Washington State Institute for Public Policy
110 Fifth Avenue SE, Suite 214 • PO Box 40999 • Olympia, WA 98504 • 360.586.2677 • www.wsipp.wa.gov

September 2015

Washington's Coordination of Services Program for Juvenile Offenders: *Outcome Evaluation and Benefit-Cost Analysis*

Over the past two decades, the Washington Legislature has moved toward greater use of evidence- and research-based interventions in the juvenile justice system. The Washington State Institute for Public Policy (WSIPP) has evaluated several juvenile justice interventions to determine whether these programs reduce recidivism and whether the benefits outweigh program costs.

One of these interventions, Coordination of Services (COS), is implemented in Washington's juvenile courts and serves low-risk offenders. WSIPP's initial evaluation of COS was completed in 2004.[1] The implementation of COS has since changed. In 2010, the Washington Association of Juvenile Court Administrators, in collaboration with the Juvenile Rehabilitation Administration (JRA), implemented a quality assurance process to ensure program integrity.[2] The program also expanded from one to seven courts.

In this report, we present the results of WSIPP's new evaluation of COS.

Summary

Coordination of Services (COS) is an educational program for low-risk juvenile offenders that provides information about services available in the community. The program is designed to help juvenile offenders avoid further involvement with the criminal justice system. COS currently serves about 600 youth per year in Washington State.

The Washington State Institute for Public Policy (WSIPP) first evaluated COS in 2004 following its first year of implementation. As part of ongoing work to identify research- and evidence-based programming in juvenile justice, WSIPP re-evaluated COS to determine its current impact on recidivism.

In this new evaluation, we compared 699 COS participants to a group of 699 similar low-risk juvenile offenders from courts that did not offer COS during the study period.

Based on the results from both of WSIPP's evaluations of COS, we estimate that the program reduces recidivism by about 3.5 percentage points (from 20% to 16.5%). Our benefit-cost analysis of this result indicates that the benefits of COS outweigh its costs. The program costs about $412 per youth, and the benefits total $9,614 for a benefit-cost ratio of $23 to $1. We tested the uncertainty in this estimate and find that benefits outweigh costs 96% of the time.

Thus, for low-risk juvenile offenders, the program appears to be an effective way to reduce recidivism.

Citation: Fumia, D., Drake, E., & He, L. (2015). *Washington's Coordination of Services program for juvenile offenders: Outcome evaluation and benefit-cost analysis* (Doc. No. 15-09-1901). Olympia: Washington State Institute for Public Policy.

[1] Barnoski, R. (2004). *Outcome evaluation of Washington State's research-based programs for juvenile offenders* (Doc. No. 04-01-1201). Olympia: Washington State Institute for Public Policy.

[2] Department of Social and Health Services, Juvenile Rehabilitation. (2014). *Juvenile Court Block Grant Report*.

The report is organized as follows. Section I provides background on research-based interventions in juvenile justice and COS specifically. Section II outlines our methodology, while Section III summarizes the key findings from our evaluation and meta-analysis. Finally, Section IV presents our benefit-cost analysis of COS. A Technical Appendix is provided for supplemental analysis and technical detail.

I. Background

In 1997, the Washington State Legislature passed the Community Juvenile Accountability Act (CJAA).[3] This act establishes a goal of using research-based programs that will cost-effectively increase juvenile accountability and reduce criminal recidivism.[4] Funding for research-based programs is administered by the JRA through a block grant. The funding formula used to distribute the block grant allocates 25% of these funds to programs defined as "evidence-based."[5]

In 2004 and 2006, WSIPP completed evaluations of research-based programs for juvenile offenders including Aggression Replacement Therapy, COS, Family Integrated Transitions, Functional Family Therapy, and Multisystemic Therapy.[6] These programs constitute the set of research-based programs funded through the block grant.

At the time of these evaluations, one of these programs, COS, was implemented in only one court (Snohomish County). WSIPP found that participation in this program significantly reduced felony

[3] RCW 13.40.500 – 540.
[4] RCW 13.40.500 – 510. In this context, research-based means a program has sufficient scientific evidence to conclude that it can reduce recidivism if properly implemented.
[5] Department of Social and Health Services, Juvenile Rehabilitation, (2013). *Juvenile Court Block Grant Report*. "Evidence-based" under the funding formula generally means those programs that demonstrate a statistically significant reduction in recidivism. This definition more closely aligns with the "research-based" definition in RCW 71.36.010. We therefore use research-based throughout this report to reflect the legislative definition.
[6] Barnoski, (2004) and Aos, S., Miller, M., & Drake, E. (2006). *Evidence-based public policy options to reduce future prison construction, criminal justice costs, and crime rates.* (Doc. No. 06-10-1201). Olympia: Washington State Institute for Public Policy.

recidivism and had small but insignificant impacts on misdemeanor and total recidivism.

Since WSIPP's initial evaluation, the use of COS has expanded to seven courts and nine more have signed on to offer the program in 2015. In 2010, JRA contracted with a quality assurance specialist to develop tools to measure program adherence and ensure that each court's implementation maintains program integrity. Given these changes in COS implementation, we re-evaluated the program as part of WSIPP's ongoing work to determine what works and what does not work in Washington State to reduce juvenile recidivism cost-effectively.

COS Program Description

COS began in Washington State in 2000 as the "WayOut" program in the Snohomish County Juvenile Court. COS is an educational program for low-risk juvenile offenders and their parents or a connected adult. Program providers developed a COS manual that includes policies and procedures that ensure consistent implementation. COS has two primary goals: (1) to serve as an early intervention to prevent further criminal justice system involvement and (2) to provide youth and their family members with information about available services.[7]

After committing an offense, arrested youth are assessed using the Washington State Juvenile Court Assessment. To be eligible for COS, youth must be assessed as low-risk for re-offense and a parent or connected adult must also be available to attend program sessions.[8] Individual courts may impose other eligibility criteria, such as allowing only juveniles with certain offenses or without language barriers to attend.

COS is typically delivered through a 12-hour seminar over two or three days. Seminar sessions are run by various community juvenile justice and service providers such as the juvenile court, substance abuse prevention programs, Department of Social and Health Services, YMCA, and employment services programs. We estimate that COS costs about $412 per participant.[9]

[7] Tolan, P.H., Perry, M.S., & Jones, T. (1987). Delinquency prevention: An example of consultation in rural community mental health. *Journal of Community Psychology, 15*(1), 43-50.

[8] Prior to 2013, a youth's assessment also had to specify that the youth "usually obeys" parental authority and control to be eligible for COS. However, we find that more than 25% of COS participants did not satisfy this condition of eligibility. Thus, we do not impose this eligibility criterion in this analysis.

[9] Barnoski, R. (2009). *Providing evidence-based programs with fidelity in Washington State juvenile courts: Cost analysis* (Doc. No. 09-12-1201). Olympia: Washington State Institute for Public Policy.

II. Evaluation Methodology

Estimating the effect of COS on recidivism rates requires comparing COS participants (treated group) to a sufficiently similar group of individuals that are eligible for, but did not receive, COS (comparison group). Ideally, we would estimate this effect using an experimental research design where COS-eligible youth are randomly assigned to either the treated or comparison group. In a well-implemented experimental design, assignment to the treated and comparison groups occurs only by chance; thus any differences in later outcomes could be confidently attributed to COS.

Without random assignment, however, we must consider that those who participate in COS could be less (or more) likely to recidivate even in the absence of participation due to some other factor unobservable to the researcher. For example, youth who are most motivated to reduce their criminal behavior may be more likely to participate in the program. We would expect these youth to have lower recidivism rates because of their higher motivation regardless of participation.

Because random assignment did not occur in implementing COS, we rely on observational data to evaluate the program. We use an advanced statistical technique, propensity score matching, which can approximate group comparability on observed factors achieved with random assignment. We recognize, however, that propensity score matching may not eliminate all differences in unobservable characteristics.

