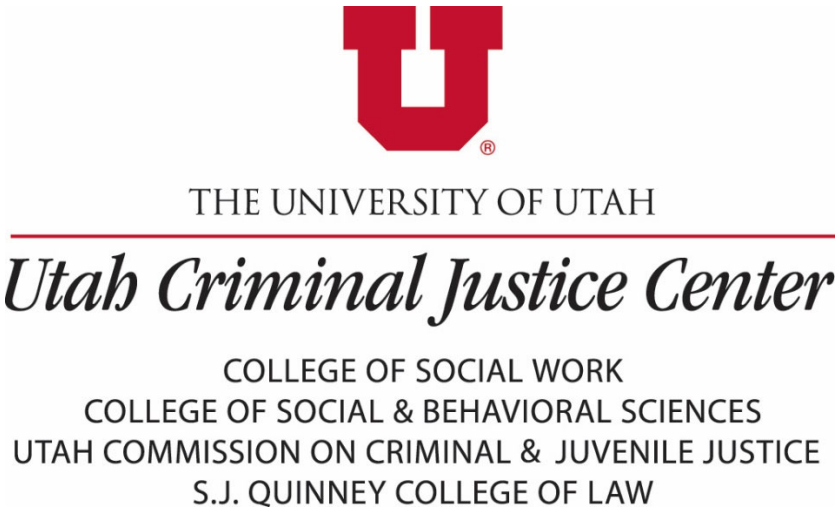
Evaluation of the Utah Department of Correction's (UDC) Implementation of the Statewide Adult Recidivism Reduction (SRR) Program:
Phase One Report



Evaluation of the Utah Department of Correction's (UDC) Implementation of the Statewide Adult Recidivism Reduction (SRR) Program: Phase One Report

Kort Prince, Ph.D
Derek Mueller, M.S.
Christian Sarver, Ph.D.

September 2020

Table of Contents

Acknowledgments

The evaluation research contained in this report was made possible by a Bureau of Justice Assistance Grant (Number 2017-CZ-BX-0210) awarded to the Utah Department of Corrections (UDC). Data for the analyses in the report were provide to the evaluation team at the Utah Criminal Justice Center (UCJC) by UDC. UCJC would like to thank (in alphabetical order) the following staff at UDC for both providing the data essential to the project and for exceptional support and collaboration throughout the evaluation process: Julie Christenson, Dennis Franklin, Gary Jensen, and Phillip Stireman.

Background

Program Introduction

The Bureau of Justice Assistance awards funding to federal and state correctional agencies for vital programs and systems reform aimed at improving the reentry process through the Second Chance Act (SCA). The SCA legislation was signed into law on April 9, 2008 and designed to help states take a systematic, sustainable approach to establish policies and practices that will improve recidivism outcomes for individuals returning from federal and state prisons, local jails, and juvenile facilities. In order to receive funding, correctional agencies must invest in implementing evidence-based programs and practices that have been shown to reduce recidivism. The initiatives selected by correctional agencies must address three primary areas: 1) use of risk/needs assessments to inform resource-allocation decisions and individual case plans, 2) evaluate recidivism-reduction programs, practices, and trainings and ensure that they are implemented with fidelity, and 3) implement community supervision policies and practices that promote successful reentry.

The Utah Department of Corrections (UDC) was awarded funding through the SCA to implement an initiative to improve the reentry process for individuals returning to the community from prison. It was through this funding that UDC designed and implemented the Statewide Adult Recidivism Reduction (SRR) initiative. The main goal of SRR is to reduce the criminogenic risks of justice-involved individuals reentering the community. In order to achieve this, UDC is implementing more frequent/timely LS/RNR assessments as well as developing individualized case action plans for parolees based on current risk/needs assessment results. The case action plans consist of evidence-based programming, career building opportunities, and strong support systems that begins upon intake into prison, evolves throughout their time in custody, and continues upon release back into the community.

As of 2014, approximately 62% of UDC inmates released from prison return within 3 years of release. SRR was designed to have the greatest impact on high or intensive risk individuals given that they are responsible for the majority of returns to prison in Utah. High to intensive risk individuals accounted for 84% of returns to prison in 2014. UDC has set a goal to reduce statewide recidivism by 10% within the first two years of implementing SRR initiatives and 25% within a 5-year period.

Evaluation Plan and Objectives

The Utah Criminal Justice Center contracted with the Utah Department of Corrections (UDC) to evaluate several aspects of the Statewide Adult Recidivism Reduction (SRR) Programs' initiatives and their hypothesized effects. This report is divided into sections that address each of the following objectives related to SRR:

1. Examination of whether risk/needs scoring fidelity (quality assurance), overall and across UDC Adult Probation and Parole locations (when sample size allowed), improved from pre- to post-SRR adoption;

2. Examination of whether the frequency of risk/needs assessments administered within 60 and 90 days of release pre- and post-SRR differed;
3. An analysis of whether client risk/needs assessment (overall and by domain-specific score) trajectories improved (owing to improved services) between pre- and post-SRR periods;
4. Examination of whether the case action plans (CAPs) occurring/updated within 60 and 90 days of release improved between the pre- and post-SRR periods; and
5. An analysis of whether case action plans' alignment with needs identified by the LS/RNR risk assessment improved from pre- to post-SRR adoption.

As noted in more detail in the relevant section below, the fourth and fifth objectives could not be addressed as planned in the current evaluation due to data limitations. Analyses related to goals four and five are tentatively planned for a Phase II evaluation, if funded. The nature of the problems surrounding CAP data are discussed in more detail in Section Four, CAP-related issues.

Note that this report is a Phase One report evaluating the aforementioned objectives of the program. Pending future funding, two additional phases will examine recidivism and treatment outcomes in addition to the five objectives listed above (including the CAP analyses, if possible). Because of the limited amount of time since SRR implementation, an analysis of recidivism was not viable at this stage and will be covered in a Phase Two report if funded. Additionally, UDC is currently working on a system to better track treatment and treatment dosage. These system-improvements are expected to be in place by the time a Phase Two report would be written; treatment outcomes will be examined at that time.

General Analytic Caveats

When examining all analytic models that follow, it important to keep in mind that the evaluation examines changes in patterns associated with the adoption of SRR. While it is tempting to assume a causal relationship between any improvements and adoption of SRR, the design is not a randomized control trial (RCT) and cannot infer such causality. Practically speaking, it was not feasible for UDC to utilize an RCT. While SRR-related enhancements or usual practices could have been randomly assigned to AP&P locations, doing so was not feasible in practice for several reasons. First, AP&P offices across the state represent drastically different populations, with differing offense types and resources. Second, if random assignment did occur, carryover or contamination effects would be expected, as training is not region-specific. Third, and perhaps most importantly, the adoption of SRR was intended to fundamentally change the way UDC and AP&P offices operate. This means SRR applies to UDC as a whole and not specific regions.

While somewhat comprehensive in scope, the reasons for lack of an RCT design listed above are not intended to be exhaustive. They are intended to remind the reader that the evaluation can only speak to patterns of association rather than causality. To the extent that patterns are consistent, one might infer an effect was reasonably associated with SRR. However, it remains true that factors aside from SRR also differ across time; these include, as examples, person (i.e., different clients) and system-level (i.e., different policies) factors. These cannot be ruled out as alternative explanations to improvements associated with SRR.

Changes in LS/RNR Quality Assurance Scores

Purpose

In an effort to improve the validity and fidelity of the LS/RNR assessments conducted by Adult Probation and Parole (AP&P), correctional case managers within the programming department, and by those in the inmate placement program, UDC has implemented an LS/RNR Quality Assurance (QA) scoring assessment. The assessment is performed by a coach and evaluates the capabilities of LS/RNR assessors on their ability to accurately administer the assessment. Evaluation of assessment accuracy contains three domain scores:

- Motivational Interviewing
- Interviewing Skills
- Integrity of Scoring/Case Plan

These sections have 8, 6, and 8 items, respectively. Each item in a domain is scored on a 1 to 5 scale, where higher scores are preferable. Scores within each domain are averaged such that each domain score has a maximum value of 5; the three domain scores are then summed. Across the three domains, the lowest possible score is three and the highest possible score is 15.

Analytic Approach

QA scores were provided to UCJC by UDC staff. The QA assessments were conducted on a quarterly basis from July 2017 through the quarter ending in September 2019. Beginning in October 2019, assessments were, in some cases, collected more than once per quarter. For analysis purposes, when more than one assessment was collected in these quarters, the scores were averaged to produce a quarter-based average.

Once data were cleaned for analysis purposes, the maximum number of quarters for which a person had a QA assessment was 11, with a minimum of 1 and a mean of 6.3. In all, 430 LS/RNR assessors had assessments available for analysis, providing 2,707 total assessments across the 11 available quarters (ending after the first quarter of 2020).

The structure of the data necessitated an analytic approach that address the nested nature of the data and the dependencies such data create. In these data, assessments are nested within individual LS/RNR assessors, which are further nested within AP&P locations; in other words, the assessments form a hierarchy where locations have multiple people and people have multiple assessments. Typical analytic approaches assume such dependencies do not exist and are not robust against violations of this type. In these data, one can better appreciate the nesting structure by considering dependencies at each level. Here, dependencies exist because one assessor's QA scores are more likely to be similar to her own scores at different times than to another assessor's scores. Similarly, assessors within specific locations, owing to regional practices and client differences, should also be more similar when compared to assessors in other locations.

Initial examination of the distributional form of the QA assessment scores indicated the data conformed well to a normal distribution[1]. Given the nested structure, data were modeled using multilevel discontinuous growth models. The term "multilevel" simply indicates the model will address the data dependencies described above. Discontinuous growth models[2] are a special form of regression that look for shifts in a general pattern (i.e., trajectories over time) that might be tied to specific events – in this case, SRR[3].

It is difficult to conceptualize what a discontinuous growth model does in practice without seeing it visually. Therefore, to help set concepts, Figure 1 below provides some examples of hypothetical discontinuities we might expect to observe in the QA data. For expositional purposes, the four examples in the figure are denoted by A, B, C, and D.

Possibility A suggests that QA scores were improving over time, but the adoption of SRR did not alter that general trend. In that case, we would determine that SRR had no immediate discernable effect above that of time. Pattern B suggests that, while scores were improving over time, SRR produced a sudden jump in elevation of the line (i.e., a discontinuity). In that case, SRR would be associated with a "sudden" improvement in QA scores; after the jump in elevation, the rate of change (slope) remains the same.

Pattern C suggests that there is no sudden shift in QA scores associated with SRR, but that adoption of SRR is associated with a change in the slope or trajectory of scores. This pattern suggests that scores begin to improve more quickly with adoption of SRR. Finally, pattern D suggests that SRR is associated with both a sudden improvement in QA scores and a faster improvement that alters the slope. In most cases, pattern D is the most favorable for a program such as SRR, but patterns B and C are also favorable. While pattern A can also be favorable, given that SRR was rolled out over time, it is difficult to attribute pattern A to a single event, such as SRR.

These patterns are not intended to be exhaustive; instead, they are examples of the trends we might expect to see with these data. Among other possibilities, a fifth possibility (not shown) is a completely flat trajectory that assumes no change in QA scores over time or as a function of SRR. Though not pictured in the figure, this possibility is automatically considered by the model. If the

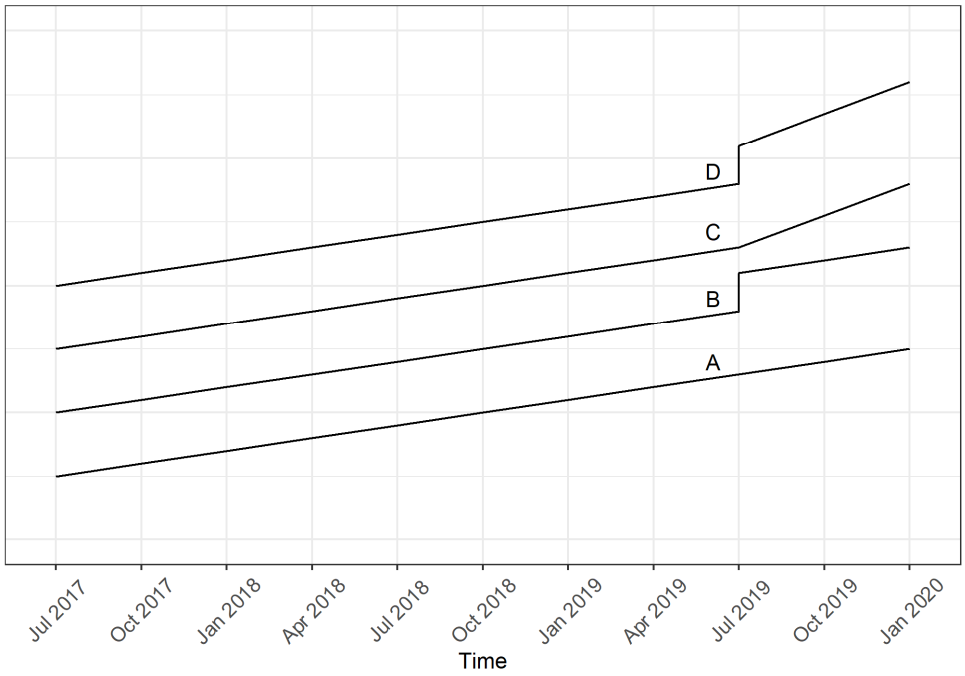
[1] Technically, the raw data do not have to be normal (i.e., Gaussian distributed) for the analysis to proceed with an assumed normal distribution. What matters is whether the residuals derived from the model, conditioned on predictors, are normally distributed. However, when the raw data are normal, the residuals are almost certain to be normal as well except for under specific, unusual circumstances.

[2] Singer, J. D., & Willet, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* Oxford: Oxford University Press.

[3] It is not necessary to understand the model form to interpret the output, but sometimes providing the model form can be helpful. For the interested reader, the equation for the model is as follows: $Y_{tij} = \gamma_{000} + \gamma_{100}Time_{tij} + \gamma_{200}SRR_{tij} + \gamma_{300}Time_{tij} * SRR_{tij} + \mu_{00j} + r_{0ij} + e_{tij}$. In this case, all predictors are time varying, or at level-1. To account for dependencies in the data, the model contains random intercepts for location and person in addition to an overall error term. Though not shown, model fit criteria indicated models that included predictors for the variance were a better fit. When predictors for the variance are included, the model allows for heterogeneity of variance; this regression assumption would be violated without inclusion of these terms.

coefficient for time and SRR were both zero, a flat line would be observed when predicted values were graphed. Other possibilities include patterns where SRR led to decreases in QA scores. Though not shown, the analytic model automatically addresses this possibility as well.

Figure 1: Examples of Potential Discontinuous Growth Patterns for QA Scores



Notice in the figure that the changes in elevation and slope patterns are centered on July 2019. This was the date by which UDC staff indicated SRR and related policies should be fully implemented. However, it is also true that adoption of SRR and related policies is not a discrete event. Clearly not all changes related to SRR were adopted on July 1, 2019. Instead, it was an implementation process where related changes were expected to be fully in place by that date. Discontinuous growth models do not assume the discontinuity occurred exactly at the point in the figure, but the changes do need to occur near the point. An alternative analysis was performed to determine whether another point in time, perhaps associated with changes unrelated to SRR, better explained any elevation or slope changes. No other time point leading to a discontinuity was discovered and so these models are not discussed further.[4] Though non-linear trends were also investigated, these preliminary models indicated a non-linear form was not necessary; accordingly, only the results of the linear model are discussed.

[4] A completely different type of model is needed when one has no a priori reason to believe the shift in elevation or slope should occur at a specific time. There are several ways one can model the data without such a hypothesis, but the approach considered by UCJC was a generalized additive mixed model (GAMM). The details of these models are not germane to this report, but the general process occurs as follow: first the model is fit to the data with a focus on the time variable; then the first-order derivatives are extracted from the model. One then plots the derivatives and looks for periods where the derivatives do not touch 0 (or no change). In these data, the derivatives only changed at the hypothesized time; accordingly, the more specific approach described in the body of the report was adopted instead of the GAMM.

