a simple scan of the rating scales. The second factor would be elevated if the rating for substance use is appreciably higher than other ratings. This might be useful for identifying adolescents for referral to or placement in a service for substance use problems. Then again, such a general decision rule could probably be implemented by probation officers from a scan of the ratings as a set. Also, it is important to remember that this second factor alone does not have an independent effect on recidivism. Finally, the third factor identifies the likelihood that an adolescent has a prior offense history, hangs with antisocial peers and has little involvement in community activities. Again, such a classic pattern of an adolescent likely to go on to future offending could probably be identified by a scan of the highest ratings from the set of possibilities.

The findings from this exercise could prove useful in focusing an examination of the pattern of risk/needs in the juvenile justice population more broadly. However, it is certainly not a blueprint for an automated system that would provide estimates of the appropriateness of particular services or the likelihood for recidivism for particular adolescents. The potential for identifying particular patterns of risk/needs, however, could be a general consideration for probation officers in making recommendations.

3.5.4 *Cluster Analysis of cases based on risk/need scores*

The factor analysis reported above examined how well each of the risk/needs ratings was associated with each of the other risk/need ratings. It examined how well the different ratings "hung together" across the whole SPEP'd sample. It also seems plausible, however, that there might be groups of youth who have similar risk/need patterns across the whole set or ratings (meaning, they score high or low on a similar set of risk/need factors). Cluster analysis was used to identify relatively homogeneous groups of cases ("cluster groups") that have a distinct similarity over the whole set of risk/need ratings.

A cluster analysis does largely what the name implies; it systematically "agglomerates" cases that have close patterns and values of ratings. Using the whole set of each adolescent's risk/need ratings, this procedure looks for a small number of cases whose pattern of scores appear very similar. It then keeps adding newly identified cases to these groups by seeing how much they look like a member of one group versus the others. The optimal solution to such an analysis is the one that most distinctly identifies a set of "clusters" in the sample of cases.

Three groups were identified in this sample of risk/need scores. These three groups of cases provided an adequate depiction of how the sample of adolescents could be divided according to the values of their risk/needs rating. The values of the average of each risk/need rating in the three clusters is provided in Table 9 below.

Table 9: Risk/Need Scores by Cluster

	Cluster 1	**Cluster 2**	**Cluster 3**
Family Score	1.3	3.7	3.7
Education Score	1.6	2.0	4.4
Peer Score	1.7	3.3	3.0
Substance Score	1.2	2.2	1.9
Leisure Score	1.2	2.0	1.8
Personality Score	1.7	3.0	4.8
Attitudes Score	0.6	2.0	3.6
	n = 1,307	n = 1,178	n = 1,500

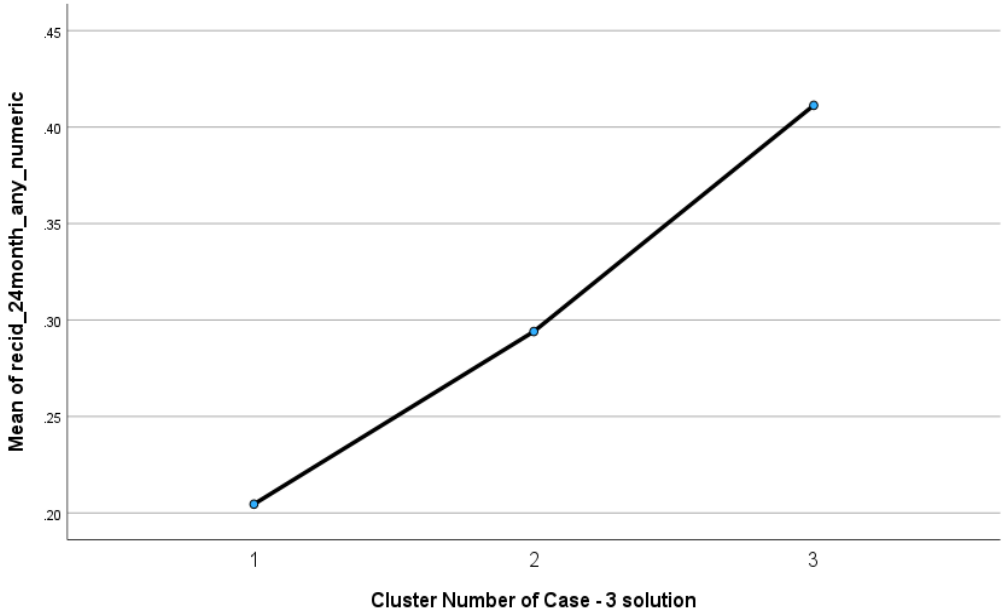
This solution illustrates some notable differences among the three identified groups. In this sample of youth in the SPEP'd services, family, education, personality, and attitude ratings follow a consistent difference in the mean rating score across the three clusters; Cluster 1 having the lowest mean score, cluster 2 a midrange value and Cluster 3 having the highest value. Peer, substance use, and leisure activities all show an increasing mean value from Cluster 1 to Cluster 2, and then a decreasing or no difference between Cluster 2 and 3. All averages increase between Cluster 1 and Cluster 3 by a relatively large amount.