Study Groups

The "treated group" are those individuals who participated in COS between January 1, 2011 and December 31, 2012; these youth come from the seven "COS courts." We include all youth who participated in COS regardless of completion.[10] The "comparison group" comes from the population of low-risk juvenile offenders from the remaining courts that do not offer COS ("non-COS courts") and were assessed between January 1, 2011 and December 31, 2012.

Youth in both the treated and comparison groups were excluded from the data if they were older than 18.5 at the start date, were participating in one of the other block grant funded research-based programs at the time of assessment, or had prior treatment from a research-based program.[11] Treated group participants were also excluded if they had no assessment prior to starting the program or if they were assessed more than 180 days prior to starting COS.[12] Finally, we excluded ten youth with missing data.[13]

[10] We estimate the treatment effect on the treated. We also retain all treated youth regardless of completion to avoid biasing the results toward the treated group because those that complete the program may also be more motivated and less likely to recidivate.

[11] We assume that a youth could commit a juvenile offense and have their juvenile risk assessment just prior to age 18. They could then participate in COS up to six months later making our cutoff 18.5 years of age.

[12] We chose 180 days because in some courts, COS is offered twice per year meaning a youth could have to wait up to 180 days for COS to be available depending on when they are assessed relative to when COS is offered. The average time between assessment and program start was about six weeks for the treated group.

[13] Prior to propensity score matching there were 864 youth in the treated group and 3144 in the comparison group.

We chose to use a comparison group from non-COS courts primarily because, as noted above, we are unable to control for unobserved factors that could impact a youth's participation in COS. Youth with access to COS who do not attend may do so for reasons that could also explain their likelihood to recidivate such as a preference for criminal behavior, lack of motivation, or lack of access to a parental figure or juvenile probation counselor who is able to help them participate. These unobserved factors cannot be addressed through propensity score matching alone. By drawing a comparison group from courts that do not offer COS, however, comparison group assignment is based on location rather than self-selection or selection on the part of others, such as parents or juvenile probation counselors.[14]

Recidivism Measure

We define recidivism as any offense committed in the 18-months following the "at-risk" start date that results in a Washington State conviction.[15] We define the "at-risk" date as the program start date for the treated group and the assessment date for the comparison group. The at-risk date for the treated group is the program start date while the at-risk date for the comparison group is the date they were assessed.

We chose to start the recidivism follow-up period for the treated group at the time of program start to avoid associating recidivism that occurs prior to program participation with the program itself. However, the disadvantage of choosing the program start date as the "at-risk" date is that any offense occurring after assessment but prior to program start will not count as recidivism. The average time between assessment and program start is six weeks, while the maximum time allowed by the study design is six months.

[14] We also explored using a comparison of youth from COS courts but found that 40% of youth that did not participate in COS were referred to other programs making it difficult to obtain a no treatment or standard treatment comparison group. Youth that did not participate in COS for reasons related to program availability would also constitute a strong comparison group; however, sample sizes for this group were too small to analyze. See the Technical Appendix for more information about the sensitivity of our results to our comparison group selection.

[15] The recidivism measurement period includes the 18-month follow-up plus an additional six months to allow for adjudication. We consider juvenile diversion a conviction for the purposes of measuring recidivism. This definition was established in a legislatively directed study; see Barnoski, R. (1997). *Standards for improving research effectiveness in adult and juvenile justice.* (Doc. No. 97-12-1201). Olympia: Washington State Institute for Public Policy.

If youth tend to recidivate shortly after assessment, we may observe reduced recidivism rates for COS youth simply because we fail to capture early offenses that occur between assessment and program start. To address this potential concern, we conduct a sensitivity analysis where we start the recidivism follow-up at the time of assessment for both treated and comparison youth; for further discussion, see the Technical Appendix.

Methods

We use propensity score matching to select the matched comparison group from the pool of COS-eligible youth—i.e., low-risk youth—in non-COS courts. Propensity score matching has three steps.

First, we estimate the propensity score—defined as the probability that a youth participates in COS—using a statistical model controlling for demographic, criminal, and behavior characteristics as well as county and court factors (see Exhibit 1 for the list of variables).

Second, we randomly sort the individuals and match each treated individual to the nearest comparison group individual with a similar propensity score. After matching, our final sample is 699 treated and 699 comparison group youth.[16]

Third, we perform an outcome analysis using this matched sample. We employ logistic regression to estimate the likelihood that a youth will recidivate, conditional on COS participation, as well as other variables included in the propensity score model.

Exhibit 1 reports the means and percentages for all variables used in the analysis for the treated and comparison groups before and after matching. After matching, the two groups were very similar on all observed characteristics.

[16] We use 1:1 nearest neighbor caliper matching without replacement but including ties—i.e., we include all comparison group youth with identical propensity scores that match a given treated individual within a defined minimum distance. However, in practice, we did not have any ties meaning each treated youth only had one comparison group match. We also exclude individuals for whom no good match could be found. More detailed methods for this evaluation are described in the Technical Appendix.

Exhibit 1

Study Group Characteristics

Variable	Before matching		After matching	
	COS	Comparison group	COS	Comparison group
Age	16.07	16.08	16.05	16.04
Percent male	69%	70%	68%	68%
Percent white	75%	73%	77%	75%
Percent black	10%	6%	8%	9%
Percent Latino	8%	17%	9%	10%
Percent other race	8%	5%	6%	7%
Criminal history score[1]	3.55	3.62	3.56	3.57
Social history score[1]	2.95	3.24	3.09	2.95
Whether youth is law abiding (0/1)[1]	70%	75%	73%	71%
Whether youth is anti-social (0/1)[1]	35%	24%	30%	31%
Whether youth demonstrates verbal aggression (0/1)[1]	42%	37%	38%	34%
Whether youth demonstrates physical aggression (0/1)[1]	54%	47%	49%	49%
Whether youth demonstrates violent or sexual aggression (0/1)[1]	20%	19%	21%	19%
Whether youth started program in 2011 (0/1)[1]	42%	51%	45%	41%
County juvenile arrest rate in 2010[2]	1.69	2.05	1.85	1.81
County population aged 0-17 in 2010 (logged)[3]	11.57	10.57	11.33	11.31
Court caseload in 2010[4]	1020	597	844	866
Percent of caseload that is low-risk in 2010[4]	33%	36%	33%	33%
Number of youth	864	3144	699	699

Notes:

[1] These measures come from the juvenile risk assessment; see Barnoski, R. (2004). *Washington state juvenile court assessment manual, version 2.1* (Doc. No. 04-03-1203). Olympia: Washington State Institute for Public Policy.

[2] From the Summary Reporting System obtained from the Office of Financial Management Washington State Criminal Justice Data Book (http://wa-state-ofm.us/CrimeStatsOnline/index.cfm).

[3] Obtained from the Office of Juvenile Justice and Delinquency Prevention "Easy Access to Juvenile Populations" (http://www.ojjdp.gov/ojstatbb/ezapop/).

[4] Calculated from juvenile justice system administrative data obtained from the Administrative Office of the Courts.