Analytic Results

The results of the modeling process are provided in Table 1[5]. The column marked "Predictor" shows the predictors included in the model; in this case, the models consider an effect of Time (i.e., QA scores might be expected to change over time), SRR (i.e., the immediate elevation effect associated with SRR that produced increased scores)[6], and the interaction between SRR and Time (i.e., changes associated with time may differ before, relative to after, SRR).

The column labeled "Estimate" shows the coefficient associated with each predictor. The value for the intercept is usually an ignorable component of the model. This coefficient indicates the average value of the QA score when time and SRR are 0. In this case, time is 0 at the point of full SRR implementation (July 2019) and SRR is coded as 0 and 1, where 0 indicates pre-SRR and 1 indicates post-SRR. Therefore, the average QA score pre-SRR, but exactly at the point of full SRR implementation, is expected to be 10.9 out of 15.

The estimate for time indicates that for every one-unit change in time (one quarter in this case), QA scores are expected to increase by 0.28 points, holding all other variables constant. The effect for SRR provides the elevation effect associated with full implementation of SRR. Thus, at the point of full implementation, QA score are expected to increase by 0.20 points. The interaction coefficient for time and SRR is 0.16; however, interactions are much easier to interpret graphically, so we defer discussion of this term until presentation of a graph of the model-predicted values below in Figure 2.

The column labeled "Std. Error" provides an estimate of the uncertainty around the point estimate in the "Estimate" column. The 95% confidence intervals (CI) provide the confidence level around the point estimates. In this type of model, the estimate is significant (at $p \leq .05$)[7] when the CIs do not cross 0. The specific level of significance is provided in the column labeled "P-value". In this case, all effects are significant.

However, despite the significant p-values, it is important to note that one of the limitations of p-values is that it is easy to obtain significance with large samples. In this case, 2,707 assessments across 430 assessors and 42 locations reflects a large sample, and, though it is always useful to consider effect sizes, this becomes particularly important in large samples where trivial effects can be significant.

---

[5] Note that, for simplicity, random effects and predictors associated with the variance are not shown. These parameters are not essential to understanding the model and are omitted for that reason.

[6] It is important to note that time was centered at the point of full SRR implementation, or July 1, 2019, making that value of time 0. This is essential to obtaining the elevation effect that SRR had at the time it was fully enacted. If time were not centered such that enactment of SRR reflected time 0, the SRR coefficient would not capture the effect of SRR.

[7] Despite the fact that p-values have serious limitations, they are still widely used and are generally useful as long as other aspects of the model, such as the CIs, are also considered. R.A. Fisher, whose work is synonymous with significance testing did not argue there was anything special about the value p≤.05, but .05 has been widely adopted as the standard against which to determine significance (See Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A, 22,* 309–368).

Table 1: Coefficients from Discontinuous Growth Model of QA Scores

Predictor	r	Estimate	Std. Error	95% CI	P-value
(Intercept)	--	10.89	0.05	10.78 – 10.99	**<0.001**
Time	0.44	0.28	0.01	0.26 – 0.30	**<0.001**
SRR	0.06	0.20	0.07	0.06 – 0.34	**0.005**
Time * SRR	0.05	0.09	0.04	0.02 – 0.16	**0.017**

N_{person} = 430, $N_{location}$ = 42, Assessments = 2,707

In contrast to a p-value, which is impacted by sample size, an effect size is less influenced by large samples[8]. It also provides a standardized measure that allows one to compare effects across predictors on different scales such as time (coded -9 to 2 here where each one-unit change is a quarter) and SRR (coded 0 or 1 here). Effect sizes are also a measure of theoretical importance. That is, they indicate whether an effect is theoretically important whether or not it is significant. Here, effect sizes are provided as correlation coefficients in the column labeled "r"[9]. In the social sciences, these can generally be interpreted, in accord with Cohen (1988)[10], such that small, medium, and large effects correspond to values of .1, .3, and .5, respectively. Small to medium effect sizes are most common in the social sciences; large effect sizes are fairly rare.

From the "r" column, one can see the largest effect is associated with Time. This value (.44) corresponds to about a medium-to-large effect. In the case of SRR and the interaction, the effects are below "small" using Cohen's standards.

Typically it is easier to understand what a model is indicating by graphing the predicted values from a model; this is particularly true for the interaction term. Figure 2 provides a graphic interpretation of the model in Table 1. Time is on the x-axis and the predicted QA score is on the y-axis. The SRR elevation effect (that is, the main effect of SRR, denoted "SRR Elevation Effect") is annotated and a solid black line and indicates the immediate shift in elevation of QA scores associated with full implementation of SRR.

In the figure, it is easier to see the impact of the significant interaction. In addition to an elevation effect for SRR, the figure makes it clear that the slope also changes at full SRR implementation (denoted "SRR Slope Effect"). That is, in addition to the immediate shift of 0.20 associated with SRR, after SRR implementation, an additional increase of 0.09 is associated with each additional increase in time (measured in quarters). Seen graphically, it is also easier to appreciate the importance of effect sizes. Although significant, the effect of SRR and the interaction are relatively small, especially compared with Time. A caveat to this interpretation is provided in the next section, which focuses on location-specific models.

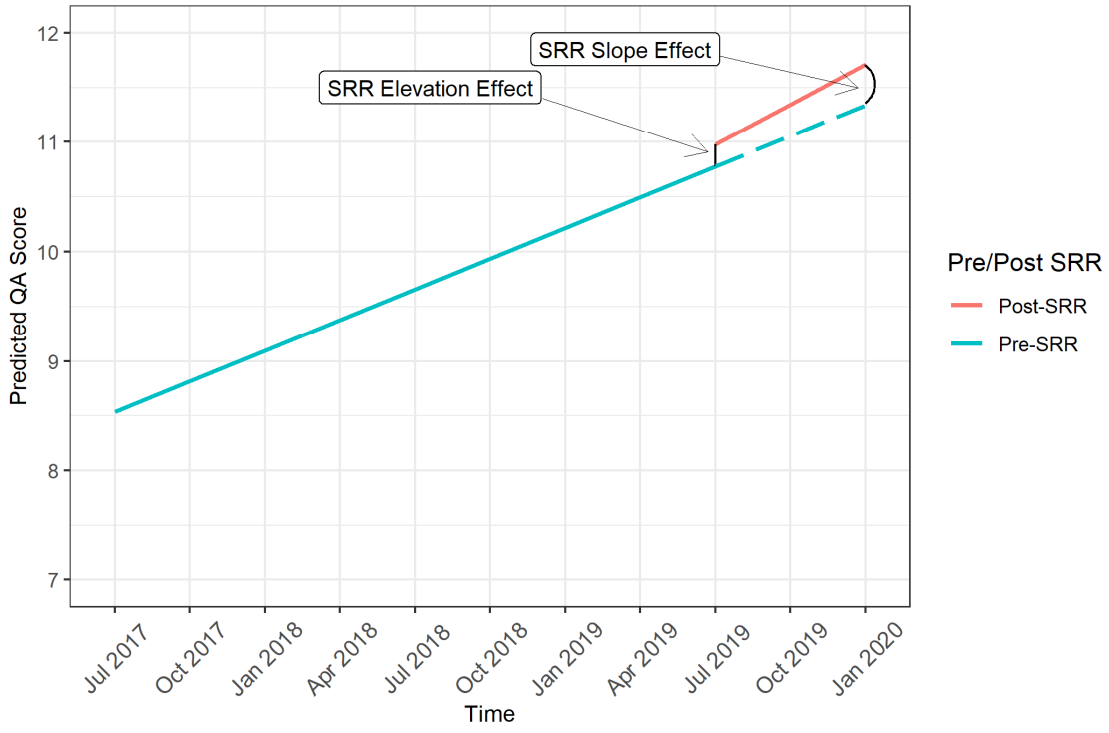
[8] Effect sizes are "mostly" independent of sample size, but, in small samples, the CIs around them can be very uncertain, suggesting a lack of confidence in the effect size. Here, the sample is quite large and this is not a concern.

[9] Effect sizes in multilevel models are not universally agreed upon and different methods can lead to drastically different results owing to the presence of random effects. However, because it is important to think about effects in terms of effect sizes, we include r as a means to assess this. The reader should keep in mind, however, that aspects of multilevel models bring the accuracy of these estimates into question.

[10] Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Erlbaum.

Note that, at July 2019, the blue line reflecting pre-SRR becomes a dashed line. The dashed line represents what is known as a counterfactual. Clearly, the post-SRR projections for the pre-SRR group are outside of the observed data because it is impossible to have an observed pre-SRR slope once SRR has been implemented. Nevertheless, the model can project this counterfactual based on trends up to the point of SRR implementation. Given they are projections, they can be inaccurate because what might have happened to the pre-SRR slope in the absence of SRR can never be known given that SRR was, in fact, implemented.

Figure 2: Pre- and Post-SRR QA Score Trajectories From Discontinuous Growth Model



Considering the model and the figure concurrently, one would conclude that full implementation of SRR was associated with a small increase in QA scores as well as an increase in the rate at which the scores themselves increased after implementation; this is similar to pattern D in Figure 1. Time is the most important predictor, revealing a medium-to-large effect. It should be noted, however, that Time might also be an indicator of the effect of SRR. As noted earlier, SRR was not a discrete event that occurred exactly on July 1, 2019. To the extent that some policies were enacted to improve QA assessments before full implementation, the effect of Time might also contain some of the effect of SRR and it is clear QA scores have been improving since SRR was implemented.

Location-Specific Models

Originally, an additional goal when assessing changes in QA scores was to examine how changes differed by AP&P office location. However, such an analysis was not possible for every location owing largely to a small number of LS/RNR assessors within certain locations. In some cases,

there were 10 or fewer assessors per location, which, in conjunction with a small number of assessments, does not provide enough cases to produce stable estimates.

Only four locations (Region Three Field, Region Four Field, Northern Utah Region Ogden, and the Draper AP&P location) had over 20 assessors. When modeled, two of these locations, Northern Utah Region (NUR) Ogden and Draper, had negative slopes for QA assessment scores post-SRR, meaning that their scores actually became worse after full implementation. An examination of the number of assessments in these locations over the quarters for which data were available indicated many more assessors were receiving QA assessment scores post-SRR rather than pre-SRR.

Normally missing data (such as missing pre-SRR assessments) is not a major challenge for multilevel models, but, in this case, the data were not missing at random as the models assume. Instead, they were systematically missing for earlier time points. Whether newly hired assessors were joining these locations post-SRR or whether efforts to collect QA assessments were improving over time, leading to more assessments post-SRR, the lack of a sufficient number of pre-SRR assessments by person in these locations made models unstable, but also misleading. If attributable to new hires, for example, one might not expect these new assessors to be proficient with the LS/RNR at the same level as someone who had been administering it for years, even if the new hires (assuming that is the cause) received better training. By attempting to compare new assessors (as in this example) to those who had been using the assessment for years, one could reasonably expect a negative slope where QA scores decline over time and after SRR implementation.

More succinctly, the pre- and post-SRR analysis, if performed with vastly different number of people in each period, would not provide a test of SRR. Of course other scenarios could also explain the negative slopes found in some of these models, but the point remains that models without sufficient representation pre-SRR would not be an accurate model of the effect of SRR. This would be true whether slopes were positive and favorable as well; sometimes the data simply cannot support a model.

Instead of modeling data that would produce ambiguous models of change, the analysis in this section focused on the two locations (Region Three Field and Region Four Field) with the most steady rates of assessment; that is, the regions where most of the assessors contributed to both the pre- and post-SRR trends.

Region Three provided 100 assessors and 541 assessments while Region Four provided 40 assessors and 317 assessments. The details of the modeling process are identical to the model of all locations combined above and are not repeated here. Results from the models are provided in Table 2. Because these are identical models derived from a population of data, the models coefficients can be compared across locations and to the overall model above that combined all locations.

The estimate for Time indicates that, for every one-unit change in Time (one quarter in this case), QA scores are expected to increase by 0.14 points in Region 3 and 0.16 points in Region 4, holding all other variables constant. The effect for SRR provides the elevation effect associated with full implementation of SRR. Thus, at the point of full implementation, QA scores are expected to

increase by 0.61 points in Region 3 and 0.57 points in Region 4. The interaction coefficient for Time and SRR is 0.26 in Region 3 and .09 in Region 4; however, we again defer discussion of this term until presentation of a graph of the model predicted values below in Figure 3.

From the "r" column, one can see the largest effect is associated with Time in both regions. In Region 3, encouragingly, all effects represent small-to-medium-sized effects. In Region 4, the effect size for Time represents a medium-sized effect and, for SRR, it represents a small-to-medium-sized effect. The interaction effect size in Region 4 is well below small (and the coefficient was also not significant).

Table 2: Coefficients from Discontinuous Growth Model of QA Scores

Location	Predictor	r	Estimate	Std. Error	95% CI	P-value
Region Three	(Intercept)	--	9.81	0.13	9.55 – 10.07	**<0.001**
	Time	0.24	0.14	0.03	0.09 – 0.20	**<0.001**
	SRR	0.17	0.61	0.18	0.27 – 0.96	**0.001**
	Time * SRR	0.14	0.26	0.09	0.09 – 0.43	**0.003**
Region Four	(Intercept)	--	11.00	0.12	10.76 – 11.25	**<0.001**
	Time	0.34	0.16	0.03	0.11 – 0.21	**<0.001**
	SRR	0.19	0.57	0.19	0.20 – 0.95	**0.003**
	Time * SRR	0.02	0.04	0.11	-0.17 – 0.25	0.713

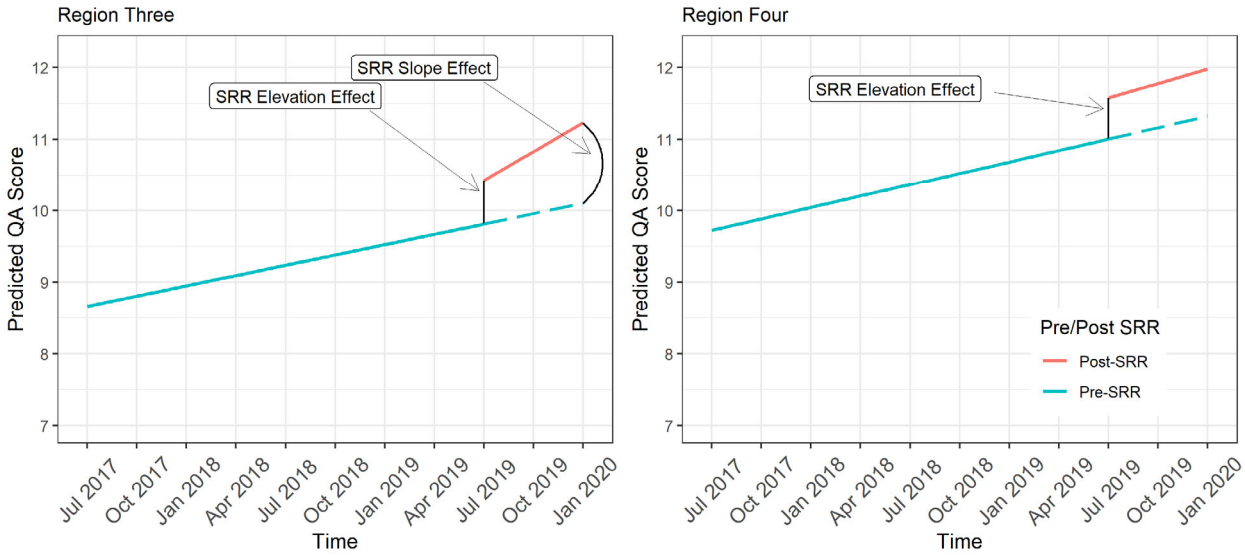
$N_{person,\ region\ 3}$ = 100, $Assessments_{region\ 3}$ = 541; $N_{person,\ region\ 4}$ = 40, $Assessments_{region\ 3}$ = 317

As before, we can examine a graph of the predictions derived from the models (see Figure 3). The figure has the same properties as the figure for all locations above, but combines the two individual location models into one figure. It is easier to see the impact of the significant interaction in Region 3. In addition to an elevation effect for SRR, the figure makes it clear that the slope also changes at full SRR implementation (denoted "SRR Slope Effect"). That is, in addition to the immediate shift of 0.61 associated with SRR, after SRR implementation, an additional increase of 0.26 is associated with each additional increase in time (measured in quarters).