The clearest pattern of these average scores seems to be an increase in the mean ratings between Cluster 1 and Cluster 3. All of the rating increases between these two clusters are substantial, except for the substance use and leisure scores. Cluster 2 averages appear to be mainly in the mid-range between those of Cluster 1 and 3.

This pattern illustrates two points. First, the most efficient method for grouping these ratings appears to be simply according to the seriousness of the ratings across all the rating scales. This implies that the ratings are working in unison at identifying more serious from less serious cases (as would be expected given intent and empirically demonstrated functioning of the YLS/CMI™). Second, the differences in this overall pattern are when the mean values for Cluster 2 do not substantially or consistently differ from Cluster 3. The ratings for Cluster 2 that are not conforming to a straight increase across all three groups are those for peer, substance use, and leisure activities. This would seem to indicate that these three ratings also define a set of cases with midrange values that do not neatly follow the pattern of distinctively low or high values seen more generally.

The slightly different pattern for the ratings of *peer, substance use*, and *leisure activities* conform closely to the findings of the factor analysis, where one factor (Factor 2) loaded heavily on these values. It is important to remember, however, that Factor 2 did not have a significant independent association with recidivism at 24 months. It could be instructive therefore to test the association of the cases in each cluster with this outcome variable of interest. Figure 13 shows results of the relationship of the cases in each cluster with the 24-month recidivism rate.

Figure 13: 24-month Recidivism by Cluster

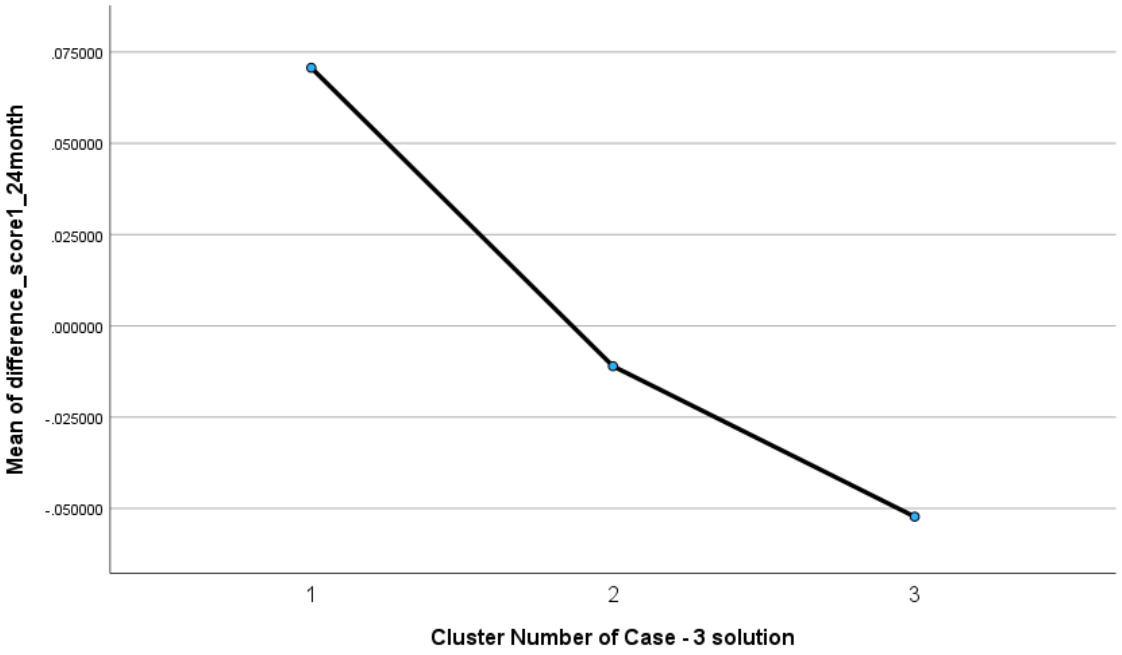


This figure shows a clear relationship of cluster group membership and the likelihood of recidivating at 24 months (ANOVA p<.001 across all three cluster groups). The recidivism rate for the members of Cluster Group 1 is about 20%, while the rate of offending for members of Cluster Group 3 is slightly over 40%. The cases in Cluster Group 2 have about a 30% chance of recidivism. These figures indicate a rather strong association between group membership and subsequent offending.