III. Evaluation Findings

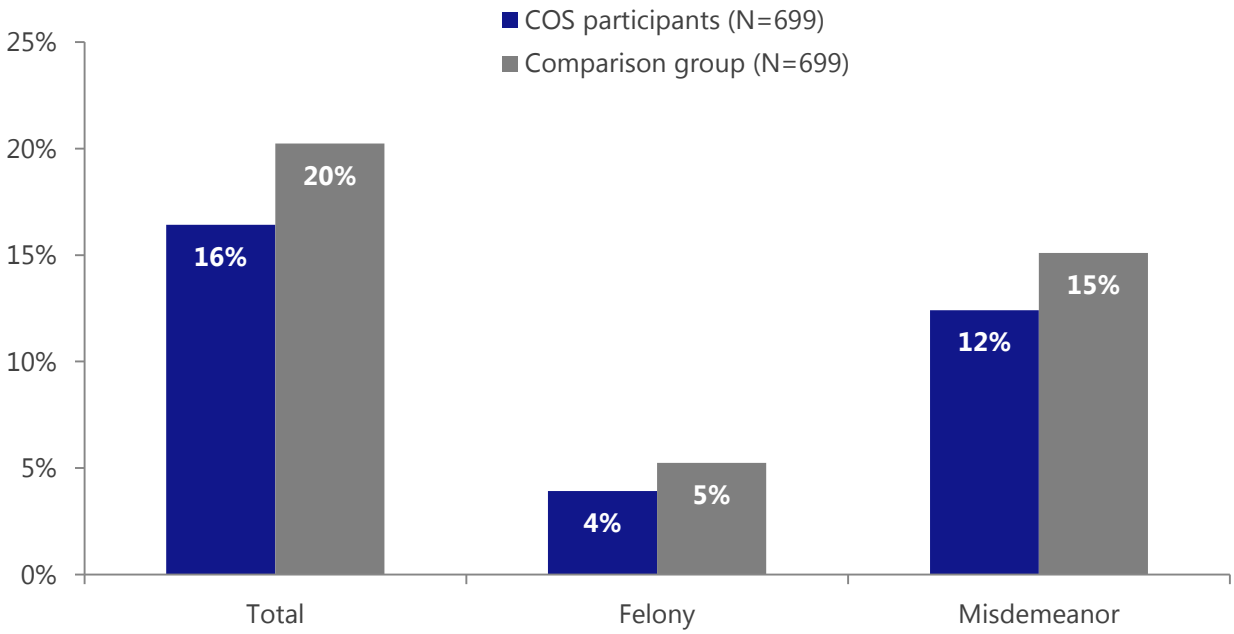
We analyze the effect of COS participation on the following outcomes:

- Total recidivism (any misdemeanor or felony conviction),
- Felony recidivism, and
- Misdemeanor recidivism.[17]

We display our regression-adjusted recidivism rates in Exhibit 2. We find that COS reduces the likelihood of recidivism across all three recidivism measures, although the results are not statistically significant.[18]

Exhibit 2

18-month Adjusted Recidivism Rates across Treatment Status



None of these difference are statistically significant at p <0.10 based on bootstrapped standard errors.

[17] Because only 28 youth had a violent felony conviction, violent felony recidivism is too infrequent to analyze.

[18] We used bootstrapping to arrive at the standard errors for determining statistical significance. Bootstrapped standard errors, as opposed to analytical standard errors, allow us to take into account the fact that the propensity score is estimated in our outcome analysis. Analytic standard errors were smaller than those from bootstrapping; using analytic standard errors would result in statistically significant effects of COS as seen in Exhibit A4 in the Technical Appendix. However, we believe bootstrapped standard errors are appropriate in our main analysis as discussed in the Technical Appendix.

Overall, COS reduces recidivism by four percentage points (16% for the treated group compared to 20% for the comparison group). The effect of COS is larger for misdemeanor recidivism than for felony recidivism.[19]

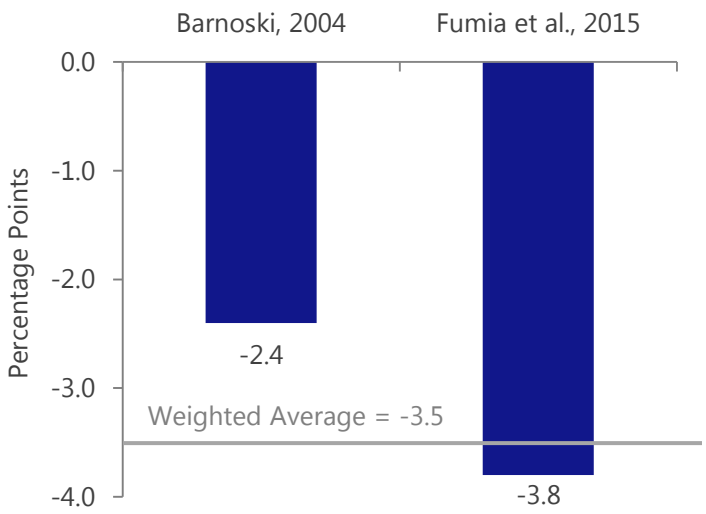
Meta-Analytic Results

Whenever possible, we conduct a systematic review of all rigorous evaluations that measure the impact of a particular program. We then conduct a meta-analysis to determine the average effect of a program given the collective evidence.[20]

We reviewed the literature evaluating COS, and the only two rigorous evaluations of COS as implemented in Washington State are WSIPP's 2004 and 2015 evaluations.[21] Although the program added a quality assurance process after the initial evaluation, implementation has remained generally consistent over time. Thus, we pool the results from both evaluations to arrive at our current estimate for the average effect of COS on recidivism.

Exhibit 4 displays the effects of COS on total recidivism from both evaluations. We find that, on average, COS reduces total recidivism by 3.5 percentage points (from 20% to 16.5%; $p < 0.06$).

Exhibit 4
Effect of COS on Total Recidivism across Studies



[19] We attempted to conduct a subgroup analysis by racial group to determine whether COS is effective in heterogeneous populations. However, we were unable to find a sufficiently matched group for nonwhite treated youth to proceed with the analysis. We also attempted to examine if the delivery format of COS—that is, whether it is delivered in two 6-hour or three 4-hour sessions—impacts the estimated effects of the program. Unfortunately, because of other differences between the courts, we were unable to match treated and comparison groups adequately once we subdivided the data by COS format. Thus, we cannot identify whether effects differ by COS format. However, we are able to conduct correlational analyses with respect to COS delivery format and present that analysis in the Technical Appendix.

[20] In general, we follow the procedures in Lipsey, M.W., & Wilson, D. (2001). *Practical meta-analysis*. Thousand Oaks: Sage Publications. For more information about we compute effect sizes, see our Technical Documentation: http://www.wsipp.wa.gov/TechnicalDocumentation/WsippBenefitCostTechnicalDocumentation.pdf

[21] Although COS was evaluated by Tolan et al. (1987), the program as evaluated in Tolan et al. does not reflect the program as implemented in Washington State. See Barnoski, (2004), p. 12.

IV. Benefit-Cost Analysis

WSIPP also considers the benefits and costs associated with implementing a program. WSIPP's benefit-cost model provides an internally consistent monetary valuation so program and policy options can be compared on an apples-to-apples basis.[22] Our benefit-cost results are expressed with standard financial statistics: net present values and benefit-cost ratios.

In benefit-cost analyses of juvenile justice programs, reductions in recidivism produce benefits to program participants, crime victims, taxpayers, and other people in society. Reductions in recidivism also produce benefits through avoided costs of crime. Crime produces many costs, including those associated with the criminal justice system as well as those incurred by crime victims. When crime is avoided, these reductions lead to monetary savings or benefits to victims and taxpayers. WSIPP's benefit-cost model estimates the number and types of crimes avoided (or incurred) due to the effect of a policy and the monetary value associated with that reduction.

Juvenile offenders who reduce their probability of recidivism can experience increases in high school graduation, which benefits both the offenders themselves through increased employment and others in society through greater tax revenue and other positive "spillover" effects. Higher rates of high school graduation can also lead to changes in healthcare coverage, as those with high school diplomas are more likely to use private or employer-sponsored health insurance rather than publicly-provided healthcare.

Finally, to account for the inherent uncertainty associated with any statistical or benefit-cost analysis, we perform a "Monte Carlo simulation" in which we vary key factors in our calculations. We can then estimate the degree of risk associated with our estimates. More details on our benefit-cost analysis methods can be found in the Technical Documentation.[23]

Exhibit 5 shows the results of our benefit-cost analysis. COS costs $412 per youth (in 2014 dollars).[24] Participation in COS results in total benefits from avoided crime and increased high school graduation of $9,614 shown in Exhibit 5. Thus, we estimate net benefits of $9,202. Our risk analysis indicates that COS will yield positive net benefits 96% of the time.