Looking at Region 4, the graph makes it clear why the interaction was not significant, as the blue and red lines are largely parallel (indicating no change in slope post-SRR implementation). There is, however, a significant elevation effect associated with SRR. This is reflected by the immediate shift in elevation of 0.57 at full SRR implementation.

Considering the model and the figure concurrently, one would conclude that full implementation of SRR was associated with a small-to-medium-sized effect on QA scores in both regions. Time was also important in both locations, showing small-to-medium-sized effects, and indicating that QA scores have been improving over time ignoring SRR. In Region 3, the medium-sized interaction indicates that an increase in the rate at which the QA scores increased is also associated with SRR. The fact that scores are generally higher in Regions Three and Four may be of some practical value to UDC if assessors who perform well in this region can also be used as coaches in the future.

Figure 3: Pre- and Post-SRR QA Score Trajectories From Discontinuous Growth Models: Regions Three and Four Field



Section Summary

Though, for reasons outlined above, causality cannot be assumed, there are changes in QA scores associated with the timing and implementation of SRR. In the model including all locations, a small effect size was observed for SRR and a near small effect size for the interaction. However, the location-specific models indicate that outcome in the overall model needs to be interpreted with caution. The effects in the overall model, Region 3, and Region 4 can all be compared. Because the location-specific effects are larger (see tables) than in the overall model that combines all locations, and because these two regions are the largest in terms of LS/RNR assessors, one can infer that much of the improvement in QA scores associated with both time and SRR is due to assessors and QA improvements in these locations. In fact, as noted above, some other locations (for which models could not be created) saw tentatively negative effects for SRR, but these were noted above as being attributable to having largely different assessors in the pre-SRR and post-SRR periods for some locations (which is why models of change could not be created). Evidence from the modeled regions (3 and 4) suggests that it is not only training, but good training and practice that matters, as newer assessors (that is, newer to a location) did not do as well on QA assessments.

Changes in LS/RNR Client Trajectories

Purpose

Adoption of the SRR is theorized to create better agreement between clients' risk/needs and their treatment and supervision. It is hypothesized that this will change the trajectories of scores, leading to greater reductions in overall and dynamic domain scores over time in the post-SRR period relative to the pre-SRR period. It is presumed that this effect, if observed, would be partially attributable to better meeting client needs through treatment; however, development and collection of treatment and dosage data is not fully implemented yet within UDC. If additional years of funding are secured to study SRR, future evaluations will explore treatment completion and dosage and concomitant changes in risk/need scores.

Analytic Approach

The analyses in this section focus on changes in LS/RNR score trajectories that may be associated with SRR. Similar to the examination of changes in QA scores above, this section utilizes discontinuous growth models to identify shifts in a general pattern (i.e., trajectories over time) that might be tied to SRR[11].

The nature of the LS/RNR data produced some analytic challenges that should be considered as one evaluates the models that follow[12]. These include:

- The number of available assessments
- The number of dynamic items included in each domain score
- The distributional form of the outcomes (particularly LS/RNR domain scores)
- Notable non-linearity in trends

Beginning with the number of available assessments, 3,676 parolees were identified for the sample with LS/RNR test dates dating back to July 1st, 2017. This two year "look back" was intended to establish a trend in LS/RNR scores before implementation of SRR that could be used to compare to scores collected after SRR implementation. The initial plan was to divide each analysis by baseline risk category to account for the fact that individuals in different risk categories should get

[11] As with the QA section of the report, it is not necessary to understand the model form to interpret the outputs that follow, but sometimes providing the model form can be helpful. For the interested reader, the equation for the total score model is: $Y_{ti} = \pi_{00} + \pi_{10}Time_{ti} + \pi_{20}SRR_{ti} + \pi_{30}Time_{ti} * SRR_{ti} + \mu_{0i} + e_{ti}$. The model form for all domain score models is $\eta_{ti} = \log\left[\frac{\pi_{ti}}{1-\pi_{ti}}\right] = \pi_{00} + \pi_{10}Time_{ti} + \pi_{20}SRR_{Ti} + \pi_{30}Time_{ti} * SRR_{ti} + \mu_{0i}$. In both cases, all predictors are time varying, or at level-1. To account for dependencies in the data, the models contain random intercepts for person. The binomial models for domain scores, described more below, contain a logit link function and lack an error term, e_{ti}, because the variance and mean are linked in these models.

[12] The data dependency issue occurs here as it did for QA scores. Here, LS/RNR assessments are nested within clients. Because this issue was discussed in the QA section, it is not repeated here, but the models do account for it via the random intercept term.

different treatment, with those at greatest risk receiving the most intensive services; this, in turn, might lead to the greatest reductions in LS/RNR scores being observed for the higher risk groups.

Table 3 provides the mean and median number of assessments by baseline risk category (i.e., first available LS/RNR category score) in the pre- and post-SRR periods. Because the table is broken down by baseline risk, the mean and median numbers include only those parolees who had an LS/RNR assessment. The first thing to note in the table is that, on average, there are more assessments in the pre-SRR period than in the post-SRR period. This was expected because the pre-SRR period includes two years while the post-SRR period includes only one, given the timing of this report. When one considers that the pre-SRR period is twice as long as the post-SRR period, the average number of assessments is actually increasing in the post-SRR period. While 38 of the 3,676 parolees identified for the sample had no assessments at all, improving the number and timing of assessments is an issue UDC is working to address and the increased rate of assessment in the post-SRR period suggests improvement is occurring. Notice also that the average number of assessments increases slightly as the baseline risk category increases. This too was expected because higher risk individuals should be assessed more often.

Table 3: Mean and Median Number of LS/RNR Assessments by Pre/Post SRR and Baseline Risk Level

Period	Baseline Risk Category	Mean	Median
Pre-SRR	Low	1.8	2
	Moderate	2.0	2
	High	2.0	2
	Intensive	2.1	2
Post-SRR	Low	1.3	1
	Moderate	1.5	1
	High	1.5	1
	Intensive	1.6	1

The biggest challenge for the analytic models of changes in LS/RNR scores is the small number of assessments within each baseline risk group post-SRR. Across all groups post-SRR, the median number of assessments per person was one. The highest average number of assessments per group was 1.6 in the intensive risk group. The fact that most people had only one assessment post-SRR presents a challenge to any model attempting to assess change because a trajectory for change cannot be built at the person-level when most people have only one assessment; that is, generally speaking, change cannot be observed unless a second assessment occurs.[13]

[13] To some extent, this is an oversimplification of how multilevel models work. Models of change can be built when some individuals have only one assessment because multilevel models also consider the group-level (or fixed) effect that represents the average trajectory across individuals; however, in these data, and when divided by baseline risk group, too many individuals lacked the additional assessments needed to produce stable models of change in the post-SRR period.

The small number of assessments when divided by risk category made it necessary to combine the groups to produce stable models. In combining the groups, care was taken to ensure that the trajectories were not vastly different across combined groups such that combining them would obscure an effect in one group relative to another. Looking at the initial trajectories, the best combination to address this concern led to combining low and moderate as one group and high and intensive as another. This is not ideal given that these groups are not necessarily similar in terms of risk or supervision. If funding permits, analyses conducted in the future can split the risk groups when more post-SRR assessments are available.

A second issue for the analysis of changes in LS/RNR scores is that one needs to adjust expectations regarding the extent of change expected depending on how many items are dynamic within each domain and for the total score; this is not an analytic issue, but it is relevant to understanding the patterns one might expect trajectories to take. Dynamic items are items that can change over time and with treatment. These include, as examples, getting a job or taking responsibility for criminal acts. Static items, in contrast, cannot change. For example, an item in the criminal history section of the LS/RNR (sections/domains are discussed in more detail below) asks whether the individual was ever incarcerated upon conviction. Once that is true, it cannot change.

Table 4 provides each of the domains from the LS/RNR below and specifies how many items in each section are dynamic. Given the criminal history domain has only one dynamic item, trajectories for this section would not be expected to decline very much over time. Instead, one might hope they would stay flat, indicating risk is not becoming greater over time. For domains with mostly dynamic items, one might reasonably expect greater declines in risk scores. This should be kept in mind as one reviews model results below.

Table 4: Total Number of Items and Number of Dynamic Items by LS/RNR Domain

Domain	Dynamic Items	Total Items	% Dynamic
Criminal History	1	8	12.5
Education/Employment	8	9	88.9
Family/Marital	3	4	75.0
Leisure/Recreation	2	2	100.0
Companions	4	4	100.0
Alcohol/Drug Problems	6	8	75.0
Procriminal Attitudes	4	4	100.0
Antisocial Patterns	4	4	100.0
Total Score	32	43	74.4

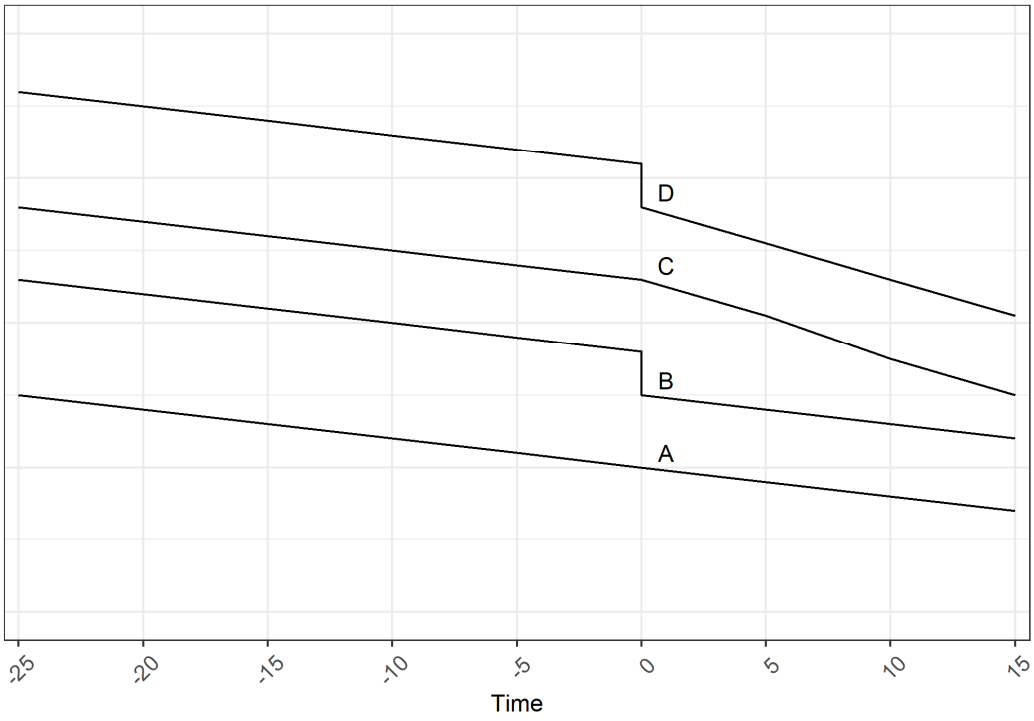
Similar to the QA assessments, this section utilizes discontinuous growth models to detect changes in overall patterns for LS/RNR total and domain scores as a function of both time and SRR. Figure 4 shows the pattern of discontinuities we might expect for LS/RNR scores. Time is again on the x-axis, but here it is measured in months before and after July 1, 2019.

Unlike the QA assessments, here, we expect the pattern to show decreasing rather than increasing scores. Because SRR was not implemented as a discrete event, we might expect some effect for Time, where scores decrease generally closer to SRR; this pattern is illustrated by trajectory A in the figure.

Trajectory B in the figure shows a sudden drop in risk at the point where SRR was fully implemented. Again, the model does not require the shift to occur exactly at this point; it merely needs to be centered near this point. Pattern C shows a change in the slope associated with SRR. In this case, there is no sudden shift, but scores decrease more quickly following full implementation. Finally, pattern D combines the two: a sudden shift in risk scores in addition to a quicker decrease in risk scores following full implementation.

As with QA assessments, a flat line (not shown) would indicate no effect associated with SRR. This pattern might be expected in the case of domain scores with few dynamic items, such as the criminal history domain. In that case, no change would be desirable because it would show no increased risk occurred over time.

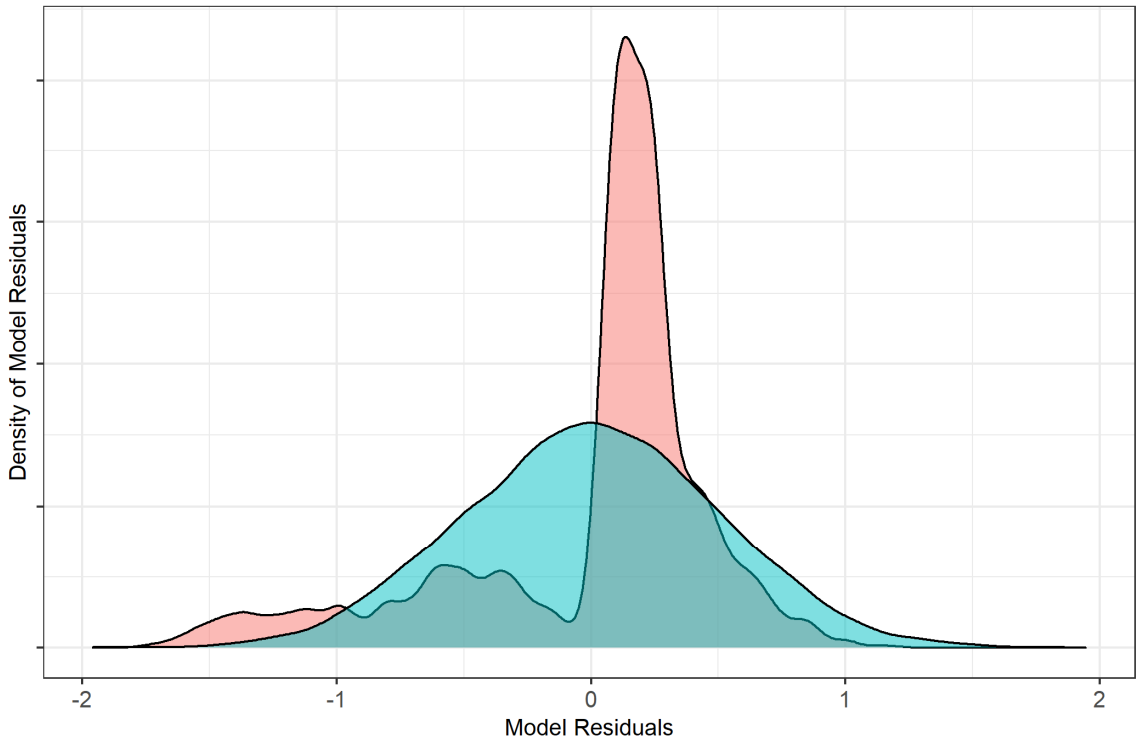
Figure 4: Examples of Potential Discontinuous Growth Patterns for LSRNR Trajectories



The third issue relevant to modeling changes in scores over time is the distribution of the overall (total score) and domain scores. Typical regression approaches assume that the model residuals (i.e., how much error there is in the model after it is fit) are normally distributed with a mean of 0 and a standard deviation of one. A value of zero implies perfect model prediction, or no error. Most residuals should be within plus or minus one standard deviation of this value after the model is fit. Figure 5 below shows a density plot of model residuals from an analysis of high/intensive risk parolees' leisure and recreation domain scores. The blue density in the figure represents an

ideal scenario, while the red density represents the actual residuals observed from a model that assumed a normal distribution. A degree of transparency has been added to the figure so that one can "see through" one shade to another.