Another relationship between cluster group membership and success in service involvement is also evident in the study data. Figure 14 below shows the relationship between cluster group membership and the difference score between observed and expected recidivism (the recidivism reduction effect) of service involvement. This figure shows a clear relationship between cluster membership and effectiveness of the intervention provided to the adolescent (ANOVA p<.01 across all three cluster groups).

The largest difference between observed and expected recidivism (the lowest negative value or greatest recidivism reduction) is found for Cluster Group 3 at -.05. Group 1 shows a positive mean value, indicating a recidivism rate higher than expected (at about .75). Cluster group 2 has a slightly negative value (indicating almost no reduction in the likelihood of reoffending).

Figure 14: 24-month Difference Score by Cluster Group



Even though Cluster Group 3 (those with the highest risk/needs) has the highest recidivism rate at the 24-month follow-up (about 40% based on Figure 14), this mean difference between the observed rate and expected rate seen above is still about 5% less than what would be expected. Meanwhile, Group 1 (those with the lowest need/risk ratings) have a 24-month recidivism rate of about 20%, a rate about 7.5% higher than would be expected after receiving services. While the raw recidivism rate of the Group 3 adolescents is higher than the other groups, it is lower than expected; the raw recidivism rate of Group 1 is lower than the other groups, but it is higher than would be expected.

This is a demonstration that service involvement has a more beneficial effect on adolescents with the most risk/needs, and a seemingly detrimental effect on adolescents with the lowest level of needs. This finding replicates a prior finding by Lipsey and his colleagues (Lipsey & Chapman, 2017), but with a diverse, statewide sample receiving a variety of services. Intervention with adolescents with a low level of risk/need appears to raise the chances of recidivating, while services with adolescents with a high level of risk/needs appears to reduce recidivism.

Summary of risk/needs analyses

This set of analyses provided some limited insights into the patterns of risk/needs seen in this sample. The analyses addressed three issues: the relationships of the risk/needs scores to the service type the adolescent received, the interrelationships of the risk/needs ratings to each other, and possible methods to develop a unified system of more general scores to portray risk/needs patterns. We portray these efforts as providing “limited” insights because they have shown some interesting results, but few clear-cut methods to improve the risk/needs rating exercise.

Given that the major goal of this study is to examine the validity of the SPEP™ process, we first looked at the relationship of the average risk/need scores of adolescents referred to types of services and whether individuals with particular risk/needs were systematically going to more or less quality services. We then explored whether particular risk/needs retrospectively predicted the types of services in which the youth ended up.

As visualized in Figure 12 (risk /need by level of SPEP™ score), there are no practically relevant differences in the risk/needs of youth who ended up in SPEP'd services receiving a high or medium score; the high and mid-level scoring programs received youth with a similar range and average level of risk/needs while the services with low SPEP™ scores (typically job-related training or individual counseling) had youth with statistically (but not practically) lower mean scores. The exception to this pattern is substance use – youth with substance use needs are equally distributed across high, medium and low-scoring services.

Similarly, the logistic regressions models that were run with each of the individual risk/need scores predicting service type showed no clearly discernible patterns. A risk factor that significantly predicted receipt of one service type might also significantly predict two or three other service types. There did not seem to be any apparent pattern in the type of service received by youth with a particular set of risk/needs.

Two types of analysis were conducted to see if the risk/need rating could be combined into a smaller number of scores that gave a more succinct view of these associations. The factor analysis looking for associations of scores that hung together produced weighted combinations of scores that could reasonably (but not strongly) reduce the large set of risk/need ratings so that further investigation of appropriate service matches could be examined more efficiently. However, one of the factor scores (Factor 2) was not significantly related to 24-month recidivism.

The cluster analysis tested whether cases could be placed in homogenous groups with roughly similar patterns of scores. Three groups were identified in this sample of youth in SPEP'd service, with the general pattern being an increase in overall ratings with each cluster. The mid-level group (Group 2) in this solution mirrored some of the factor analysis results, identifying a group with a mix of ratings differing on the earlier ratings identified in the factor analysis. In both analyses, adolescents with risk/need scores elevated on substance use, limited social involvement, and a resistant attitude were likely to have about a 30% chance of reoffending at 24 months, but unlikely to benefit from any of the services offered (at least in terms of reduced recidivism).

In short, in this sample of youth involved with SPEP'd services, these analyses of risk/need ratings show some limited consistency, with adolescents with substance use, limited community involvement, and a resistant attitude being a group not particularly well served by the services provided. Identifying this group using a practical method for combining or assessing the pattern of risk/need scores, however, does not emerge cleanly from these analyses. The methodology needed to integrate these results into an automated assessment practice would seem to require considerable effort with limited return.