[22] Washington State Institute for Public Policy (2014). *Benefit-cost technical documentation*. Olympia, WA: Author.

[23] http://www.wsipp.wa.gov/TechnicalDocumentation/WsippBenefitCostTechnicalDocumentation.pdf

[24] Barnoski, (2009).

The legislature has identified a three-tiered classification to identify effective programs for children and youth: evidence-based, research-based, and promising practices. Research-based programs are defined as those that have "some research demonstrating effectiveness, but that does not yet meet the standard of evidence-based practices."[25] Based on this definition and the findings from this evaluation, we define COS as a research-based program. That is, the weight of the evidence indicates a significant reduction in recidivism. Additionally, COS produces cost-beneficial outcomes.

Exhibit 5

Benefits and Costs per Participant for COS vs. Comparison Group in 2014 Dollars

Program cost		
COS participants		
Additional cost per participant for transportation, court services (interpreter services, rent, supplies, etc.), indirect costs (quality assurance and administrative), oversight, and case management		$412
Comparison group costs		$0
	(1) Net COS cost	**-$412**
Recidivism effects		
Decreased taxpayer costs due to decreased recidivism		$1,318
Decreased crime victim costs due to decreased recidivism		$3,484
Health care-related effects		
Increased healthcare insurance costs to participants due to moving from public to employer or private insurance		-$38
Decreased healthcare insurance costs to taxpayers due to movement from public to employer or private insurance		$298
Increased costs to private or employer-sponsored insurance programs		-$217
High school graduation effects		
Increased income to participants due to increased labor market participation		$2,168
Increased tax revenue to taxpayers due to increased labor market participation		$925
Positive externalities to society due to greater number of high school graduates		$1,071
Deadweight cost of taxation		$604
	(2) Total benefits	**$9,614**
Bottom line:		
Net benefits (cost) per participant	(3) Net (benefits – costs)	**$9,202**
Benefit-to-cost ratio		**$23.34**
Probability of positive net benefits (risk analysis)		**96%**

[25] RCW 71.36.010 requires that to be designated as evidence-based, a program must demonstrate effectiveness in "multiple site random controlled trials." Neither of the two evaluations included in our review of COS were randomized trials.

Appendix

A. I. Study Group Selection & Matching Procedures

In an ideal research design, offenders eligible for COS would be randomly assigned to COS or an untreated comparison group. With a successfully implemented random assignment, any observed differences in recidivism could be attributed to the effect of COS. Unfortunately, as is the case in many real world settings, random assignment was not possible for this evaluation.

Instead, we use observational data and rely on a quasi-experimental research design. Unlike random assignment, this type of design cannot eliminate the risk that selection bias or unobserved factors may threaten the validity of the findings. For example, juvenile probation counselors (JPCs), parents, or the youth themselves can base participation on the youth's likelihood of success or motivation. If youth that participate in COS are more motivated, this unobserved factor would bias the results in favor of the treated group. However, if youth in COS are referred to the program because they are perceived as somehow worse off than non-treated youth, this selection would bias the results toward the comparison group.

To infer causality from this quasi-experimental study, selection bias must be minimized. To do so, we implement a variety research design methods and statistical techniques that provide the ability to test the sensitivity of our findings. In this section of the Technical Appendix, we describe the study groups and statistical methods we use to arrive at estimates of the effects of COS.

Study groups

We draw our COS participant pool from youth that started COS between 1/1/2011 and 12/31/2012 as reported by the juvenile courts that offered COS at the time: Clallam, Cowlitz, King, Kitsap, Snohomish, Spokane, and Whatcom ("COS courts"). Comparison group youth are drawn from the population of low-risk youth assessed between 1/1/2011 and 12/31/2012 in courts that did not offer COS during that time (i.e., "non-COS courts").

We draw the comparison group from a pool of low-risk offenders in non-COS courts to ensure that program availability is the primary driver of comparison group assignment. That is, youth from non-COS courts do not participate in the program primarily because it is not offered. If we draw a comparison group from COS courts, however, we cannot easily determine the reasons that an untreated youth did not participate. Youth may not participate in COS for many reasons. If the researcher cannot observe the

reasons for nonparticipation, however, then selection bias would be a serious concern. However, youth from non-COS courts do not have the option to participate and so we can infer that their nonparticipation has to do with their location rather than unobserved characteristics.

Drawing comparison youth from non-COS courts does not completely eliminate selection bias. Those that participate in the treated group may still be more motivated than the average comparison group youth, for example. In other words, using youth from non-COS courts as the comparison group allows us to mitigate selection bias among nonparticipants, but we still cannot prevent selection into the treated group on the part of JPCs, youth, parents, lawyers, etc. Thus, it is important we are able to balance the treated group with comparison youth from non-COS courts on all observables.

Propensity Score Matching

While using a comparison group from other locales and selection bias pose possible threats to the validity of a study, we attempt to minimize these influences using propensity score matching. Propensity score matching allows us to match treated individuals with similar comparison group individuals to obtain balance on observed covariates. This method has many benefits over standard regression analysis, which is often used to control for differences between treated and comparison groups.

First, the outcome plays no part in matching the treated and comparison groups. This emulates an experimental design by separating the research design stage—where we test various matching procedures to obtain a sufficiently matched sample—from the analysis stage—where we estimate the effect of the treatment using our matched sample. Second, matching can limit the importance of functional form in regression analysis.[26] Third, by imposing common support restrictions, we ensure that the comparison group does not differ substantially in their likelihood to participate in COS, i.e. we are not comparing treated youth to youth who we would never expect to participate in COS. Finally, by conducting a logistic regression on the matched sample using the covariates from the matching model, we further reduce any residual bias that may remain after matching and account for any correlation between matched pairs.

We match on the logit of the propensity score defined in the equation below:[27]

$$(1)\ \ logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \ log\left(\frac{e^{(\alpha+\beta_1 X_{1i}+\cdots+\beta_k X_{ki})}}{1+e^{(\alpha+\beta_1 X_{1i}+\cdots+\beta_k X_{ki})}}\right) = \ \alpha+\beta_1 X_{1i}+\cdots+\beta_k X_{ki}$$

In equation (1), p_i represents the probability that individual i receives treatment (i.e., the propensity score), α represents the intercept of the model, β_j represents the parameter of the model for covariate X_j, and e is the base of the natural logarithm. We match on the logit of the propensity score as suggested by Rosenbaum and Rubin and because we calculate the optimal caliper using the logit of the propensity score as discussed below. [28]

[26] Ho, D.E., Imai, K., King, G., & Stuart, E.A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, *15*(3), 199-236.

[27] The propensity score was estimated using the pscore command and the matching procedures were performed using psmatch2 in STATA.

[28] Rosenbaum, P.R., & Donald B.R. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33-38 and Austin, P.C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, *10*(2), 150–161.

Exhibit A1 below reports the results from the coefficients from the first stage model estimating the likelihood of COS participation. We control for demographic characteristics, criminal history and social history scores, and behavior variables from the assessment data. Additionally, to address the fact that treated and comparison group youth come from different locales, we control for various county and court characteristics measured in 2010 including county juvenile aged population (0-17), juvenile arrest rate, court caseload size, and percent of caseload that was low-risk.

Exhibit A1

Logit Model Estimating the Likelihood of COS Participation

Covariate	Coefficient	SE
Age	-0.140 ***	0.031
Male (0/1)	-0.141	0.096
Black (0/1)#	-0.380 **	0.163
Latino (0/1)#	-0.830 ***	0.154
Other race (0/1)#	0.218	0.181
Criminal history score	-0.228 ***	0.028
Social history score	-0.180 ***	0.028
Whether youth is law abiding (0/1)	-0.129	0.117
Whether youth is anti-social (0/1)	0.471 ***	0.112
Whether youth demonstrates verbal aggression (0/1)	-0.007	0.114
Whether youth demonstrates physical aggression (0/1)	0.551 ***	0.112
Whether youth demonstrates violent or sexual aggression (0/1)	-0.324 ***	0.117
Whether youth started program in 2011 (0/1)	-0.547 ***	0.088
County juvenile arrest rate in 2010	-0.506 ***	0.085
County population aged 0-17 in 2010 (logged)	1.313 ***	0.120
Court caseload in 2010	-0.001 ***	0.000
Percent of caseload that is low-risk	0.007 ***	0.004
Constant	-0.001 ***	0.000
N	4008	
Pseudo R2	0.20	
AUC	0.79	

Notes:
Stars indicate statistical significance; * p< 0.1; ** p<0.05; *** p<0.01
#Reference group is white youth.