Figure 5: Residuals Derived from Leisure and Recreation Domain Model with Assummed Normal Distribution



It should be immediately clear that the residuals from the model do not conform at all to a normal distribution. The blue density plot contains the same number of cases as the red one, but the red one looks like it has more cases simply because most cases have a residual between 0 and 0.5. There are also a large number of positive residuals (i.e., where the predicted value is less than the observed value) and very few negative residuals. It is not necessary to understand why this assumption is important for the regression approach, but it is important to understand that violating the assumption as notably as above necessitates the use of another distributional form.

The LS/RNR data present a serious challenges on this front. Domain scores are often bi-modal (i.e., have multiple peaks) and do not conform well to any particular distributional form. Several different distributional forms were examined to identify the best fitting distribution for the total and domain scores. For the total score model, the best-fitting distribution was a skew normal distribution; this distribution assumes the data are distributed much like the blue density in Figure 5, but allows for the mean to be shifted away from zero.

For the domain scores, a binomial process was used[14]; this model treats each point on the domain scores as a yes or no out of a total value. For example, the Leisure and Recreation domain has two items, and possible scores are, therefore, 0, 1, or 2. A person who scores a one on the domain then receives a probability on the binomial scale of 0.5 (or ½). Additional details about the implications of these modeling choices are described below.

The final issue to consider is departures from linearity in the trajectories over time. Typical regression procedures assume a linear relationship between predictors and outcomes. While this works well, or reasonably well, in many contexts, it is not always an accurate representation of the relationship. In the case of the models that follow, non-linearity between Time and total or domain scores is likely. This often occurs when relationships reach an upper or lower bound. For example, a particularly high risk group might score at or near the top end of one of the domain scores. While their scores might increase over time, once they reach the upper bound of the scale, they cannot increase any further and the relationship must flatten or change direction.

A real example from the current analysis will help set concepts. Figure 6 shows the relationship between time and the LS/RNR total score for the low/moderate risk group both pre- and post-SRR. The model is a discontinuous growth model similar to the QA section of this report. In the figure, Time is on the x-axis and is measured in months. The value zero represents the point of full SRR implementation. The scale shows scores from approximately two years pre-SRR to one year post-SRR. In this case, the focus is on the pre-SRR trajectory as well as its counterfactual (or projected) value once SRR was implemented.

The left panel shows a presumed linear relationship, which is an assumption of regular regression procedures and which worked well in the case of QA scores in section one of this report. One can test for non-linearity in a number of ways, but the preferred standard is to model the relationship using a smoothing spline[15], which detects non-linear trends. One can then compare model fit between the linear and smoothing spline models using statistical procedures. In order to ensure that the non-linear model is not just detecting noise in the data, a penalty term is added to the models. One can think of this as a means to be relatively certain a linear relationship is not adequate.

The right panel provides an alternative model of the same data as the left panel; particularly in the post-SRR counterfactual, the models predict very different scores. For the pre-SRR counterfactual in the linear model, the predicted low/moderate risk score is nearly 25 when Time reaches its maximum observed value of 15 (15 months post-SRR implementation). If the model extended further, it would continue to rise, but this is not a reasonable prediction and is an artifact of assuming the relationship is always linear over time. In fact, if this model were accurate, it would

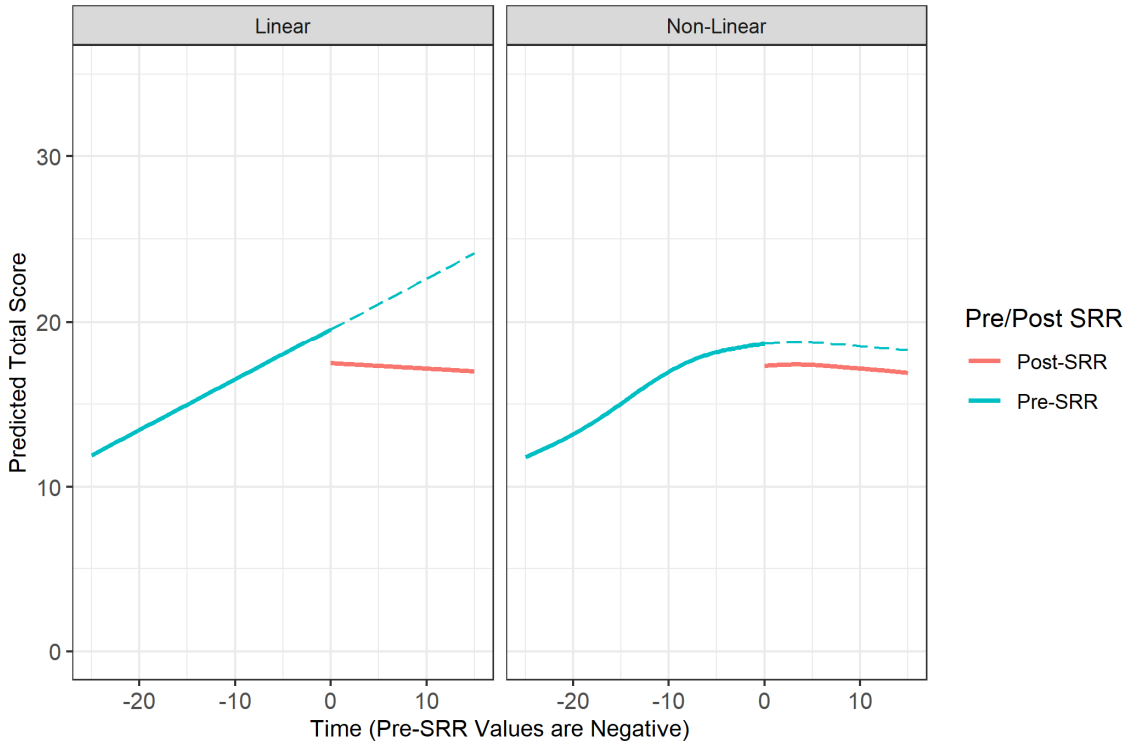
[14] Other considered distributional forms were skew normal, gamma, Poisson, negative binomial, and ordinal. Each of these produced violations of the model assumptions that were considerably worse than the binomial process model.

[15] In the past, non-linear relationships were commonly modeled using polynomials, but those provide global fit rather than local fit, as one gets from spline-based models. In short, spline-based models provide better predictions than polynomials, which can predict scores that are impossible values on a given scale. For more information, see: Wood, S. N. (2017). *Generalized additive models: An introduction with R*. Boca Raton: Chapman & Hall, CRC.

mean that the average value of the low/moderate group would put them in the high risk group (total scores for high risk range from 20 to 29) at 15 months post-SRR implementation.

In contrast to the linear model, the spline-based model recognizes the relationship is non-linear before it enters into the projected/counterfactual period[16]. Accordingly, the spline-based model levels off at about five months pre-implementation and does not make predictions not supported by the data. In this model, one can see an effect for SRR is still present, but it is much smaller than the linear model would suggest. Again, one can assess which model is more appropriate using model fit statistics. Most models that follow utilize a linear trend, but, alternative spline-based models were also considered and those models are presented when linear fit was not adequate.

Figure 6: Pre- and Post-SRR LS/RNR Total Score Trajectories (Low/Moderate Groups) Assuming Linear or Non-Linear Forms



Interpretation of Figures

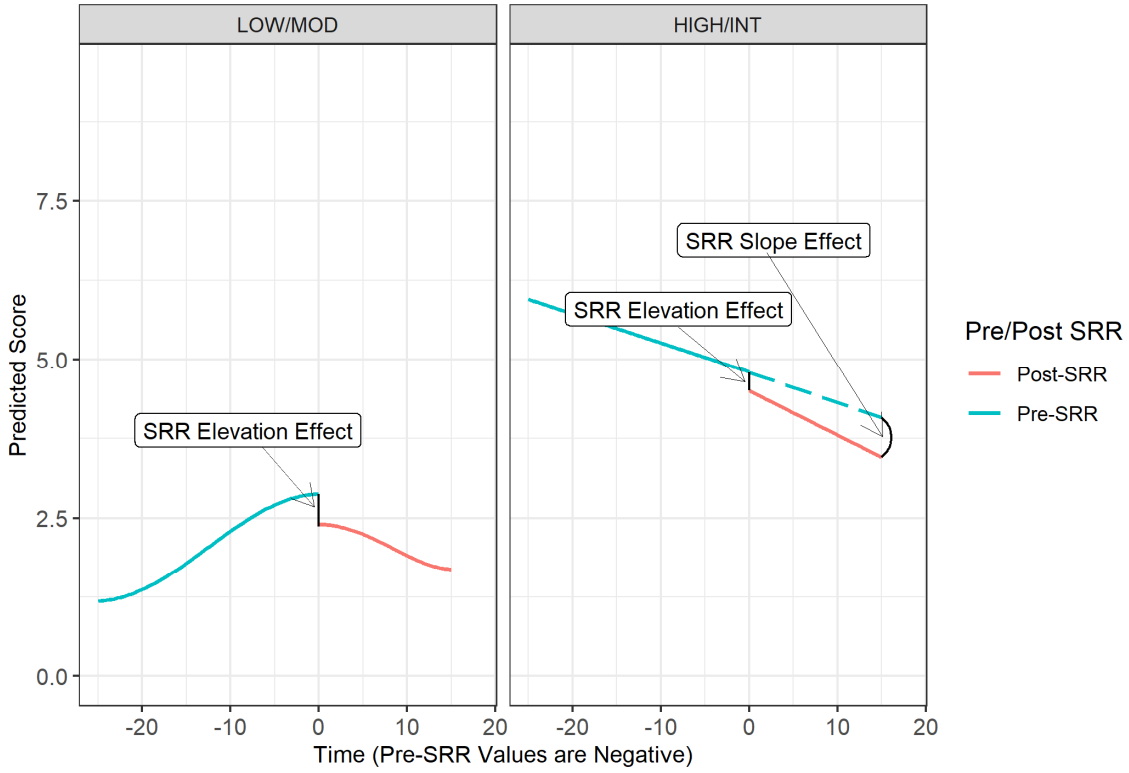
To avoid redundancy in model interpretation, an exemplar of the figures provided in this section is shown in Figure 7. This figure is derived from a domain score model shown later, below. It was selected as an exemplar because it illustrates all of the properties of figures that follow in this section.

In the left panel, one can see an annotation for SRR elevation effect demarcated by a vertical line (a pipe, "|") that illustrates the size of the effect. Note that there is no projection of the blue line

[16] Though discussed in more detail below, technically this projection should not be made in this type of model. It is used here to illustrate the notable difference in the models.

because this model is a spline-based model. Spline-based models learn from the data directly[17]. There is, therefore, no basis for projections in these models as there are for linear models, which merely extrapolate the linear trend forward. For spline-based models, only a main effect of SRR and Time will be presented, when applicable.

Figure 7: Example Figure of Trajectories by Baseline Risk Category



In contrast, the model on the right side did not require a smoothing spline and is modeled using a linear form. In this case, one can extrapolate a linear effect into unobserved time (although, as noted above, extrapolations can be inaccurate because they are projected based on a linear assumption), which makes modeling an interaction possible for these models. The model in the right panel of the figure shows both the main effect of SRR, via the vertical line, and an interaction denoted by "SRR Slope Effect" and an arc. Technically the arc is the not the size of the effect; it merely illustrates that a significant effect is present. For the figures that follow, any figure in which a particular annotation is not seen (either a vertical line or arc with text annotation), that effect is not significant. For spline models, though, one should only expect a main effect or vertical line to be shown given the lack of counterfactual projections.

Analytic Results

[17] Hemmi, M. H., Schneider, S., Müller, S., Meyer, A. H., & Wilhelm, F. H. (2011). Analyzing temporal patterns of infant sleep and negative affective behavior: A comparison between different statistical models. *Infant Behavior and Development, 34*(4), 541-551. doi:10.1016/j.infbeh.2011.06.010

LS/RNR Total Score

Initial examination of the LS/RNR total score trajectories revealed non-linear patterns that necessitated use of a smoother, or spline-based model. Results of the model are show in Table 5 and are illustrated in Figure 8. As noted above, outcomes are provided separately for individual's whose baseline risk categories were low/moderate and high/intensive. Because of the non-linear trends, there are no interactions in the total score models; the focus is on Time and SRR. As noted in the footnote in the table, 860 low/moderate baseline risk individuals provided 2,632 assessments while 2,778 high/intensive baseline risk individuals provided 9,273 assessments.

In these models, an effect size can only be computed for the linear term, SRR. Similarly, standard errors and confidence intervals are not available for the smoother terms (Time) because coefficients for smooth terms do not have the same interpretation as in linear regression and they are typically interpreted graphically rather than statistically. However, the coefficients do still communicate information about how "wiggly" or non-linear the trend is. The value in the "Estimate" of a smooth term is known as the Effective Degrees of Freedom (EDF). This value is roughly equivalent to the polynomial that would be required to model the relationship plus one[18]. One can, therefore, obtain the approximate polynomial form by taking the EDF and subtracting one.

As an example, consider the effect for "Time" in the low/moderate group. The value of 3.8 – 1 = 2.8. This is roughly equivalent to a cubic trend. Values near two would be roughly equivalent to a quadratic polynomial[19]. In the case of the low/moderate group, the effect for Time is not significant. However, full SRR implementation was associated with a small drop in overall risk level (1.36 points). Despite the significance, however, the effect size (r) was very small at .04.

Table 5: Coefficients from Discontinuous Growth Model of LS/RNR Total Scores

Baseline Risk	Predictor	r	Estimate	Std. Error	95% CI	P-value
Low/Moderate	(Intercept)	--	15.77	0.43	14.92 – 16.62	**<0.001**
	SRR	.04	1.36	0.63	0.11 – 2.60	**0.033**
	Time	--	3.80			0.186
High/Intensive	(Intercept)	--	26.79	0.37	26.06 – 27.52	**<0.001**
	SRR	.00	-0.02	0.50	-0.99 – 0.96	0.972
	Time	--	4.02			**<0.001**

N_{person} = 860 (low/mod)/2,778 (high/intensive), Assessments = 2,632/9,273

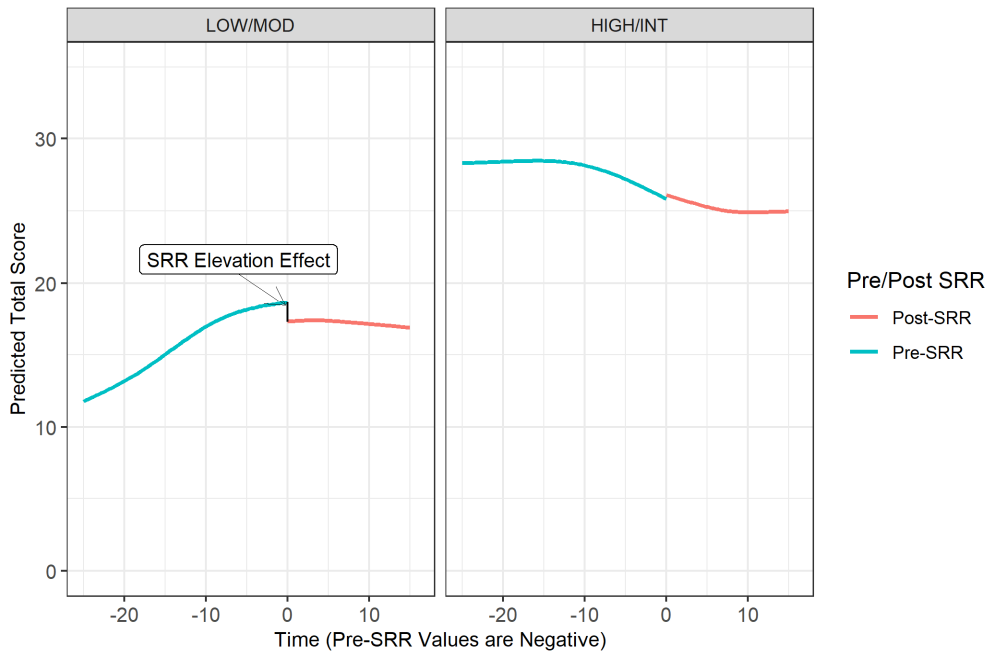
No effect for SRR was observed in the high/intensive baseline risk group, but, as seen in the figure, Time was significant and corresponds roughly to a cubic polynomial (4.02 – 1 = ~3). Because the smooth terms are interpreted graphically, one can discern from the figure that, overall, scores were

[18] Sullivan, K. J., Shadish, W. R., & Steiner, P. M. (2015). An introduction to modeling longitudinal data with generalized additive models: Applications to single-case designs. *Psychological Methods, 20*(1), 26-42. doi:10.1037/met0000020

[19] An exception to this rule occurs when the EDF value is less than three. A value of 1.0 is equivalent to a linear trend and a value of 0 means no effect.

decreasing prior to SRR. The rate at which scores were declining slowed soon after SRR and then leveled off thereafter. Again, because SRR was not a discrete event, some of the Time effect could be associated with SRR; specifically, scores in the high/intensive group began to decline at about 10 months before full implementation of SRR; however, SRR policies were being enacted during this period.