We would be remiss if we did not end with a cautionary note. As noted earlier, this examination of risk/need-service dyads was undertaken to determine if there were any observable patterns that might explain some of the findings from the SPEP™ revalidation (e.g., why there isn't a consistent reduction in recidivism at 24-months for all service types). The available data thus allows for a retrospective look at the risk/needs of youth who ended up in services that were SPEP'd; the risk/need association with services and the groupings of risk/need (as in the cluster

analysis) may look different if the data were from a different sample. Furthermore, our exploration is a very different task than it would be to examine risk/need for the full scope of juvenile justice-involved youth and/or any investigation of the most productive match of interventions to risk/need in the juvenile system as a whole.

4. Conclusions

Summary Point #1: ***The investment of time and resources into SPEP™ has been demonstrated to be worthwhile.***

Conclusion: Two validation efforts have now demonstrated that total SPEP™ scores and/or POP scores are statistically significantly linked to meaningful reductions in recidivism. Across both validation studies, total SPEP™ and POP scores have shown statistically significant associations with meaningful reductions in recidivism. Notably, services that underwent reassessment—i.e., received more than one SPEP™ rating over time—demonstrated improved recidivism outcomes related to increases in their SPEP™ scores. This pattern across multiple assessments provides compelling within-service evidence that strengthening SPEP™-aligned practices can contribute to better youth outcomes. In this revalidation, we also observed that programs in the mid-performance tier exhibited the greatest variability in scores between assessments, suggesting they may be particularly responsive to targeted quality improvement efforts. As shown in Figure 11, increases in SPEP™ scores were significantly associated with improvements in recidivism difference scores at 24 months, reinforcing the role of reassessment not only as a quality control mechanism but also as a catalyst for sustained program growth.

Implications: The investment of time and resources in SPEP™ has proven worthwhile, and the SPEP™ effort should continue and expand. The value of another replication study, however, seems negligible. Examining change over time in SPEP™ services could be worth continuing, if a larger, stratified sample with more varied service types could be recruited to take part. Such a study could provide more information on the impact of improvement in certain areas of program improvement or differential success with particular types of adolescents.

The priority could now shift away from a concern with whether SPEP™ "works." Instead, it would seem more important to focus on preserving the integrity of the SPEP™ system and refining its application. Periodic re-assessment of programs would also be valuable if a larger effort could be put into this practice.

Summary Point #2. ***The relation between the POP scores and recidivism reduction, although still statistically significant, was less powerful in the second validation than in the first validation.***

Conclusion: No statistically significant differences at the 12- or 24-month follow-up period emerged among the three scoring tiers ("low," "medium," and "high") identified in the first evaluation and adopted as program target goals subsequently. Programs classified as "high-performing" continued to be significantly associated with reduced recidivism in this revalidation, reinforcing the practical merit of this higher benchmark for effectiveness. The "low" and "medium" scoring range offered less distinction, as outcomes varied widely across services at each of these levels. Upon further investigation, it was clear that there was a difference in the distributional shapes of the SPEP™ and POP scores in the two studies. There was a much higher proportion of services that achieved higher POP scores in the second validation sample. There were two positive factors driving this difference: a) more services were improving in their overall performance, and b) more adolescents were being directed to more effective services.

An analysis of adolescent YLS/CMI™ scores across the distribution of POP scores in this revalidation indicated that adolescents with significantly higher YLS/CMI™ scores were being sent to services with higher POP scores. As more serious-risk youth were directed to these high-performing services, it may be increasingly difficult for such programs to sustain further reductions in recidivism—what some have termed a "ceiling effect."

Implications: In this revalidation, a higher proportion of the services in the sample were at the high end of their SPEP/POP-related ratings; a good result in terms of improving overall service performance. As a result, though, the distinction between low-performing and medium-level performing services may no longer be meaningful, as few services remain in the lower tier, and outcomes among medium scorers vary widely. As a result, a **two-level classification system** may now be more empirically justifiable: one group consisting of **"high-performing"** services, and a second combining the current **"medium" and "low"** tiers. In addition, having adolescents with more serious risk/need profiles being regularly referred to higher performing services, it may be more difficult to differentiate service performance at the higher level of POP scoring. If a large proportion of programs are already performing at or near maximum effectiveness with increasingly serious groups of adolescents, they may have limited capacity to reduce recidivism further. Some consideration might be given to identifying metrics (e.g., adolescent improvement on specified additional scales) beyond POP score to differentiate positive dimensions of high-scoring services.