Our preferred matching procedure for the main analysis is 1:1 nearest neighbor matching without replacement with a caliper of 0.2 times the standard deviation of the logit of the propensity score.[29] We also allow ties, meaning that a treated youth will be matched with all closest comparison group youth with identical propensity scores. Using 1:1 matching can reduce the bias between the treated and comparison groups by only matching treated individuals with the most similar comparison group individual. By using a caliper, comparison group matches must fall within the caliper distance from a treated individual to be included. The caliper ensures that treated individuals are not matched with comparison group youth that are too dissimilar and also ensures sufficient overlap between the treated and comparison groups (i.e., a common support region).

[29] Austin, (2011).

However, 1:1 caliper matching without replacement can also lead to a smaller common support region by excluding treated individuals for whom no good match can be found.[30] Furthermore, the variance of the estimated effect is higher with 1:1 matching leading to larger confidence intervals. Thus, we tested numerous matching procedures and chose 1:1 matching without replacement based on the balance achieved. We examine the differences across matching methods in the estimated effects of COS in Section A.III. of the Technical Appendix.

Other matching procedures include 1:1 caliper matching with replacement where a comparison group youth can be used more than one time. We also tested numerous methods that allow for more than one match per treated individual. These methods can be more efficient as they have a larger sample size. We tested 1:3 caliper matching where each treated individual is matched to three comparison group youth; radius matching where each treated individual is matched to all untreated individuals within the caliper (note that we use a smaller caliper for radius matching); and kernel matching where treated individuals are matched to each untreated youth within a bandwidth, and untreated youth are weighted based on the distance from the treated individual on the propensity score.

We used various diagnostics to determine the extent to which the propensity score matching improved balance between the treated and comparison groups. A common measure of balance is the standardized difference (or bias) calculated as the difference in the mean/proportion for the treated and comparison groups divided by the pooled standard deviation for each covariate prior to matching. This measure is preferred to traditional t-tests as the standardized difference is not influenced by the study's sample size. Additionally, t-tests are used for making inferences about a population based on a sample; balance, on the other hand, is an in-sample property. Standardized bias values greater than 0.10 usually indicate moderate imbalance while greater than 0.25 indicates severe imbalance.[31] Exhibit A2 displays the percent standardized bias for each covariate in the propensity score model before and after matching as well as the p-value as a reference. After matching, most differences were greatly reduced and the bias for all covariates fell below the 0.10 threshold.

[30] With our preferred matching method, 165 treated participants were excluded because no good matches were found. We examined the characteristics of these excluded participants and found that they were more often male and nonwhite and had more behaviors associated with criminality such as less likely to be law abiding and more likely anti-social. However, they also tended to have lower social history scores and live in areas with lower juvenile arrest rates. Overall, excluded youth also had lower raw recidivism rates. Results using 1:3, radius, and kernel matching methods show the estimated effects if these youth were not excluded. The results using other matching methods tend to be slightly smaller than using our preferred method, although all estimates are generally in the same range (see Exhibit A5).

[31] Austin, P.C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine, 28*(25), 3083-3107 and Stuart, E.A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *25*(1), 1–21.

Exhibit A2

Matched Study Groups Characteristics

Variable	After matching			Absolute standardized difference (d)	
	COS	Comparison group	p-value	Before matching[1]	After matching
Propensity score	0.32	0.32	0.51	118.80	3.35
Age	16.05	16.04	0.85	0.48	1.01
Percent male	68%	68%	0.73	0.76	1.86
Percent white	77%	75%	0.35	5,91	4.88
Percent black	8%	9%	0.64	15.43	2.71
Percent Latino	9%	10%	0.64	28.64	2.19
Percent other race	6%	7%	0.58	11.18	2.97
Criminal history score[2]	3.56	3.57	0.88	3.37	0.76
Social history score[2]	3.09	2.95	0.15	16.10	7.43
Whether youth is law abiding (0/1)[2]	73%	71%	0.55	11.97	3.22
Whether youth is anti-social (0/1)[2]	30%	31%	0.91	23.86	0.63
Whether youth demonstrates verbal aggression (0/1)[2]	38%	34%	0.20	9.48	6.74
Whether youth demonstrates physical aggression (0/1)[2]	49%	49%	1.00	14.56	0.00
Whether youth demonstrates violent or sexual aggression (0/1)[2]	21%	19%	0.46	1.19	3.96
Whether youth started program in 2011 (0/1)[2]	45%	41%	0.12	17.73	8.35
County juvenile arrest rate in 2010[3]	1.85	1.81	0.28	51.59	4.98
County population aged 0-17 in 2010 (logged)[4]	11.33	11.31	0.77	91.75	1.40
Court caseload in 2010[5]	844.21	866.18	0.39	74.56	3.87
Percent of caseload that is low-risk in 2010[5]	33.13	32.83	0.64	24.38	2.32
Number of youth	699	699			

Notes:

[1] Red text indicates severe imbalance, |d| > 0.25; orange text indicates moderate imbalance, |d| >0.1.

[2] These measures come from the juvenile risk assessment developed by the Washington Association of Juvenile Court Administrator and the Washington State Institute for Public Policy.

[3] From the Summary Reporting System obtained from the Office of Financial Management Washington State Criminal Justice Data Book (http://wa-state-ofm.us/CrimeStatsOnline/index.cfm).

[4] Obtained from the Office of Juvenile Justice and Delinquency Prevention "Easy Access to Juvenile Populations" (http://www.ojjdp.gov/ojstatbb/ezapop/).

[5]Calculated from juvenile justice system administrative data obtained from the Administrative Office of the Courts.

Other diagnostic tests include the mean and median standardized bias across all covariates, Rubin's B which is the standardized difference in the mean of the linear prediction of the propensity score, and Rubin's R which is the ratio of variance of the treated and comparison group for the linear prediction of the propensity score.[32] Average and median bias below 0.25 indicate relatively strong balance overall. Rubin's B and R values should be less than 0.25 and between 0.5 and 2, respectively, to indicate sufficient balance.[33]

[32] These diagnostics were calculated using the pstest command in STATA.

[33] Rubin, D.B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*(3-4), 169-188.

Exhibit A3 reports balance measures from other matching methods. Note that all matching models show improvement across all balance diagnostics from the unmatched sample except on Rubin's R where 1:3, radius, and kernel matching all have Rubin's R values outside the preferred range.

Exhibit A3

Overall Model Balance across Different Matching Methods

Matching method	Treated N	Comparison group N	Rubin's R	Rubin's B	Median bias	Mean bias
Unmatched	864	3,144	0.92	115.83	14.99	22.38
1:1 nearest neighbor without replacement[1]	**699**	**699**	**1.44**	**22.96**	**2.84**	**3.29**
1:1 nearest neighbor with replacement[1]	850	536	1.85	33.60	6.78	7.05
1:3 nearest neighbor[1]	850	1,172	2.52	30.09	5.69	6.15
Radius matching[2]	850	3,144	3.52	26.91	3.61	5.59
Kernel Matching[3]	850	3,144	3.32	27.49	3.50	5.70

Notes:

Bolded text identifies chosen matching method.

[1]We use a caliper equal to 0.2 times the logit of the propensity score.

[2]We use a caliper equal to 0.05 times the logit of the propensity score.