Figure 8: Pre- and Post-SRR LS/RNR Total Score Trajectories by Baseline Risk Category



Criminal History Domain Scores

As discussed above, the remainder of the models (those considering LS/RNR domain scores) utilized a binomial process model where each point on the scale was considered out of the total possible points one could accumulate. This class of models falls under a special case of regression known as generalized linear models. Unfortunately, the coefficients from these models are not interpreted the same as in a general linear model, and so they require some additional explanation below.

Because the models for the criminal history domain are linear models, rather than smoother models, they offer an opportunity to introduce generalized linear models without the additional complexities of spline-based models. Table 6 below shows the coefficients derived from a model of the criminal history domain scores over time. The column "Estimates" provides the expected change in the outcome (in log odds) given a one-unit increase in the predictor. This is identical to the linear models in the QA analyses except that the scaling is different. Because the outcome variable is not continuous, a link function (in this case the logit link) is used to transform the relationship between a predictor and the outcome to become linear on the transformed scale. A side effect of the rescaling means that the coefficient for SRR does not represent the change in points on the domain score; instead, the coefficient represents change in log odds.

Though not significant, consider the "Estimate" for SRR in the low/moderate group. The value of .04 means that the pre-SRR group has a score .04 higher (on the log odds scale) compared to the post-SRR group. Thinking in terms of log odds is not natural for most people, but these types of models also provide a statistic known as the odds ratio (OR)[20], which is a ratio of the probability in one group over the probability in another group. This statistic is easier to interpret. Again, though not significant, the OR of 1.04 for SRR in the low/moderate risk group indicates that the pre-SRR scores on the criminal history domain were 4% higher than the post-SRR scores.

An added benefit of ORs is that they are also effect sizes, but only in the case of a binary predictor. Here, SRR is a binary predictor because it can only take on two values; in this case, it is coded as 0 or 1 (pre- or post-SRR). In that case, one can roughly interpret ORs of 1.32, 2.38, and 4.70 as small, medium, and large, respectively[21]. Unfortunately, however, there is no similar interpretation for continuous predictors (e.g., Time) because the OR in that case depends on the scaling. Therefore, one can speak of theoretical importance of effects for the SRR predictor, but not time or the interaction.

The lack of significance for the Time predictor in these models is actually an important finding. Because this domain contains only one dynamic item (out of eight possible), not much change over time would be expected unless the population was changing in terms of their overall risk or risk was getting worse. For that reason, the finding of a significant interaction for the high/intensive group is somewhat surprising until one considers the size of the effect by looking at Figure 9 below[22]. In the figure, the effect appears relatively small and this is especially the case when one considers the extremely large sample size for the number of assessments in the high/intensive group (9,273; see table footnote). In the figure, one can see that the interaction is indicating a change in the slope of the trajectory post-SRR for the high/intensive group. Scores tend to level off (rather than rising slightly) after full implementation and this effect may be associated with SRR initiatives. The same trend is seen in the low/moderate group, but there it is not significant owing partly to the discrepancy in sample size (2,632 assessments in the low/moderate group compared to 9,273 in the high/intensive group).

Despite the general lack of significant effects for this domain, a couple of points about the figure should be noted. Unlike models with smoothers or splines, linear models throughout this section have counterfactuals shown by the dashed line that projects the pre-SRR trend line into the post-SRR period. As discussed above, this is possible in linear models because the projection is merely extrapolating the linear trend[23]. Like the figure below, all linear models in this section also have points (triangles for pre-SRR and circles for post-SRR) to facilitate interpretation. These points are spaced evenly over five-month intervals. Technically this is not necessary, but it is sometimes

[20] The odds ratio is also the exponentiated b-coefficient or "Estimate".

[21] Oliver, J., May, W. L., & Bell, M. L. (2016). Relative effect sizes for measures of risk. *Communications in Statistics - Theory and Methods, 46*(14), 6774-6781. doi:10.1080/03610926.2015.1134575

[22] Again, effect sizes are not available for the interaction, but one can visually judge the size of the effect given the scale (y-axis) of the figure.

[23] In some linear models, the reader will notice the pattern is not linear and may have a curve. This occurs because the models are built using the aforementioned logit link, which transforms the relationship so that it is linear on the log scale. When back-transformed to the original scale, the relationship does not have to be linear.

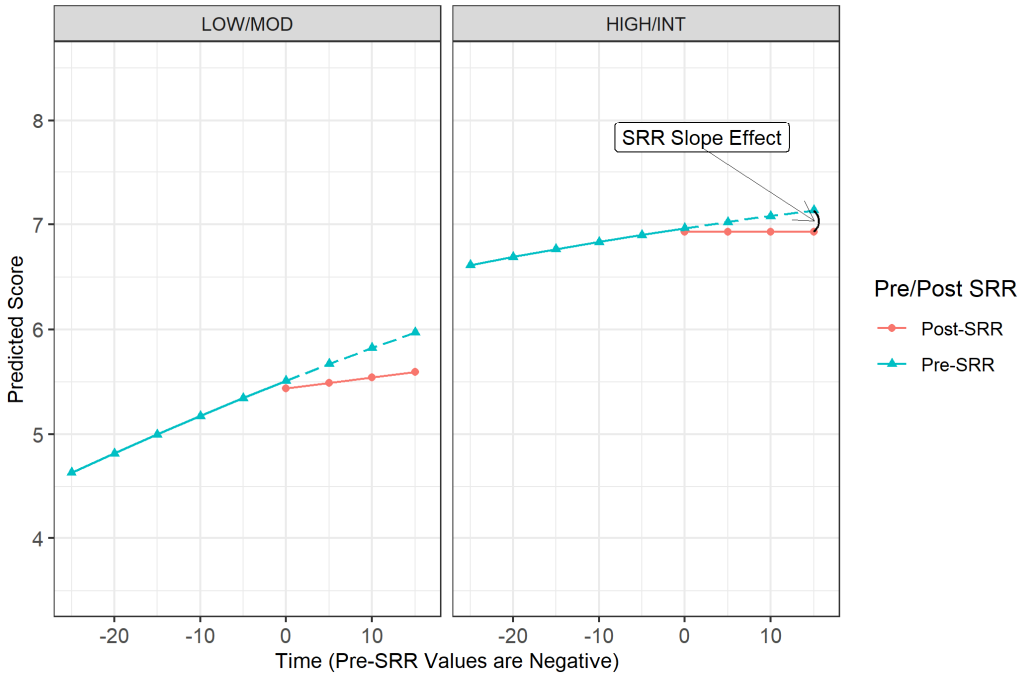
helpful as a stylistic choice. Use of these points is not helpful for spline-based models because they require a dense grid on the x-axis (e.g., estimates every month rather than every five months); in that case, use of the points would clutter the figure.

Table 6: Coefficients from Discontinuous Growth Model of Criminal History Scores

Location	Predictor	OR	Estimate	Std. Error	95% CI	P-value
Low/Moderate	(Intercept)	2.12	0.75	0.08	0.61 – 0.90	<0.001
	SRR	1.04	0.04	0.08	-0.11 – 0.20	0.587
	Time	1.01	0.01	0.01	-0.01 – 0.02	0.464
	Time * SRR	1.01	0.01	0.01	-0.00 – 0.03	0.150
High/Intensive	(Intercept)	6.50	1.87	0.04	1.79 – 1.95	<0.001
	SRR	1.03	0.03	0.05	-0.06 – 0.13	0.474
	Time	1.00	0.00	0.01	-0.01 – 0.01	0.978
	Time * SRR	1.01	0.01	0.01	0.00 – 0.02	**0.011**

N_{person} = 860 (low/mod)/2,778 (high/intensive), Assessments = 2,632/9,273

Figure 9: Pre- and Post-SRR LS/RNR Criminal History Score Trajectories by Baseline Risk Category



Education/Employment Domain Scores

Having established how one interprets the tables and figures for spline-based models (using the total scores above) and binomial process models (using criminal history scores), the rest of this section moves more quickly through model interpretation. For the education/employment model, a spline-based model was necessary for the low/moderate risk group owing to non-linearity, while a linear model worked well for the high/intensive group. For that reason, the low/moderate group model has no interaction, but the high/intensive model does. Scores on the scale range from zero

to nine and eight items are dynamic. Table 7 shows the coefficients from the models and Figure 10 shows the predicted values of the models.

Beginning with the low/moderate group, the effect for SRR is significant. The OR indicates a small effect and suggests that pre-SRR, the probability of endorsing any one item (that is, a score of one on an item) on the education/employment domain was 32% higher[24] than post-SRR; this is represented by the immediate elevation shift in the left panel of the figure. The effect for Time is also significant and shows that scores increased in the pre-SRR period, shift downward at full implementation, and then maintain a downward trajectory.

The model for the high/intensive group reveals what is, in some ways, an ideal pattern for SRR. Though the OR is small (1.14), SRR is associated with a significant downward shift in education/employment risk scores and also a significant change in the slope; that is, scores were declining pre-SRR, declined further at full implementation, and declined faster post-SRR. Because this domain is comprised of nine items, eight of which are dynamic, this domain offers one of the best opportunities to detect changes in risk that may be associated with SRR and, though they are small, they are also not negligible.

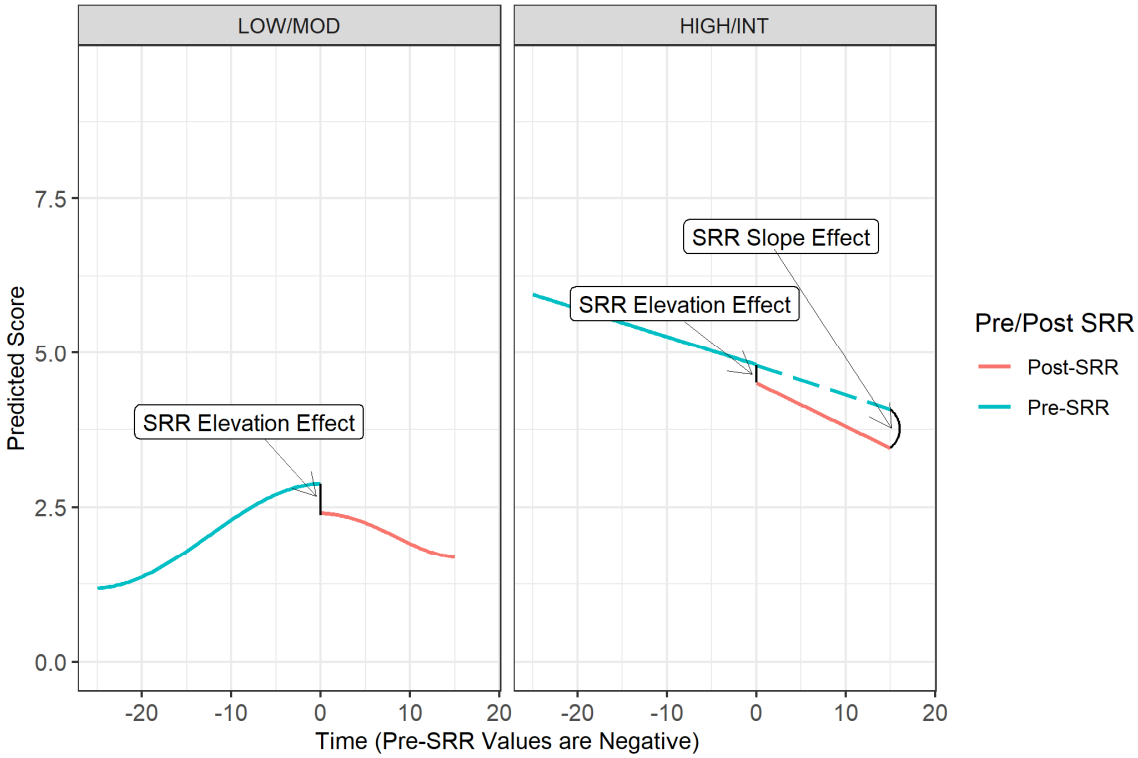
Table 7: Coefficients from Discontinuous Growth Model of Education/Employment Scores

Baseline Risk	Predictor	OR	Estimate	Std. Error	95% CI	P-value
Low/Moderate	(Intercept)	0.26	-1.36	0.09	-1.54 – -1.18	<0.001
	SRR	1.32	0.27	0.14	0.01 – 0.54	**0.045**
	Time		6.64			**<0.001**
High/Intensive	(Intercept)	1.00	0.00	0.03	-0.06 – 0.07	0.894
	SRR	1.14	0.13	0.03	0.06 – 0.20	**<0.001**
	Time	0.97	-0.03	0.00	-0.04 – -0.02	**<0.001**
	Time * SRR	1.01	0.01	0.00	0.00 – 0.02	**0.009**

N_{person} = 860 (low/mod)/2,778 (high/intensive), Assessments = 2,632/9,273

[24] Odds ratios can be quite large in some cases and can, at first glance, seem to indicate an enormous effect. They can be misleading, though, when baserates are very low or very high because they are a relative metric. For example, suppose a probability was 2% on an item in the pre-SRR period but 1% in the post-SRR period. On an OR scale, this represents a value of 2 (2/1) or a 200% increase in odds. The same difference near the middle of the probability scale would lead to a small OR. For example, a 51% probability pre-SRR and 50% probability post-SRR still represents a 1% difference, but now the OR is 1.02 (51/50), revealing only a 2% increased relative probability. This is not a flaw in ORs; when baserates are low, going from 1% to 2% can be an important effect. They just need to be interpreted with caution and with consideration of the baserates.

Figure 10: Pre- and Post-SRR LS/RNR Education and Employment Score Trajectories by Baseline Risk Category



Family and Marital Domain Scores

The family and marital domain scores were modeled using linear models. Scores on the scale range from zero to four and three items are dynamic. Table 8 shows the coefficients from the models and Figure 11 shows the predicted values of the models.