Summary Point #3: ***Particular dimensions of SPEP™ operations appear to be more consistently related to successful recidivism reduction than others.***

Conclusion: Within the SPEP™ process, certain dimensions appear to drive reductions in recidivism more predictably than others. Service quality was especially influential in differentiating high- and low-performing services, whereas service type was only significantly associated with recidivism reduction using the three-category classification of restorative, counseling, and skill-building services. Items related to the amount of service (both dosage and duration) showed some associations with recidivism reductions (in the analyses of outlier cases), but these effects were not powerful or consistent. It is likely that any effect of the amount of service is affected substantially by characteristics of the adolescent and their adjustment to the service, rather than the direct effect of the amount of service on recidivism.

Implications: These findings raise the possibility of streamlining the assessment of service quality to better support improvements in lower-scoring programs, while also differentiating the top-performing ones. It may be possible to create a simplified system of assessment related specifically to quality that could prove more effective at boosting the performance of low-scoring services. Given the centrality of service duration to juvenile justice system operations, further research might also examine whether service duration benefits areas other than recidivism or interacts with distinct youth characteristics in more complex ways. The effects of service amount might be found in the assessment of outcomes other than recidivism (e.g., job placement; family reunification).

Summary Point #4. ***There were no distinct profiles evident in the YLS/CMI™ risk/needs ratings of the adolescents represented in this sample of SPEP'd services.***

Conclusion: Analyses of YLS/CMI™ risk/needs revealed no distinct "profiles" of adolescents based on specific needs (e.g., family, attitudes). Most youth in this sample showed an inconsistent pattern of multiple risk/needs. In addition, no particular service type had a clear pattern of risk/needs reflected in the youth referred. There was some indication that adolescents

with substance use, leisure, and attitude problems may be somewhat distinct. Identification of a group of adolescents with this constellation of risk/needs, however, does not coincide with a reduced level of offending or a heightened impact of service involvement on recidivism. Overall, the sample of youth involved in the SPEP'd services was most accurately characterized as having generally low, medium, or high levels of needs. Moreover, whether the adolescent had low, medium, or high level of needs was related to the amount of recidivism reduction, with the highest risk/needs group having the highest level of recidivism reduction (replicating an earlier finding by Lipsey and his colleagues).

Implications: Assuming that this sample of youth involved in SPEP'd services is similar to the broader juvenile justice population, most adolescents present a variety of risk/needs. Thus, the practice of matching a narrowly focused service to a single need is difficult to implement. The possibility of developing an automated system of integrating assessment of risk/needs into the choice of intervention also seems rather low. Better approaches might involve alternative ways to structure and guide probation officers' use of these ratings.

An assessment of the overall level of risk/needs would seem more congruent with the logic and practice of probation officers making this assessment, as is the current practice. An overall "score" representing the level of risk/needs appears more in line with the goals of assessment and decision making than trying to formulate "packages" of correlated needs (as in the factor analysis and cluster analyses we described above). Alternatively, a useful system could be developed that requires a judgment on the probation officer's part regarding the priorities of an adolescent's identified risk/needs and then focusing service assignment to the most seemingly important needs (e.g., a revamped county-level service matrix for use in disposition decision making). Together, these practices could focus service recommendations more productively.

Summary Point #5. ***The next step in improving service outcomes is to address the question of how to improve the match between an adolescent and the type of service provided.***

Conclusion: The results of these validation studies support the contention that the SPEP™ framework can validly rate service provision. The SPEP™ system provides a framework for promoting more effective service provision in the juvenile justice system by widespread application as part of continuing quality improvement. The next step is to improve how youth are matched to specific services.

Not all services work equally well for every adolescent, but a comprehensive, uniform system for characterizing which risk/needs align best with which service types remains underdeveloped. The next challenge could then be how to devise systems that effectively match an adolescent to a particular service or facility with the highest likelihood of success for that adolescent. Certain services work better with particular types of adolescents, but we currently have no uniform system for characterizing adolescents or services on the most relevant dimensions. Pennsylvania is in a unique position to tackle this issue because the infrastructure to do so is already in place. YLS/CMI™ tool is already used statewide (offering systematic information on youth risk/needs), and the Juvenile Court Justice Commission (JCJC) has detailed information on placement, service involvement, and recidivism.

Implications: The use of SPEP™ and JCJC information could inform a number of activities. The use of SPEP™ assessments alone (or augmented with simplified, validated instruments for methods to tap relevant dimensions of operation) could be integrated as a routine requirement of services or facilities contracting with the courts. A service or facility could be required to

demonstrate a base level of competence to qualify as a contractor. In addition, an ongoing system for assessing services and/or facilities over time could be put into place (as the SPEP™ process does now). Eventually, recognition levels for achievement of quality markers could be developed and made available to other courts and the public, similar to the system used in early childhood education in Pennsylvania (https://www.pakeys.org/keystone-star).