[3]We use a bandwidth of 0.19. We calculate the optimal bandwidth using leave-one-out cross validation.[34]

[34] Black, D.A., & Smith, J.A. (2004). How robust is the evidence on the effects of college quality? Evidence from matching. *Journal of Econometrics, 121*(1), 99-124.

A. II. Methods to Estimate the Effects of COS

Logistic Regression Analysis on Full (Unmatched) Sample

We begin our outcome analysis using traditional multivariate logistic regression analysis on the full (i.e. unmatched) sample. Regression analysis allows us to control for observed covariates in estimating the treatment effect. However, regression analysis has several limitations. First, regression analysis can only control for observed factors. Second, if treated and comparison group covariate distributions do not overlap, then any causal inferences for regions with few treated or control group members must be based on extrapolation, leading to less precise estimates. Third, to approximate an experimental design, the research design stage of an evaluation should be separate from the outcome analysis stage. With standard regression analysis, the outcome of interest is necessarily part of the regression model and determining model fit requires repeatedly estimating the treatment effect.[35] This can lead to model selection based on the observed treatment effect and also suffers from the multiple comparisons problem, where the likelihood of finding a statistically significant result increases with the number of statistical tests performed. Finally, regression analysis requires making assumptions about functional form, which can increase bias if the wrong functional form is used.

While regression analysis has several limitations, it can outperform matching methods if important unobserved covariates are omitted from the analysis. In this case, regression analysis will produce a less biased estimate than propensity score matching. For this reason, we first estimate the relationship between COS participation and recidivism using standard logistic regression. Row (7) of Exhibit A5 reports the regression-adjusted recidivism rates for the unmatched sample. The effects using standard logistic regression indicate that COS participation reduces recidivism by about 4.7 percentage points, a slightly larger reduction than in our chosen matched sample of 3.8 percentage points (Row (8)). Generally the results from the standard regression analysis do not differ substantially from the effects using the various matching methods reported in rows 8-12 of Exhibit A5.

Outcome Analysis: Logistic Regression on Matched Sample

Our preferred analysis uses logistic regression on the matched sample to estimate the effect of COS on total, felony, and misdemeanor recidivism. Our outcome model uses the same covariates included in the matching model and reported in Exhibit A4 below. The logistic regression model is weighted using the normalized weight based on the number of times an untreated youth was matched to a treated individual. Exhibit A4 reports the results from our preferred model that uses the matched sample from 1:1 nearest neighbor caliper matching without replacement.

[35] Rubin, D.B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in medicine*, *26*(1), 20-36.

Exhibit A4

Logistic Regression Estimating Effect of COS on Recidivism

(COS participant N = 699, Comparison group N = 699)

Covariate	Total recidivism		Felony recidivism		Misdemeanor recidivism	
	Odds ratio	p-value[#]	Odds ratio	p-value[#]	Odds ratio	p-value[#]
COS participation	0.763	0.014	0.725	0.075	0.790	0.066
Age	0.963	0.369	1.157	0.180	0.909	0.006
Male (0/1)	1.365	0.165	6.403	0.000	0.958	0.848
Black (0/1) [##]	1.682	0.015	2.584	0.000	1.216	0.487
Latino (0/1) [##]	1.044	0.859	0.902	0.714	1.091	0.783
Other race (0/1) [##]	1.147	0.544	1.126	0.764	1.143	0.641
Criminal history score	1.113	0.051	1.150	0.187	1.090	0.103
Social history score	1.256	0.000	1.168	0.066	1.256	0.000
Whether youth is law abiding (0/1)	0.986	0.929	0.704	0.253	1.117	0.507
Whether youth is anti-social (0/1)	0.789	0.262	0.757	0.380	0.824	0.467
Whether youth demonstrates verbal aggression (0/1)	1.335	0.076	0.909	0.739	1.456	0.032
Whether youth demonstrates physical aggression (0/1)	1.408	0.014	1.410	0.275	1.372	0.065
Whether youth demonstrates violent or sexual aggression (0/1)	1.017	0.899	1.533	0.258	0.852	0.383
Whether youth started program in 2011 (0/1)	1.168	0.145	1.198	0.227	1.139	0.291
County juvenile arrest rate in 2010	0.785	0.051	0.871	0.568	0.804	0.080
County population age 0-17 in 2010 (logged)	0.857	0.354	0.823	0.276	0.946	0.779
Court caseload in 2010	1.000	0.939	1.000	0.115	1.000	0.531
Percent of caseload that is low risk in 2010	0.999	0.938	0.977	0.003	1.007	0.458
Constant	0.739	0.887	0.005	0.048	0.532	0.783
Pseudo-R2	0.062		0.121		0.047	

Notes:

[#]P-values based on analytical standard errors clustered at court level rather than bootstrapped standard errors. The analytical standard errors are smaller than those from bootstrapping yielding in statistically significant results; however, our main findings use bootstrapped standard errors.

[##]Reference group is white youth.

A. III. Sensitivity Analyses

We tested the sensitivity of our estimates of the effect of COS to different matching procedures; the results are reported in Exhibit A5 below. Rows (1) – (6) report difference in mean recidivism rates on the unmatched and matched samples without adjusting for covariates. Rows (7) – (12) show the regression-adjusted effects where we control for covariates using the unmatched and matched samples.

Regression-adjusted effects of COS are relatively consistent regardless of matching method, although it is important to note that these sensitivity analyses were performed after choosing our preferred matching method to maintain separation between the research design and analysis stages of the study. We find that COS reduces total recidivism by about 2 to 4 percentage points regardless of which matching method is used. Furthermore, COS participation reduces misdemeanor recidivism by 2 to 3 percentage point in recidivism rates and by about 1 percentage point for felony recidivism across methods.

Exhibit A5

Effects of COS Using Various Matching Methods[1]

	Matching method	Total recidivism				Felony recidivism				Misdemeanor recidivism			
		COS	Comparison	Percentage point difference[2]	SE[3]	COS	Comparison	Percentage point difference[2]	SE	COS	Comparison	Percentage point difference[2]	SE[3]
	Raw recidivism rates												
(1)	Unmatched	15.9%	22.1%	-6.2***	0.016	3.8%	5.7%	-1.9**	0.009	12.1%	16.3%	-4.2***	0.014
(2)	1:1 nearest neighbor without replacement	16.6%	20.0%	-3.4*	0.021	4.0%	5.2%	-1.1	0.011	12.6%	14.9%	-2.3	0.018
(3)	1:1 nearest neighbor with replacement	16.0%	17.9%	-1.9	0.024	3.9%	4.4%	-0.5	0.010	12.1%	13.5%	-1.4	0.023
(4)	1:3 nearest neighbor	16.0%	18.8%	-2.8	0.021	3.9%	4.2%	-0.3	0.010	12.1%	14.6%	-2.5	0.019
(5)	Radius matching	16.0%	20.4%	-4.4***	0.016	3.9%	5.4%	-1.5*	0.009	12.1%	15.0%	-2.9**	0.014
(6)	Kernel Matching	16.0%	19.9%	-3.9**	0.018	3.9%	5.4%	-1.5*	0.009	12.1%	14.9%	-2.8*	0.016
	Regression adjusted recidivism rates												
(7)	Unmatched	17.1%	21.7%	-4.7***	0.016	3.8%	5.8%	-1.9**	0.008	13.2%	16.0%	-2.7*	0.015
(8)	1:1 nearest neighbor without replacement	16.4%	20.2%	-3.8	0.034	3.9%	5.2%	-1.3	0.027	12.4%	15.1%	-2.7	0.031
(9)	1:1 nearest neighbor with replacement	15.8%	18.2%	-2.4	0.033	3.8%	4.4%	-0.6	0.031	11.9%	13.7%	-1.8	0.027
(10)	1:3 nearest neighbor	15.8%	19.1%	-3.2	0.023	3.8%	4.3%	-0.5	0.017	12.0%	14.8%	-2.8	0.022
(11)	Radius matching	16.3%	20.1%	-3.7	0.026	3.8%	5.5%	-1.7	0.013	12.4%	14.6%	-2.2	0.020
(12)	Kernel Matching	16.1%	19.9%	-3.8	0.025	3.8%	5.2%	-1.4	0.014	12.2%	14.8%	-2.5	0.020

Notes:

[1]Unweighted sample sizes are as follows:
Unmatched raw (Treated N = 864, Comparison N = 3144); Unmatched regression adjusted (Treated N = 864, Comparison N = 3144); 1:1 Nearest neighbor without replacement (both raw and regression adjusted ("both"), Treated N = 699, Comparison N = 699); 1:1 Nearest neighbor with replacement (both, Treated N = 850, Comparison N = 536); 1:3 Nearest neighbor (both, Treated N = 850, Comparison N = 1172); Radius matching (both, Treated N = 850, Comparison N = 3144); Kernel matching (both, Treated N = 850, Comparison N = 3144).