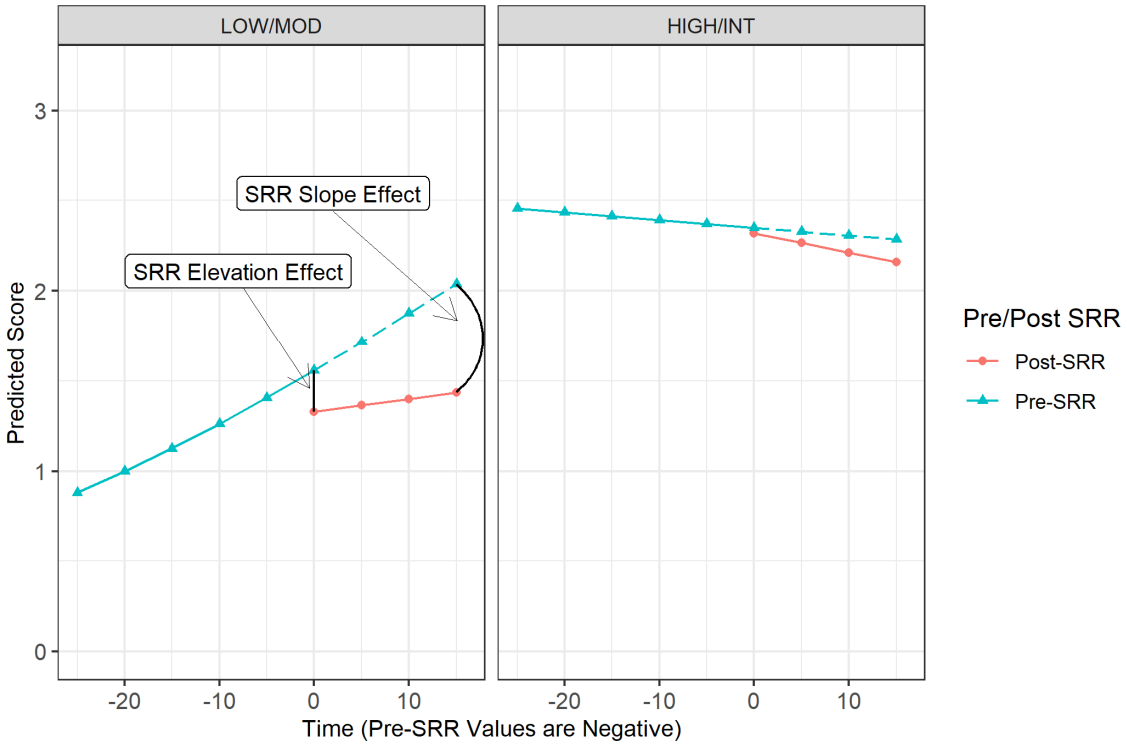
For the low/moderate group, family and marital domain scores revealed a significant effect for SRR and a significant interaction. The OR indicates a small effect and suggests that pre-SRR, the probability of endorsing any one item (that is, a score of one on an item) on the education/employment domain was 28% higher than post-SRR. The interaction shows that scores were increasing pre-SRR, but full implementation was associated with a reduction, and even a leveling off, in the rate of increase. The model for high/intensive cases revealed no significant effects.

Table 8: Coefficients from Discontinuous Growth Model of Family and Marital Scores

Location	Predictor	OR	Estimate	Std. Error	95% CI	P-value
Low/Moderate	(Intercept)	0.50	-0.70	0.09	-0.88 – -0.52	**<0.001**
	SRR	1.28	0.25	0.10	0.04 – 0.45	**0.017**
	Time	1.01	0.01	0.01	-0.01 – 0.03	0.488
	Time * SRR	1.02	0.02	0.01	0.00 – 0.05	**0.041**
High/Intensive	(Intercept)	1.38	0.32	0.05	0.23 – 0.41	**<0.001**
	SRR	1.03	0.03	0.05	-0.07 – 0.14	0.562
	Time	0.99	-0.01	0.01	-0.02 – 0.00	0.060
	Time * SRR	1.01	0.01	0.01	-0.01 – 0.02	0.297

N_{person} = 860 (low/mod)/2,778 (high/intensive), Assessments = 2,632/9,273

Figure 11: Pre- and Post-SRR LS/RNR Family and Marital Score Trajectories by Baseline Risk Category



Leisure and Recreation Domain Scores

The leisure and recreation domain scores were modeled using linear models. Scores on the scale range from zero to two and both items are dynamic. Table 9 shows the coefficients from the models and Figure 12 shows the predicted values of the models.

For the low/moderate group, leisure and recreation domain scores revealed a significant effect for SRR and a significant interaction. The OR indicates a small-medium effect and suggests that pre-SRR, the probability of endorsing any one item (that is, a score of one on an item) on the leisure

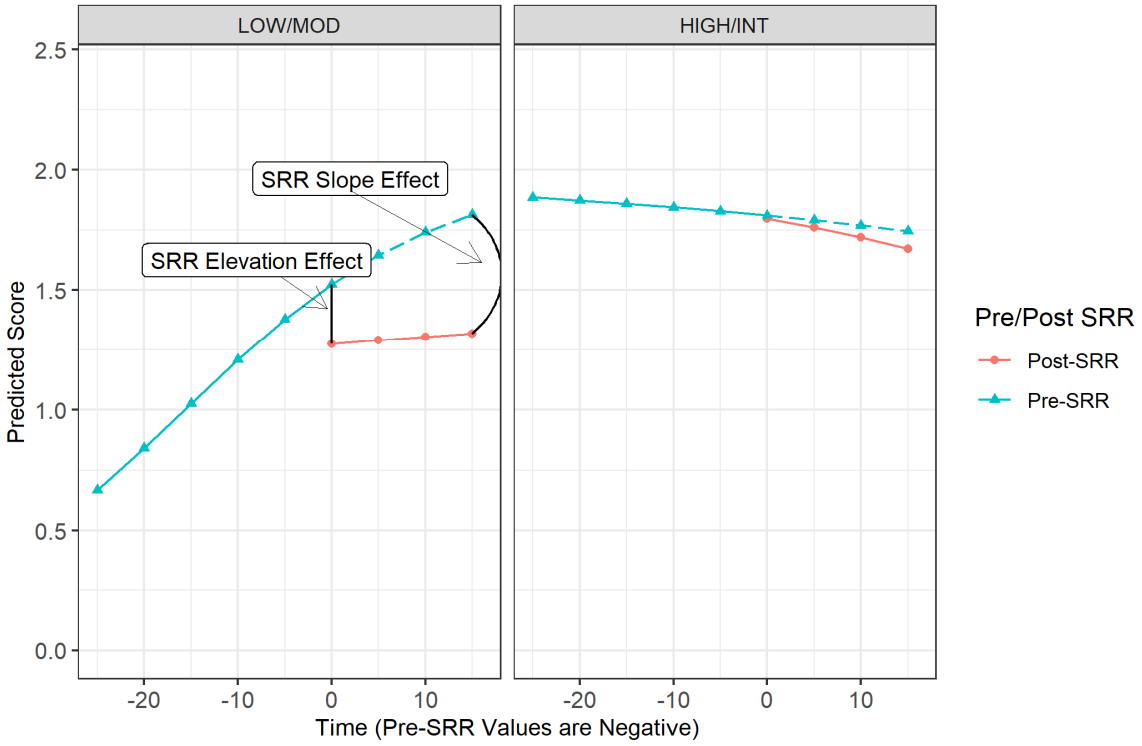
and recreation domain was 82% higher than post-SRR. The interaction shows that scores were increasing pre-SRR, but full implementation was associated with a reduction, and even a leveling off, in the rate of increase. The model for high/intensive cases revealed only a significant effect for Time. Ignoring SRR, scores were decreasingly slightly over time.

Table 9: Coefficients from Discontinuous Growth Model of Leisure and Recreation Scores

Location	Predictor	OR	Estimate	Std. Error	95% CI	P-value
Low/Moderate	(Intercept)	1.76	0.57	0.14	0.30 – 0.83	**<0.001**
	SRR	1.82	0.60	0.16	0.29 – 0.91	**<0.001**
	Time	1.01	0.01	0.02	-0.03 – 0.04	0.707
	Time * SRR	1.07	0.07	0.02	0.03 – 0.10	**<0.001**
High/Intensive	(Intercept)	8.84	2.18	0.09	2.00 – 2.36	**<0.001**
	SRR	1.08	0.08	0.10	-0.12 – 0.27	0.441
	Time	0.96	-0.04	0.01	-0.06 – -0.02	**<0.001**
	Time * SRR	1.01	0.01	0.01	-0.01 – 0.04	0.185

N_{person} = 860 (low/mod)/2,778 (high/intensive), Assessments = 2,632/9,273

Figure 12: Pre- and Post-SRR LS/RNR Leisure and Recreation Score Trajectories by Baseline Risk Category



Antisocial Companions Domain Scores

The antisocial companions domain scores were modeled using linear models. Scores on the scale range from zero to four and all four items are dynamic. Table 10 shows the coefficients from the models and Figure 13 shows the predicted values of the models.

For the low/moderate group, antisocial companions domain scores revealed a significant effect for SRR and a significant interaction. The OR indicates a small effect and suggests that pre-SRR, the probability of endorsing any one item (that is, a score of one on an item) on the leisure and recreation domain was 47% higher than post-SRR. The interaction shows that scores were increasing pre-SRR, but full implementation was associated with a leveling of the trend.

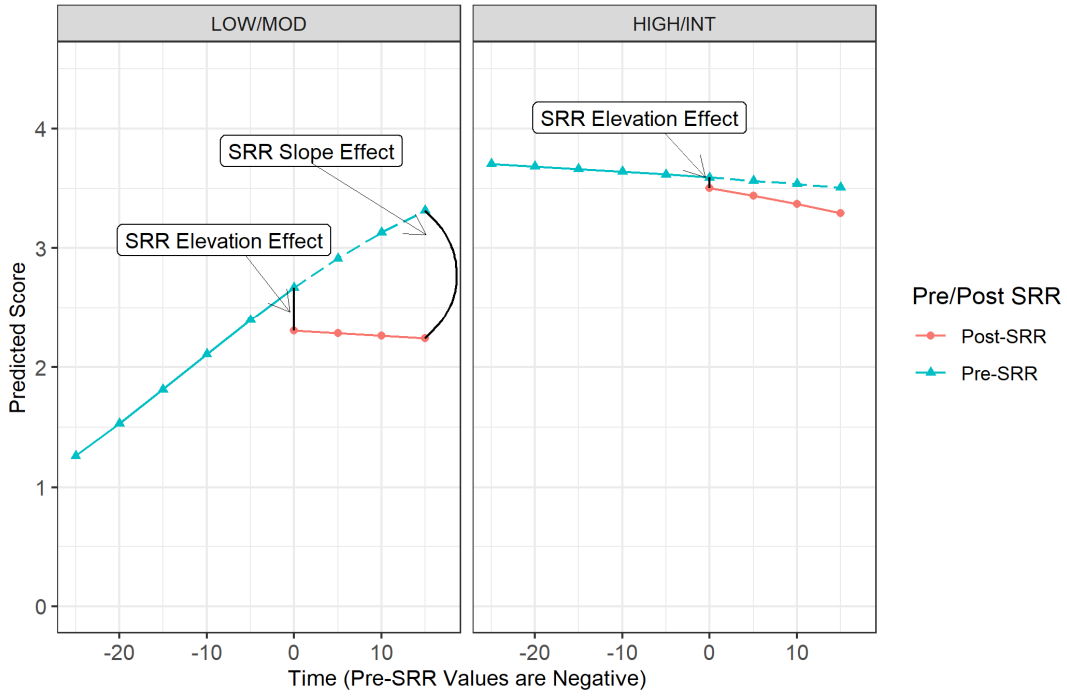
The model for high/intensive cases revealed a significant effect for SRR and Time. The OR for SRR indicates a small effect and suggests that pre-SRR, the probability of endorsing any one item (that is, a score of one on an item) on the antisocial companions domain was 24% higher than post-SRR. The effect for Time indicates that, ignoring SRR, scores were decreasingly slightly over time.

Table 10: Coefficients from Discontinuous Growth Model of Antisocial Companions Scores

Location	Predictor	OR	Estimate	Std. Error	95% CI	P-value
Low/Moderate	(Intercept)	1.36	0.31	0.10	0.11 – 0.50	**0.002**
	SRR	1.47	0.39	0.11	0.17 – 0.60	**<0.001**
	Time	1.00	0.00	0.01	-0.03 – 0.02	0.713
	Time * SRR	1.07	0.06	0.01	0.04 – 0.09	**<0.001**
High/Intensive	(Intercept)	7.07	1.96	0.07	1.82 – 2.09	**<0.001**
	SRR	1.24	0.22	0.07	0.08 – 0.35	**0.002**
	Time	0.97	-0.03	0.01	-0.04 – -0.01	**<0.001**
	Time * SRR	1.01	0.01	0.01	-0.00 – 0.03	0.076

N_{person} = 860 (low/mod)/2,778 (high/intensive), Assessments = 2,632/9,273

Figure 13: Pre- and Post-SRR LS/RNR Antisocial Companions Score Trajectories by Baseline Risk Category



Alcohol and Drugs Domain Scores

The alcohol and drugs domain scores were modeled using linear models. Scores on the scale range from zero to eight and six items are dynamic. Table 11 shows the coefficients from the models and Figure 14 shows the predicted values of the models.

For the low/moderate group, alcohol and drugs domain scores revealed a significant effect for SRR and a significant interaction. The OR indicates a small effect and suggests that pre-SRR, the probability of endorsing any one item (that is, a score of one on an item) on the alcohol and drugs domain was 40% higher than post-SRR. The interaction shows that scores were increasing pre-SRR, but full implementation was associated with a leveling of the trend.

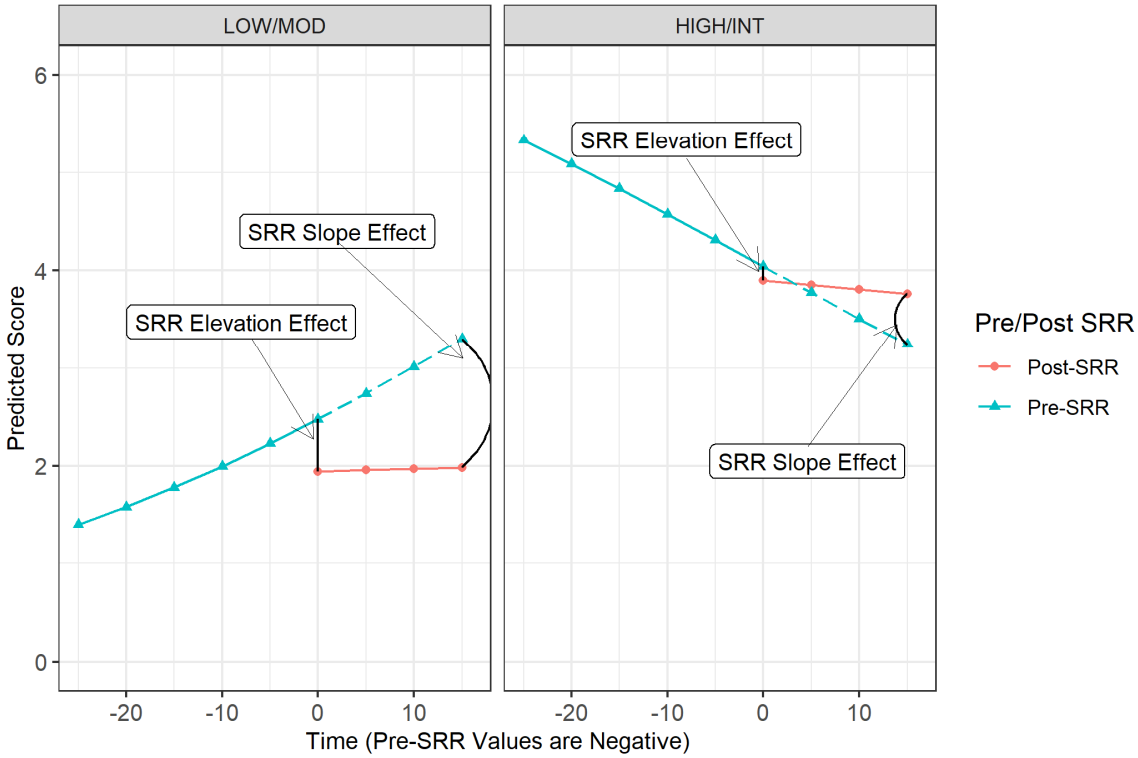
The model for high/intensive cases revealed a significant effect for SRR and a significant interaction, but an interaction that was in the unanticipated direction. The OR for SRR indicates a very small effect and suggests that pre-SRR, the probability of endorsing any one item (that is, a score of one on an item) on the alcohol and drugs domain was 7% higher than post-SRR. The significant interaction is seen more clearly in the figure. Notice that the annotation using the arc is inverted. This indicates that scores leveled off post-SRR, ceasing the rate of decline occurring pre-SRR. It is not immediately clear what would cause this effect, but scores were declining so rapidly pre-SRR that the high/intensive group was starting to reach average scores similar to the low/moderate group. It may be that the rate of decrease simply could not be maintained as it reached a lower bound for the baseline risk group.