Additional methods could be put into place to capture the fit of an adolescent to a particular service or facility. Developing a system to more elaborately characterize an adolescent's risk/needs would be necessary. The current reliance on the YLS/CMI™ items alone does not seem sufficient; work on defining and possibly expanding the scoring for the ratings could be useful. In addition, the categorization of service types could be refined and expanded. A useful change would be having agencies or facilities provide information about the "packages" of services provided to an adolescent over time, rather than an assessment of specific services (e.g., cognitive behavioral therapy). In addition, services or facilities could be required to specify the logic model or "treatment goals" for their services. This could clarify which needs they are designed to address. Eventually, the state could create a more detailed and relevant matrix that matches youth to the most appropriate service(s), aligning their profiles and providing consistent and diversified outcome tracking (beyond the current singular focus on recidivism).

Summary Point #6: ***The current data systems for assessing the impact of the SPEP™ process are not designed to take on the challenges of addressing the issue of matching adolescents to appropriate service providers.***

Conclusion: The current SPEP™, JCMS, and JCJC databases are impressive, and their consolidation is a notable effort, providing integrated information assessing service provision, processing, and outcomes throughout Pennsylvania. The data sets used to validate the SPEP™ process, however, are not structured to support complex analyses and more focused research queries. For instance, the datasets used in these studies are drawn exclusively from programs that volunteer for SPEP™, limiting broader system-wide generalizations. Future questions will likely require data sets with different information or organization. The question will have to drive the data sets used, rather than having immediately available datasets drive the questions asked.

Implications: Moving forward, Pennsylvania could leverage its existing data infrastructure to explore several questions touched upon here. JCMS data could be augmented with additional information available to JCJC staff in coordinated and focused efforts to address multiple questions of interest, e.g., the development of more sophisticated methods to measure risk/need, outcomes other than recidivism, and methods for integrating the assessment of youth with court action. The basic infrastructure needed to conduct such investigations is a valuable asset that the state already has in place and could cultivate.

Such efforts could be facilitated by a research division that addressed specific questions across the operations of JCJC, the courts, juvenile probation, and service providers. An interorganizational body could identify specific issues that would be useful for improving practice and work together with research applicants to refine the focus and designs of the envisioned studies. Working with the Juvenile Justice and Delinquency Prevention Committee, this group could identify specific research issues and solicit proposals for relevant research efforts. Sampling criteria, data construction, and analysis and interpretation could be done collaboratively by staff and researchers across agencies.

Taken together, these conclusions underscore the continued utility of SPEP™ in Pennsylvania's juvenile justice system—both as a powerful evaluative mechanism and as a catalyst for ongoing quality improvement. Future efforts would benefit from learning from and capitalizing upon the process of collaboration seen in the SPEP™ initiative.

References

Andrews, D. A., Bonta, J., & Wormith, J. S. (2004). *Level of service/case management inventory (YLS/CMI): User's manual*. Toronto, Ontario, Canada: Multi-Health Systems Inc.

Juvenile Justice and Delinquency Prevention Committee on behalf of the Pennsylvania Commission on Crime and Delinquency. (2021). *2021 Pennsylvania Juvenile Justice and Delinquency Prevention Plan*. Pennsylvania, USA: Pennsylvania Commission on Crime; Delinquency.

Liberman, A. M. (2017). Lessons from standardized program evaluation protocol demonstrations. *Journal of Juvenile Justice*, *6*, 52–70. Retrieved from https://www.ojjdp.gov

Lipsey, M. W. (2020). The role of evidence-based practices in the juvenile justice system. *Criminal Justice and Behavior*.

Lipsey, M. W., & Chapman, G. L. (2017). Standardized program evaluation protocol (SPEP™): A users guide. Nashville, TN: Vanderbilt University Peabody Research Institute.

Mulvey, E. P., Schubert, C. A., Jones, B., & Hawes, S. (2020). *A validation of SPEP™ in Pennsylvania*. Pennsylvania Commission on Crime and Delinquency. https://www.pccd.pa.gov/JuvenileJustice/Documents/A%20Validation%20of%20SPEP%20in%20PA%20Report.pdf

Redpath, B., & Brandner, C. (2017). Arizona's standardized program evaluation protocol: Implementation and impact. *Juvenile Justice Review*, *8*, 29–45. Retrieved from https://juvenilejustice.az.gov

Appendix A
Outlier Analyses

We conducted exploratory analyses to see if we could identify some of the factors that might be differentiating cohorts with similar levels of POP scores but markedly different recidivism outcomes (i.e., the difference between their observed and expected recidivism rate). This was done by identifying and comparing cohorts that, given their POP score and their recidivism outcome, seemed to have an opposite set of scores than expected. For example, this means comparing four groups (cohorts with low and high POP scores with cohorts with low and high recidivism outcomes). Such analyses (i.e., *outlier analyses*) are done to see if some similarities in these anomalous cases might indicate how the metric for judging an observed relationship (in this case between POP score and recidivism) is affected by some factor(s) other than the cohort's overall score on one of the scales.