[2]Stars indicate statistical significance; * p< 0.1; ** p<0.05; *** p<0.01.

[3]Standard errors for nearest neighbor raw recidivism rates obtained using Abadie & Imbens formula in psmatch2 program of STATA. Standard errors for other matching methods and all regression adjusted results obtained through bootstrapping.

[4]Raw recidivism rates are differences in mean recidivism rates for treated and comparison groups without regression adjustment. Matching on covariates was still used to obtain a matched raw recidivism rate.

In addition to examining the effects under various matching methods, we also test the sensitivity of our results to various specifications of the propensity score models, caliper selection, and bandwidths. Our findings were generally robust to sample selection, matching methods, and model specifications. However, three sensitivity analyses warrant further discussion: (1) comparison group selection, (2) timing of the recidivism measure, and (3) standard error estimation.

Comparison Group

We found that our results seem sensitive to the courts from which we draw the comparison group. While our final analysis used a comparison group drawn from non-COS courts, we also explored using a comparison group from within COS courts. This comparison group would be attractive because they would face the same court factors as those in the treated group. When we used untreated youth from within COS as the comparison pool, we found statistically insignificant effects that were very close to zero.

While untreated youth from within COS courts face the same court and county factors as treated youth, we chose our final comparison group from non-COS courts because we could not fully explain why untreated youth in COS courts did not receive COS when the program was available to them. In other words, we were more concerned about the potential for selection bias in the comparison group when drawing that comparison group from COS courts. Additionally, while some data exist on reasons why a youth did not start COS, this information was missing in about 30% of cases. In another 40% of cases, untreated youth did not start COS because they were referred to other programs suggesting that youth in COS courts may receive alternative treatment in lieu of COS.

As an additional sensitivity test, we attempted to limit the COS court comparison group to those that did not participate due to program availability or timing (e.g. program was full or inaccessible or there was not enough time on probation to attend). Youth who do not start solely for reasons of program availability or timing would not have the same selection concerns as youth that do not participate due to refusal or selection on the part of JPCs, judges, or lawyers. However, only 35 youth did not participate in the program for reasons of timing or availability, which is too few to analyze.

By using a comparison group from courts that did not offer COS, we cannot determine the extent to which the effect we observe is partially due to characteristics that differ between the courts that chose to implement COS and those that did not. Indeed, we also find that comparison group youth from within COS courts are less likely to recidivate than comparison group youth from non-COS courts, meaning that there may be important court or county characteristics that at least partially explain the treated group's lower recidivism rates.

Ideally, we would conduct a fixed effects regression which would control for unobserved characteristics that differ across courts, but because court is perfectly predictive of the treatment, we cannot include it in the propensity score model and will not have overlap between the treated and comparison groups in the court fixed effects in the outcome analysis. We do use county and court characteristics in both the propensity score model and outcome analysis to account for some observable differences, but we cannot account for unobserved factors that might differ across courts. Although we are confident in our findings, we do acknowledge the possibility that unobserved court characteristics may at least partially explain the effects of COS reported here.

Timing of Recidivism

Recidivism is defined as any offense committed in the 18 months after program start for participants or after assessment for comparison group youth that resulted in a Washington State conviction.[36] In addition to the follow-up period, time is needed to allow an offense to be processed in the criminal justice system. The criminal justice process also includes the adjudication period—the time period between the date recorded for the commission of a subsequent offense and the resulting conviction for that offense. This analysis allows for a 6-month adjudication period. Although this is shorter than the suggested 12-month adjudication period, we feel six months is an adequate amount of time for low-risk youth to account for court processing as it is typically more serious offenses that require longer processing times.[37]

In our sample, COS participants could wait up to six months from the time of assessment before starting the program, although on average the wait is about six weeks. In any case, offenses occurring after assessment but before starting the program are not considered recidivism for the treated group in our primary recidivism measure. For the comparison group, on the other hand, offenses occurring in the 18 months after assessment are considered recidivism. If youth reoffend shortly after assessment, our recidivism measure could underestimate the number of recidivism events in the treated group.

To determine whether our recidivism timing could impact our results, we conduct a sensitivity analysis using an alternative measure of recidivism where we define any offenses occurring after assessment and prior to program start as recidivism for the treated group. Because some COS participants recidivate prior to starting COS, we find that differences between the COS participants and the comparison group are slightly smaller when using the alternative recidivism measure. For example, the regression-adjusted total recidivism rate using the alternative recidivism measure and our preferred matching procedures is 17% for the treated group and 20% for the comparison group. As illustrated in Exhibit 2 in the main text, when using our primary recidivism measure these rates are 16% and 20%, respectively. Thus, we find a small but substantively insignificant difference in whether we measure recidivism for the treated group beginning at the time of assessment or at program start date.

Standard Errors

In propensity score matching, the problem of obtaining correct standard errors often arises. Analytical formulas for the standard error such as those from logistic regressions ignore the error associated with estimating the propensity score and the correlation of the matched sample.[38] Thus, the analytical standard errors based on matched data may be inaccurate. To address this issue, we use bootstrapping methods to estimate the standard error of the regression-adjusted effect of COS.[39] Bootstrapping means repeatedly drawing *N* random samples from the matched sample with replacement and computing an effect of COS for each sample using the methods described in the outcome analysis section above. Then, the variance of

[36] Barnoski, (1997), pg. 2.
[37] Barnoski, (1997), pg. 4.
[38] Hill, J. (2008). Discussion of research using propensity-score matching: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. *Statistics in Medicine*, *27*(12), 2055-2061.
[39] The effects of COS from the nearest neighbor matched samples that are not regression adjusted use an analytical formula for the standard errors derived by Abadie & Imbens (2006) to estimate the standard errors for nearest neighbor matching and bootstrapping for radius and kernel matching. Research indicates that bootstrapping with kernel or radius matching is probably appropriate as issues arise when using a fixed number of untreated matches. Abadie, A., & Imbens, G.W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1), 235-267.

the effect of COS is measured by estimating the variance in the estimated effects of COS across the *N* samples.

Research suggests that bootstrapping standard errors for matched data may only be appropriate in some situations. First, bootstrapping can only be used for population inference rather than in-sample estimates.[40] Second, when performing regression analyses on matched data, it may be unnecessary to employ bootstrapping when the regression analysis includes the covariates in the matching model. The correlation caused by the matched sample design will already be accounted for by regressing the outcome on the treatment and the covariates used in the matching model.[41] Finally, bootstrapping may be inappropriate for nearest neighbor matching with replacement,[42] although these concerns do not apply to matching without replacement,[43] which is our chosen method for this analysis. Given the tradeoffs between underestimating standard errors and using inappropriate methods for correction, we also examined the sensitivity of our conclusions using analytical standard errors.