Table 11: Coefficients from Discontinuous Growth Model of Alcohol and Drug Scores

Location	Predictor	OR	Estimate	Std. Error	95% CI	P-value
Low/Moderate	(Intercept)	0.32	-1.14	0.08	-1.28 – -0.99	**<0.001**
	SRR	1.40	0.33	0.08	0.18 – 0.48	**<0.001**
	Time	1.00	0.00	0.01	-0.01 – 0.02	0.822
	Time * SRR	1.03	0.03	0.01	0.01 – 0.05	**0.001**
High/Intensive	(Intercept)	0.95	-0.05	0.03	-0.11 – 0.01	0.123
	SRR	1.07	0.07	0.04	0.00 – 0.14	**0.048**
	Time	1.00	0.00	0.00	-0.01 – 0.00	0.254
	Time * SRR	0.98	-0.02	0.00	-0.03 – -0.01	**<0.001**

N_{person} = 860 (low/mod)/2,778 (high/intensive), Assessments = 2,632/9,273

Figure 14: Pre- and Post-SRR LS/RNR Alcohol and Drug Problems Score Trajectories by Baseline Risk Category



Procriminal Attitudes Domain Scores

The procriminal attitudes domain scores were modeled using spline-based models in the low/moderate group due to departures from linearity in the trajectories. The model for the high/intensive group was modeled using linear trajectories. Scores on the scale range from zero to four and all four items are dynamic. Table 12 shows the coefficients from the models and Figure 15 shows the predicted values of the models.

For the low/moderate group, only the effect for time is significant and the pattern in the figure shows that scores increased in the pre-SRR period, leveled off in the immediate post-SRR period, and then began to rise again. Notice that the OR for SRR is actually a medium-sized effect, but it is not significant. When this occurs, it is usually caused my notable uncertainty in the point-estimate. That is, the effect is associated with a great deal of variability or error in the estimation.

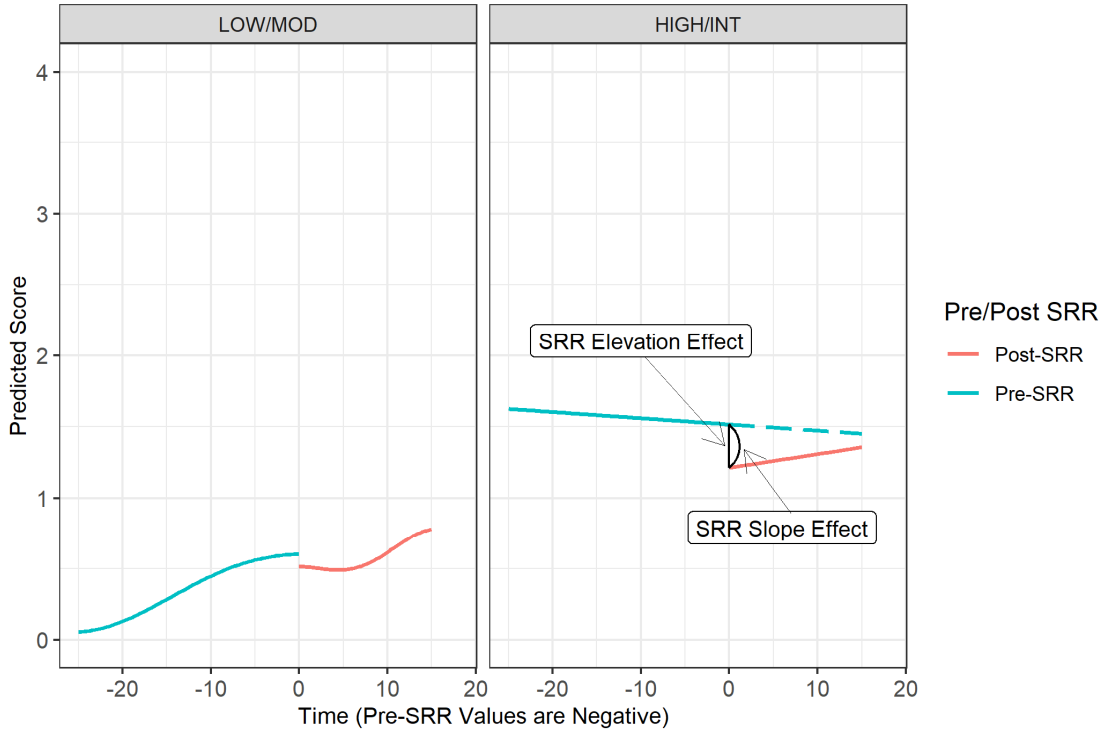
The model for the high/intensive group revealed a significant effect for SRR and a significant interaction. The OR for SRR revealed a small effect and is associated with a downward shift in scores. Pre-SRR, the probability of endorsing any one item (that is, a score of one on an item) on the domain was 40% higher than post-SRR. The interaction is in the unanticipated direction. Though SRR was associated with an immediate reduction is scores, the significant interaction seen in the figure shows that scores are again increasing in the post-SRR period.

Table 12: Coefficients from Discontinuous Growth Model of Procriminal Attitudes Scores

Baseline Risk	Predictor	OR	Estimate	Std. Error	95% CI	P-value
Low/Moderate	(Intercept)	0.06	-2.87	0.40	-3.65 – -2.09	**<0.001**
	SRR	2.19	0.78	0.47	-0.13 – 1.70	0.094
	Time	2.39	0.87	0.29	0.31 – 1.43	**0.002**
High/Intensive	(Intercept)	0.44	-0.83	0.05	-0.94 – -0.73	**<0.001**
	SRR	1.40	0.34	0.06	0.23 – 0.45	**<0.001**
	Time	1.01	0.01	0.01	-0.00 – 0.02	0.075
	Time * SRR	0.98	-0.02	0.01	-0.03 – -0.00	**0.018**

N $_{person}$ = 860 (low/mod)/2,778 (high/intensive), Assessments = 2,632/9,273

Figure 15: Pre- and Post-SRR LS/RNR Procriminal Attitudes Score Trajectories by Baseline Risk Category



Antisocial Pattern Domain Scores

The antisocial pattern domain scores were modeled using linear models. Scores on the scale range from zero to four and all four items are dynamic. Table 13 shows the coefficients from the models and Figure 16 shows the predicted values of the models.

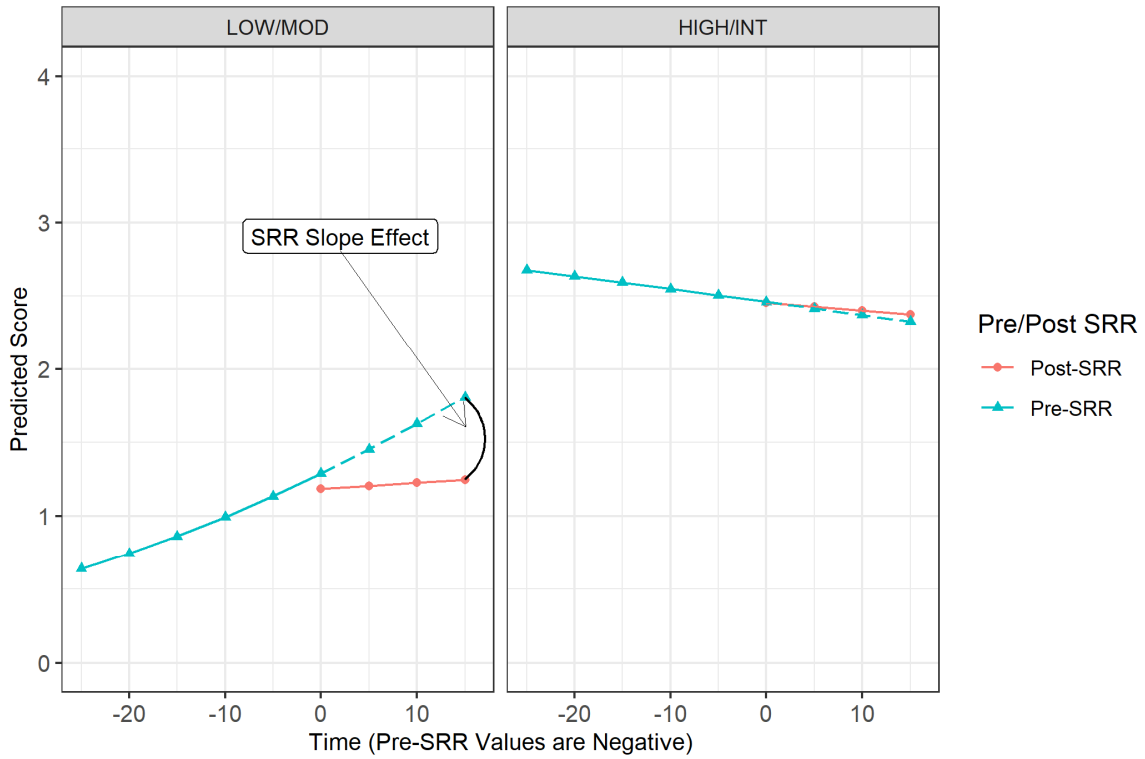
For the low/moderate group, antisocial patterns domain scores revealed a significant interaction but no main effect for SRR. Though SRR produced no immediate shift in scores, it was associated with a change in the slope. Scores were rising pre-SRR, but leveled off post-SRR. There were no significant effects in the high/intensive group.

Table 13: Coefficients from Discontinuous Growth Model of Antisocial Patterns Scores

Location	Predictor	OR	Estimate	Std. Error	95% CI	P-value
Low/Moderate	(Intercept)	0.42	-0.87	0.09	-1.04 – -0.70	**<0.001**
	SRR	1.13	0.12	0.10	-0.08 – 0.32	0.231
	Time	1.00	0.00	0.01	-0.02 – 0.03	0.657
	Time * SRR	1.03	0.03	0.01	0.01 – 0.05	**0.007**
High/Intensive	(Intercept)	1.59	0.46	0.04	0.37 – 0.55	**<0.001**
	SRR	1.01	0.01	0.05	-0.09 – 0.11	0.885
	Time	0.99	-0.01	0.01	-0.02 – 0.01	0.330
	Time * SRR	1.00	0.00	0.01	-0.02 – 0.01	0.516

N $_{person}$ = 860 (low/mod)/2,778 (high/intensive), Assessments = 2,632/9,273

Figure 16: Pre- and Post-SRR LS/RNR Antisocial Patterns Score Trajectories by Baseline Risk Category



Section Summary

This section covered a great deal of information related to LS/RNR risk and needs assessments and changes associated with implementation of SRR within UDC/AP&P. In considering the meaning of the results from this section, it is important to keep in mind the caveats mentioned at the start of the section. Primary among these would be that, because this was not an RCT, causation is not implied by the findings and one can speak only about associations between SRR and changes in risk and needs. It is also important to remember that SRR was not a discrete event; accordingly, some of the effect is captured in the main effects for SRR discussed above (which represents the

point of full implementation), but some of the effect is also captured by the Time variable and the interaction, since the program was implemented over time.

Some mixed results were found in the models above, but full implementation of SRR was often associated with significant drops in total and domain scores. These drops occurred either immediately or following full implementation and sometimes both. As mentioned above, the "ideal" pattern for SRR would reveal a main effect (elevation effect for SRR) in any model, an effect for Time in spline-based models, and an interaction in linear models. A caveat to this expectation is created by the fact that some scales, specifically criminal history and procriminal attitudes, have few dynamic items. In these cases, we would hope the patterns would not reveal scores becoming higher as a function of SRR.

Table 14 shows a descriptive summary of the findings from this section. It is an oversimplification of the results in that it neglects effect sizes and some of the additional details provided in the full models above. However, hopefully it is helpful in better understanding the patterns that were observed across all outcomes (total and domain scores).

In the table, each outcome, for each group, has two effects of interest. In this simplified presentation, a point is awarded for an outcome when either of the effects of interest were significant in the anticipated direction. Two points are awarded if both effects are significant. Because there are eight domains and a total score, the most points earned in this simplified system would be 18 for each baseline risk group.

In the spline-based models, the main effect for SRR and the effect of Time are of greatest interest. In the linear models, the main effect of SRR and the interaction are of greatest interest. Note that two effects (documented above) were significant in the wrong direction. Because these effects do not support SRR in the presumed manner, they are not counted using the point system in the table below; they are also not subtracted from the point total, however. Main effects for SRR are also known as elevation effects. These are denoted in the table by an "E". Interaction effects are denoted by an "I" (these apply to linear models) and main effects for Time are denoted by "T"; again, effects for Time apply to spline-based models.

Though the table simplifies a series of complicated models, it does reveal some interesting patterns. Most notable among those is the fact that most of the effects associated with SRR occurred in the group that was low or moderate risk at baseline. While some favorable effects were observed in the group that was high or intensive risk at baseline, half as many favorable effects associated with SRR occurred in this group. If one were to penalize SRR for the two interactions in the wrong direction (e.g., by subtracting a point for each), even fewer favorable effects would be observed.

Table 14: Number of Outcomes Favoring SRR by LS/RNR Total/Domain Score and Risk Group

	Low/Moderate		High/Intensive	
Outcome	Effects	Points	Effects	Points
Total Score	E	1	--	0
Criminal History	--	0	I	1
Education/Employment	E,T	2	E,I	2
Family/Marital	E,I	2	--	0
Leisure/Recreation	E,I	2	--	0
Antisocial Companions	E,I	2	E	1
Alcohol/Drugs	E,I	2	E	1
Procriminal Attitudes	T	1	E	1
Antisocial Patterns	I	1	--	0
TOTALS		13/18		6/18

Note: E = significant main effect for SRR; T = significant main effect for time; I = significant interaction effect

This finding is somewhat surprising because most treatment efforts and resources go into treating the highest risk individuals, so one would expect the greatest effects to occur for these people. They also have the greatest room for improvement because they start out higher on the total and domain scores relative to their lower risk counterparts. Despite these facts, there are reasons to be optimistic about the effects associated with SRR.

First, the models followed individuals from their baseline score. Higher risk individuals are more likely to be incarcerated for longer periods than lower risk ones and certain changes in risk levels cannot occur during incarceration. For example, an incarcerated individual will always receive a point for having "some criminal acquaintances" on the antisocial companions scale by virtue of being incarcerated. Second, treatment is a longer and more involved process for higher risk individuals. Changing certain patterns related to attitudes and behaviors inevitably takes longer for a group in need of more diverse and more intensive services. Because SRR had only reached full implementation for one year before data for this report were requested, it may be the case that changes will be seen as longer follow-up periods are observed.