The criteria for designating a case as an "outlier" involve some relatively arbitrary judgments. First, a cohort must be sufficiently low or high on the x-axis (in this case, the POP score) at a point with practical meaning. We chose the POP score value of .8 as the point for distinguishing high and low POP scoring because it is the target value presently used as a goal for services. Second, the group of cases designated as having outlying values cannot be too large in relation to the total number of cohorts in the quadrant examined. If the identified cohorts represent a large proportion of those in a quadrant, they could not be considered as "different" from the other cases. In the approach taken here, the group constructed as outliers for each quadrant ranged from 24% to 43% of the cohorts in that quadrant. Finally, the distance between the recidivism outcome for a given cohort must be sufficiently distinct from the outcomes of other cohorts at that POP score to indicate an anomalous result. This is a matter of subjective judgment when setting the boundaries of the outliers in each quadrant.

Figure A1. below illustrates the relationship between a cohort's POP score and the cohort's recidivism difference at 24 months after service completion. Each dot represents where a particular cohort scored on their POP rating and their recidivism difference. The solid line going downward illustrates the general overall trend of the cohorts to move toward negative recidivism difference scores as POP scores get higher. It goes down because the cohorts with higher POP scores have more negative difference scores, meaning that the observed level of rearrest was lower than the expected level of rearrest for those cohorts (service involvement *reduced* the likelihood of rearrest).

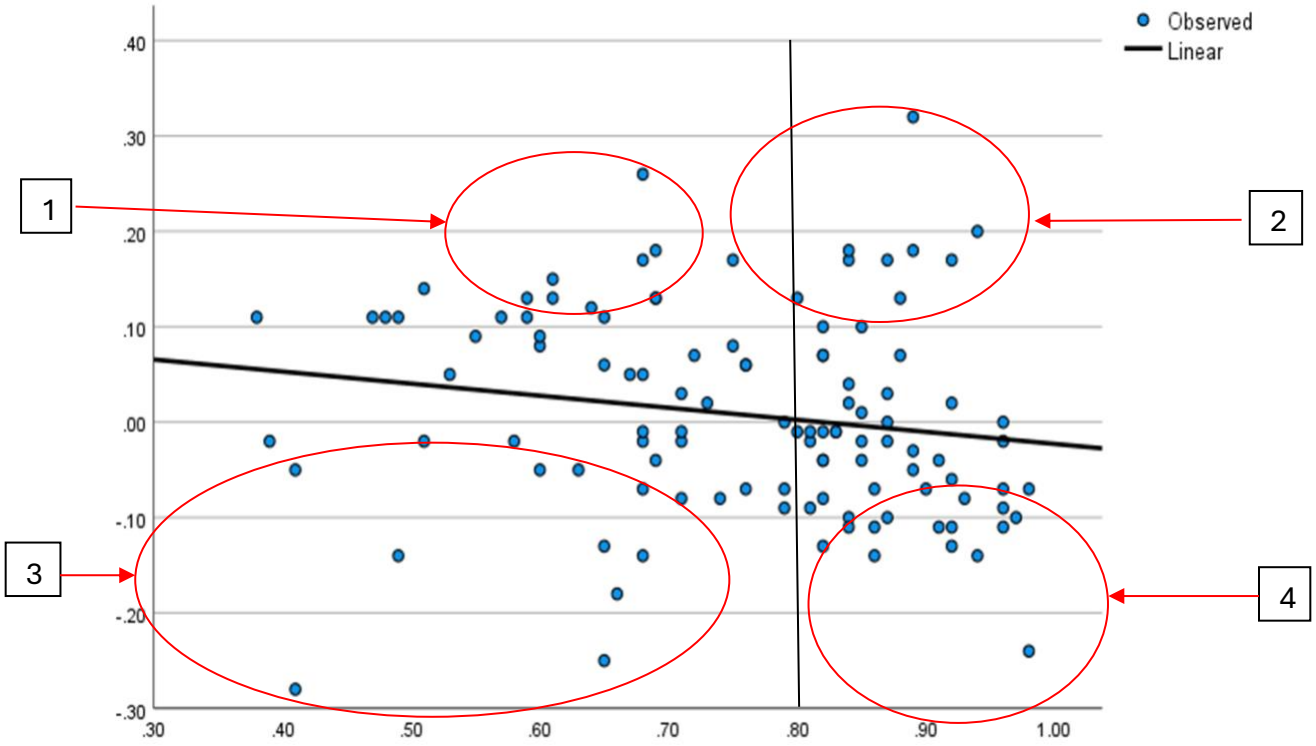


Figure A 1.