Analytical standard errors are much smaller than those obtained through bootstrapping (e.g. see Exhibit A4 where we report p-values based on analytical standard errors and find a statistically significant effect of COS). Using analytical standard errors would yield statistically significant effects of COS for total and misdemeanor recidivism for all nearest neighbor matching methods, while none of the effects are significant when using bootstrapped standard errors. Because bootstrapping does account for the error associated with estimating the propensity score and is appropriate with nearest neighbor without replacement and for matching methods that do not use a fixed number of untreated matches such as kernel or radius matching, we report bootstrapped standard errors in most of the evaluation (with the exception of Exhibit A4). It is important to note that only the standard errors and associated confidence intervals are impacted by the estimation of the standard errors; the estimated effects of COS and our meta-analytic findings would not change regardless of the methods used to estimate the standard errors.

[40] Austin, P.C., & Small, D.S. (2014). The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in medicine*, *33*(24), 4306-4319.
[41] Ho et al., (2007) and Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.
[42] Abadie & Imbens, (2006) and Abadie, A., & Imbens, G.W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, *76*(6), 1537-1557.
[43] Austin & Small, (2014).

A. IV. Does the Number of COS Sessions Matter?: A Correlational Analysis

The standard COS format consists of two 6-hour sessions ("2-session COS"), which is used by Clallam, King, and Snohomish courts (about 51% of COS participants). Cowlitz, Kitsap, Spokane, and Whatcom courts, however, generally offer three 4-hour sessions ("3-session COS"). Providers have interest in understanding whether effects of COS vary across these formats. The best way to estimate a causal effect of format would be to assign courts or youth randomly to 2- or 3-session formats. This method would ensure that selection into a particular type of format does not occur.

Without random assignment, we again should be concerned about selection bias. Courts that choose to implement 2-session COS may differ in important ways from courts that implement 3-session COS. For example, courts that choose a 3-session COS format may do so because they observed that many of the youth in their courts worked and were often unable to attend 6-hour sessions. These courts might choose the 3-session format to make it easier for these working youth to attend COS; however, these youth with jobs may also be also be less likely to recidivate because they have less free time and are more motivated to stay out of trouble. Thus, courts that have the 3-session format may have lower recidivism rates not because the 3-session format is better in general but because they have more employed youth participating in COS. In this case, selective implementation of 3-session COS would bias the results upward making 3-session COS seem more effective when in actuality, youth employment would be the main driver of the effects.

Given that the potential for selection bias persists when examining the effectiveness of COS format, we attempted to minimize the observed bias through propensity score matching. Unfortunately, we were unable to obtain strong matching on covariates using various samples and specifications limiting our ability to make any causal conclusions about the effect of COS format on recidivism. We attempted to test whether the effects vary between 2- and 3-session COS directly by comparing courts that use two 6-hour sessions to those that use three 4-hour sessions and indirectly by separately comparing each format to youth from non-COS courts and then testing whether these effects differ across format.

Both the direct and indirect comparisons require subdividing the treated youth into 2- or 3-session subgroups. In the full analysis of the effect of COS, 2- and 3-session courts are pooled. Once we subdivide the courts by COS format, we were unable to achieve sufficient balance on covariates after matching to continue with the analysis even after including interaction and higher order terms. While propensity score matching does not necessarily yield causal estimates, we believe it can approximate experimental methods in some circumstances while standard regression cannot. Thus, without the ability to perform propensity score matching, we do not feel confident in making causal inferences about the effects of 2- versus 3-session COS.[44]

Although the lack of matching prevents us from providing causal estimates of the effect of two versus three COS sessions on recidivism, we can still provide some information about the relationship between COS format and recidivism using standard logistic regression on the full (i.e. unmatched) sample. Exhibit A6 presents the results of two models regressing total recidivism on COS format and other covariates. The sample in column (1) consists of only COS youth and regresses recidivism on the full set of control

[44] We could get sufficient balance when comparing 3-session COS to youth from non-COS courts; however, this analysis does not allow for a comparison of COS format as we would also need to estimate the effect of 2-session COS compared to non-COS courts. Additionally, with common support restrictions, we could only obtain sufficient balance when we excluded more than 50% of the 2-session COS participants. Thus, we were not confident that results would be generalizable beyond a specific subgroup of youth in the region of common support.

variables. The independent variable of interest is "2-session COS" which is an indicator variable that equals 1 if a youth comes from 2-session COS court and 0 if a youth is from a 3-session COS court. Thus, results in column (1) directly compare 2-session and 3-session COS formats. The marginal effect of 2-session COS in column (1) represents the difference in the predicted recidivism rate between youth from a 2-session COS court and youth from a 3-sessoin COS youth. Youth that participate in 2-session COS have recidivism rates that are 0.8 percentage points lower than youth that participate in 3-session COS. However, this difference is not significant at any standard level (p=0.600).

Exhibit A6

Change in the probability of total recidivism for 2-session and 3-session COS

Dependent variable: Total recidivism	(1)		(2)	
Covariate	Marginal effect	SE	Marginal effect	SE
2-session COS	-0.008	0.015	-0.035	0.021
3-session COS	-	-	-0.060***	0.021
Age	0.006	0.007	-0.011*	0.004
Male (0/1)	0.052***	0.014	0.041*	0.014
Black (0/1)	0.075***	0.011	0.106***	0.024
Latino (0/1)	0.049*	0.026	0.062***	0.017
Other race (0/1)	0.043*	0.022	0.062***	0.026
Criminal history score	0.010	0.007	0.018***	0.004
Social history score	0.022***	0.004	0.029***	0.004
Whether youth is law abiding (0/1)	-0.016	0.017	0.009	0.017
Whether youth is anti-social (0/1)	0.000	0.019	0.011	0.016
Whether youth demonstrates verbal aggression (0/1)	0.024	0.024	0.022	0.016
Whether youth demonstrates physical aggression (0/1)	0.001	0.016	0.026	0.015
Whether youth demonstrates violent or sexual aggression (0/1)	-0.017	0.019	0.011	0.016
Whether youth started program in 2011 (0/1)	0.015	0.013	0.005	0.013
County juvenile arrest rate in 2010	-0.042**	0.017	-0.014	0.010
County population aged 0-17 in 2010 (logged)	-0.007	0.015	0.037	0.006
Court caseload in 2010	0.000***	0.000	0.000**	
Percent of caseload that is low risk in 2012	0.000	0.001	0.001**	0.006
Number of youth	1786		4008	
Pseudo-R2	0.05		0.04	
AUC	0.66		0.65	

Notes:

Robust standard errors reported clustered at the court level.

Column (2) indirectly examines the relationship between COS format and recidivism. The sample used in the regression reported in column (2) includes all COS youth (2- and 3-session) and a comparison group of youth from non-COS courts. Here, the 2-session COS variable indicates youth participating in 2-session COS have recidivism rates that are 6 percentage points lower than youth in non-COS courts controlling for the other variables in the model, while youth in 3-session COS have recidivism rates that are 3.5 percentage points lower than youth in non-COS courts. An indirect comparison suggests that 3-session

COS youth have recidivism rates that 2.5 percentage points lower than youth in 2-session COS court (-0.06 – (-0.035)). To determine whether this difference is significant, we test whether the marginal effect for 2-session COS differs from the marginal effect for 3-session COS. The results of this test indicate no significant difference between 2- and 3-session COS (p=0.389) meaning COS format does not appear to be correlated with recidivism.

Both direct and indirect testing indicates no significant relationship between COS format and recidivism controlling for various demographic characteristics; criminal, social, and other behavioral factors; and observed court and county characteristics.[45] Again, we do not believe that these results should necessarily be interpreted as lack of a *causal* relationship between COS format and recidivism. Courts that choose to implement 2- versus 3-session formats could differ significantly in unobserved ways that limit the causal interpretation of the above results.

[45] We found similar patterns for felony and misdemeanor recidivism.

Acknowledgements

The authors would like to thank staff at the Administrative Office of the Courts, the Juvenile Rehabilitation Administration/DSHS, members of the CJAA Advisory Board, juvenile court staff, and Desiree Cheung of Cowlitz County Juvenile Court for the data and expertise necessary to conduct this evaluation.

For further information, contact:
Dani Fumia at 360.586.2795, dani.fumia@wsipp.wa.gov

Document No. 15-09-1901