Applying the criteria outlined above, we identified some cohorts in each quadrant as outliers; i.e., cohorts that do not closely follow the pattern seen across the full sample (indicated by the solid black line above). These cohorts are those inside the numbered circles. Circle 1 indicates cohorts with lower POP scores and poorer than average outcomes (n=7); Circle 2 cohorts have higher POP scores, but poorer outcomes (n=9); Circle 3 cohorts have lower POP scores but better than average outcomes (n=10); Circle 4 cohorts have higher POP scores and higher than average outcomes (n=13). The number of cohorts in each outlier group is rather small. It would be desirable to have larger groups of distinct outliers to examine, but any strong differences among the groups should still be apparent.

Three types of possible group differences were explored. First, we examined the characteristics of the adolescents comprising each outlier group (e.g., average age). Next, we examined the types of programs (e.g., community/residential) that characterized each group. Finally, we examined differences among the groups in their dimensions of POP ratings (e.g., overall quality ratings).

Each set of analyses followed the same procedure. The statistical significance of overall group differences among the four outlier groups was assessed. If an overall statistical significance ($p < .05$) was observed across the four cohort groups, then comparisons between specific group values were examined.

Analyses of the characteristics of the adolescents comprising each outlier group indicated some differences across the four groups. Adolescents in Group 1 (low POP score service, poor recidivism outcomes) were significantly younger at the time of service than the adolescents in Group 4 (higher POP score service and better recidivism outcomes). The adolescents in Groups 1 and 3 (those in lower POP score services but with different recidivism outcomes) had lower

YLS scores, fewer prior placements, and were less likely to be white than the adolescents in Groups 2 and 4 (those in higher POP score services but with different recidivism outcomes). These observed findings seem to largely duplicate the previously reported finding of adolescents with higher risk and need profiles being assigned to services with higher POP ratings.

There were also some program-type differences among the four groups. Groups 1 and 4 (the lower POP score group with less favorable outcomes and the high POP score group with more favorable recidivism outcomes) both had higher proportions of residential programs. This seems to somewhat undercut the possible general conclusion that the outcome is overwhelmingly driven by the community/residential distinction. In addition, Groups 1 and 3 (groups with lower POP ratings, but distinctly different recidivism outcomes) were more likely to have counseling or mixed approaches, and Groups 2 and 4 (those with higher POP ratings and distinctly different recidivism outcomes) had high proportions of skill-building/CBT approaches. These results also seem to indicate that the general program type may have limited predetermination on outcome.

Particular POP-related scores were different among these outlier groups. In general, Groups 2 and 4 (those with higher POP scores but different recidivism outcomes) have higher service duration (overall amount, contact hours, weeks of service) and total quality measures. This distinction indicates that the ratings of contact and quality are appropriately linked to POP scoring but may have less ability to affect recidivism outcomes than desired at the high end of POP scoring.

The limitations of the analyses provide less than definitive findings about the strong drivers of the recidivism outcome or its relation to the POP scoring. One of the limitations is the small number of cohorts that could be clearly identified as outliers. Having a total sample of 113 cohorts and a relatively tight configuration of values around the projected line between the POP score value and the recidivism outcomes meant that only a small number of outliers (n = 39) could be identified confidently. When this sample is then divided into four groups, the differences between the groups have to be rather large to be statistically significant. Some subtler findings might have emerged from a larger set of cohorts.

These analyses of outlier groups do, however, provide some leads for inquiry and some caution about overgeneralization of previous findings. The findings about differences in background characteristics among the outlier groups provide some substantiation of the idea that the YLS scores are differentiating adolescents as intended, with more serious risk adolescents being funneled to higher performing services (without clear racial or gender bias). The findings regarding differences in program type elucidate the need to consider a broad range of issues in assessing interventions beyond program emphasis or setting for judging potential to reduce recidivism. Finally, the findings about the influence of program characteristics associated with POP scoring (e.g., quality, duration) highlight the need to consider these factors as useful, but not totally determinative, of the potential impact of a program on recidivism. In the end, these findings provide a mix of encouragement and caution about the POP score as a single metric of a program's likely success in reducing recidivism outcomes.