A final point of interest from the above models involves the pre-SRR trajectories for both baseline risk groups. Reviewing all of the figures in this section, the trends associated with Time generally indicate increasing risk in the pre-SRR period for the low/moderate group, but decreasing risk in the pre-SRR period for the high/intensive group. To the extent that SRR was being implemented over time in the pre-SRR period, it is not immediately clear why changes should be associated with increasing risk among lower risk offenders, but decreasing risk among higher risk ones. It will be important to continue to observe these trends in future analyses.

Timing of Level of Service – Risk, Need, Responsivity (LS/RNR) Assessments

Purpose

Adoption of the SRR is theorized to improve the timing of LS/RNR assessments so that they can be appropriately utilized in programming decisions, including case action plans and community-based interventions. UDC set goals of improving the frequency with which assessments occurred within 60 and 90 days of release.[25]

Analytic Approach

LS/RNR assessments were provided to UCJC by UDC staff. For purposes of the analyses that follow, anyone who was paroled from prison before July 1, 2019 (the date chosen to represent full SRR implementation) was considered as pre-SRR, while those who were paroled after that date were considered post-SRR. The sample provided by UDC was selected based on parole start dates from November 2018 through February 2020. This date range was intended to provide two cohorts of nearly equal length. The first cohort extended from November 2018 to the date of full SRR implementation, June 30, 2019. The second cohort, the post-SRR cohort, extended from July 1, 2019 through February 2020.

Because people could be released from prison to parole more than once, there was some duplication across people. As described in the previous section, this creates a dependency in the data where multiple releases are nested within persons. However, the dependencies created in this section were quite different from other sections. Consider an analysis of a person's LS/RNR scores over time as an example. It makes intuitive sense that a person's LS/RNR risk score should be more similar to his or her own score across multiple assessments than to someone else's score. That creates a data dependency. Here, however, whether a person receives an assessment near release is determined by AP&P and not the individual. For that reason, it is difficult to think of a scenario where, despite multiple releases for some people, the determination of whether an assessment was received was person-driven. Because a dependency still *might* exist, multilevel models were run to examine the possibility. The models revealed near zero person-level variance and, accordingly, single-level models were adopted.[26]

Outcome data for this metric are binary in nature; that is, a person either received an assessment within 60 or 90 days of release or did not. For analysis purposes, the window extended from 60 or 90 days prior to release to either 60 or 90 days after release. When an assessment was received within a window (60 or 90 days), the outcome was coded as a one; when an assessment was not received in the window, the outcome was coded zero. This type of outcome is suitable for binary logistic regression and models that follow utilized this approach.

---

[25] Pending future funding, subsequent reports will also examine the timing of assessments around prison starts.

[26] Preliminary models revealed a singularity for the multilevel models created by the near zero variance at the person-level. In order to obtain an estimate, the models were rerun using Bayesian methods. Bayesian methods tend to perform better when estimates are near extreme boundaries: zero in this case. The Bayesian models ran without an error but also revealed near zero variance at the person-level, which indicated single-level models could be appropriately adopted.

Analytic Results

The results of the modeling process for the timeliness of LS/RNR assessments within 60 and 90 days of release are provided in Table 15. The column marked "Predictor" shows the predictors included in the model; in this case, the only effect of interest in the variable "SRR". It is coded such that a value of 0 represents pre-SRR and 1 represents post-SRR. This means that the coefficients seen under the "OR" and "Estimate" columns reflect improvements associated with SRR.

The interpretations of the coefficients are the same as in the prior section covering client trajectories. Specifically, the "Estimate" column contains the change in log odds associated with SRR and the effect on the probability of receiving an assessment within 60 or 90 days of release. As before, the column "OR" contains the odds ratio, which is generally easier to interpret than the log odds. For the 60 day outcome, the OR of 1.17 indicates that, post-SRR, there was a 17% increased probability of receiving an assessment associated with SRR. For the 90 day outcome, the OR was 1.18, suggesting an 18% increased probability of receiving an assessment associated with SRR.

As mentioned above, when the predictor is binary (here, pre-SRR or post-SRR) ORs can also be interpreted as effect sizes. One can roughly interpret ORs of 1.32, 2.38, and 4.70 as small, medium, and large, respectively.[27] By those standards, the effect sizes associated with SRR are fairly small at this point. Future analyses (funding permitting) intend to examine a third, long-term cohort. By that time, a larger effect for SRR should be apparent if it is having the intended effect of improving the rate of assessments completed in a timely manner.

Table 15: Coefficients from Models Examining whether an Assessment was Received within 60 or 90 Days of Prison Release by Pre- or Post-SRR cohort

Location	Predictor	OR	Estimate	Std. Error	95% CI	P-value
60 Days	(Intercept)	--	0.14	0.04	0.05 – 0.22	**0.003**
	SRR	1.17	0.16	0.06	0.03 – 0.28	**0.013**
90 Days	(Intercept)	--	0.60	0.05	0.51 – 0.69	**<0.001**
	SRR	1.18	0.16	0.07	0.03 – 0.29	**0.014**

N_{person} = 3,676, $N_{yes/no\ on\ assessment}$ = 4,116

Logistic regression models can also produce predicted probabilities, which, in this case, are interpreted as the probability of receiving an assessment in the period pre-SRR or post-SRR. In the case of the model for assessments with 60 days, the predicted probability of receiving an assessment within 60 days is 57% post-SRR and 53% pre-SRR. For the 90 days model, the values are 68% and 65%, respectively. These relatively small differences in predicted probabilities underscore the currently small effect sizes associated with SRR.

[27] Oliver, J., May, W. L., & Bell, M. L. (2016). Relative effect sizes for measures of risk. *Communications in Statistics - Theory and Methods, 46*(14), 6774-6781. doi:10.1080/03610926.2015.1134575

Issues with Case Action Plans

Adoption of the SRR is expected to improve the timing of when Case Action Plans (CAPs) are generated and updated so that they can inform agents about their client's educational, program and treatment needs. SRR is also expected to improve the alignment of LS/RNR priorities identified using the risk assessment with goals in the CAPs. For example, if an individual's top priority, as identified by the LS/RNR, was antisocial patterns, we would expect to see CAP goals (including action steps, classes, and programming) that specifically targeted that prioritized need.

UCJC originally intended to examine both issues above: CAP timing and CAP-to-LS/RNR alignment. However, a preliminary examination of the data indicated some inconsistencies with, as well as changes in, how CAP events are recorded/tracked by UDC. The adoption of SRR has facilitated a number of changes in how UDC records CAP-related information. For example, beginning in October 2019, UDC made a push to increase the use of CAP action step goals. During this change UDC folded in information that previously was tracked in the classes and classes/programs data into the CAP action step goals. Other events that would previously have counted as an action step were no longer counted after system changes.

The net effect of the changes in UDC's record system regarding CAPs was that, because the changes overlapped with (or coincided with) SRR initiatives, the pre-SRR system was not directly comparable to the post-SRR system. This meant that an analysis that compared CAP timing or CAP-to-LS/RNR alignment, would not be comparing "apples" to "apples" so to speak. After discussions with UDC, both UDC and UCJC decided an analysis of CAP timing[28] and CAP-to-LS/RNR alignment would not provide a meaningful test of the effect of SRR because of the changes in how CAP events were recorded.[29]

Given the data limitations associated with the CAP data, combined with the importance of understanding how SRR changed CAP-related outcomes, UDC and UCJC discussed two analytic strategies that could be utilized for subsequent reports (pending future funding). The first approach would be to conduct the analyses using generalized additive mixed modeling (GAMMs). In the current instance, these models could be used to track changes in CAP timing and LS/RNR alignment over time after the records' system is more fully implemented. This method would allow the evaluation to address the issues associated with the changes in data collection between the pre-SRR cohort and post-SRR cohort by removing the cohort variable from the analysis and inserting Time as the only predictor. In this analysis, the findings would not speak to changes in CAP timing or CAPs mapping onto LS/RNR priorities between the two cohorts. Rather, the analysis would model these outcomes as a function of time, under the assumption that CAP timing and LS/RNR

[28] The issues around CAP events also impact CAP timing in conjunction with LS/RNR timing. Because fewer CAP events now qualify, fewer events occur within a reasonable window around the LS/RNR assessment. Relative to the pre-SRR period system, this artificially decreases the number of CAP events that can occur in close proximity to an updated LS/RNR assessment.

[29] In order to determine whether the suspected concerns did occur in the data, UCJC conducted the two analyses as planned. The analysis indicated the number of CAP events varied considerably between the pre- and post-SRR periods. The fact that significantly more events qualified as CAP-related events in the pre-SRR period masked the effect of SRR. This outcome is likely an artifact of the changes in the types of events that qualify as a CAP-related event in the current system.

alignment should continue to improve over time even after full SRR implementation. Tentative evidence supporting this hypothesis can be found in the QA scores' section of this report. In those models, QA scores showed an elevation effect associated with SRR, but the slopes also indicated continued improvement following SRR implementation. GAMMs would permit an evaluation of whether there were changes over time and whether those changes followed a non-linear trajectory.

A second approach would consist of excluding the pre-SRR cohort from the analysis and adding a long term post-SRR cohort (which was originally planned for all analyses provided future funding to support additional phases of evaluation). Removing the pre-SRR cohort would address the issue of the changes that were made to the data collection process and criteria for CAP events between the pre-SRR cohort and post-SRR cohort. The modeling approach would remain the same for the CAP timing models and CAP mapping models as described in the sections above. To analyze CAP timing, the evaluation would estimate the effects of whether a CAP event occurred within 60 and 90 days of release on cohort (i.e., short-term post-SRR and long-term post-SRR) using binary logistic regression. In the CAP mapping models, we would estimate the effects of the count of LS/RNR priorities met on cohort using a count model (e.g., Poisson, negative binomial regression). In these analyses we would expect an improvement in the CAP timing and CAP mapping models for the long-term post-SRR cohort compared to the short-term post-SRR cohort.

There is no clear preference for either approach and, realistically, both have merit. They are also not mutually exclusive analyses and both can be performed in a single model. Similar to the discontinuous growth models above, the effect of cohort (short-term post-SRR or long-term post-SRR) can be examined with a binary predictor. Time can also be included in the models in order to discern whether, above any cohort effect (post-SRR only), continued improvements in SRR-related initiatives are observed as a function of Time.

Discussion

Summary of Findings

This report set out to evaluate several outcomes related to SRR implementation. These included:

1. Examination of whether risk/needs scoring fidelity (quality assurance), overall and across UDC Adult Probation and Parole locations, improved from pre- to post-SRR adoption;
2. Examination of whether the frequency of risk/needs assessments administered within 60 and 90 days of release pre- and post-SRR differed;
3. An analysis of whether client risk/needs assessment (overall and by domain-specific score) trajectories improved (owing to improved services) between pre- and post-SRR periods;
4. Examination of whether the case action plans (CAPs) occurring/updated within 60 and 90 days of release improved between the pre- and post-SRR periods; and
5. An analysis of whether case action plans' alignment with needs identified by the LS/RNR risk assessment improved from pre- to post-SRR adoption.

Unfortunately, data limitations (discussed above) made an analysis of objectives four and five untenable at this stage of SRR-implementation. The first three objectives were addressed, however, and yielded a number of favorable associations between these outcomes and SRR.

While reviewing the summary, it important to keep in mind the caveats set forth at the start of this report. First, the evaluation examines changes in patterns associated with the adoption of SRR. While it is tempting to assume a causal relationship between any improvements and adoption of SRR, the design is not a randomized control trial (RCT) and cannot infer such causality. To the extent that patterns above were consistent, one might infer an effect was reasonably associated with SRR. However, it remains true that factors aside from SRR also differ across time; these include, as examples, person (i.e., different clients) and system-level (i.e., different policies) factors. These cannot be ruled out as alternative explanations to improvements associated with SRR.

It is also important to remember that SRR and related policies were not adopted as a discrete event. Clearly not all changes related to SRR were adopted on July 1, 2019. Instead, it was an implementation process where related changes were expected to be fully in place by that date. To the extent that some policies were enacted as a continuous process, and before full implementation, the effect of the variable Time might also contain some of the effect of SRR.

Changes in LS/RNR Quality Assurance Scores

Though causality cannot be assumed, changes in QA scores associated with the timing and implementation of SRR were observed. In the model including all AP&P locations, a small effect size was observed for SRR and a near small effect size for the interaction, indicating QA scores not only improved at full-implementation, but also continued to improve more rapidly after implementation. Notably, the location-specific models (conducted for Region Three and Four Field offices) revealed larger effects than the overall model that combined locations. Because these two regions are the largest in terms of LS/RNR assessors, one can infer that much of the

improvement in QA scores associated with both time and SRR was due to assessors and QA improvements in these locations. Evidence from these modeled regions suggests that it is not only training, but good training and practice that matters, as newer assessors (that is, newer to a location) did not do as well on QA assessments. The fact that scores are generally higher in Regions Three and Four may be of some practical value to UDC if assessors who perform well in this region can also be used as coaches in the future.

Changes in LS/RNR Client Trajectories

The analysis of changes in LS/RNR trajectories revealed some mixed findings. Most of the favorable effects associated with SRR occurred in the group that was low or moderate risk at baseline. While some favorable effects were observed in the group that was high or intensive risk at baseline, half as many favorable effects associated with SRR occurred in this group. This finding is somewhat surprising because most treatment efforts and resources go into treating the highest risk individuals, so one would expect the greatest effects to occur for these individuals. They also have the greatest room for improvement because they start out higher on the total and domain scores relative to their lower risk counterparts.

Despite these facts, there are reasons to be optimistic about the effects associated with SRR. First, the models followed individuals from their baseline score. Higher risk individuals are more likely to be incarcerated for longer periods than lower risk ones and, as noted in the body of the report, certain changes in risk levels cannot occur during incarceration. Second, treatment is a longer and more involved process for higher risk individuals. Changing certain patterns related to attitudes and behaviors inevitably takes longer for a group in need of more diverse and more intensive services. Because SRR had only reached full implementation one year before data for this report were requested, it may be the case that changes will be seen as longer follow-up periods are observed.

Timing of Level of Service – Risk, Need, Responsivity (LS/RNR) Assessments

Though the effects were fairly small, SRR was associated with a 17% increased rate of LS/RNR assessment within 60 days of release and an 18% increase within 90 days relative to the pre-SRR period. Future analyses (funding permitting) intend to examine a third, long-term cohort. By that time, a larger effect for SRR should be apparent if it is having the intended effect of improving the rate of assessments completed in a timely manner.

Overall Summary

When combined, the results above suggest SRR has had a positive association with many of the outcomes that were examined in this report and there is preliminary support that the program has achieved many of its intended goals. At present, the results remain tentative as further analyses (discussed next) are needed to help provide greater depth and context regarding the current findings.

Next Steps

Providing a more comprehensive evaluation of long-term targets will be important to determining the overall efficacy of SRR initiatives on intended goals. Expanding the evaluation to include additional goals as well as allowing more time since SRR implementation will help provide a more complete picture of the program's effect.

One of the most important aspects of the program was not addressed in this report because of the short follow-up period since SRR implementation; however, the issue of recidivism will be a primary consideration in a Phase Two report, funding permitting. The Phase Two report would examine changes in recidivism based on three cohorts (pre-SRR, short-term post-SRR, and long-term post-SRR) as well as over time. As mentioned in the "Background" section of this report, UDC has set a goal to reduce statewide recidivism by 10% within the first two years of implementing SRR and by 25% within a 5-year period. While the five-year goal will remain beyond the reach of a Phase Two evaluation, whether UDC has met the 10% reduction goal can be assessed as well as progress toward meeting the longer-term, five-year goal.

Another goal remaining to be investigated in a Phase Two report is changes in treatment. At the time of this report, UDC did not have a fully-implemented system of tracking treatment and treatment dosage, but UDC recognizes the importance of such a system and is in the process of developing one. An analysis of treatment and treatment dosage is clearly important for providing context to the results from the current report. Specifically, changes in LS/RNR scores should be connected to the appropriateness and dosage of treatment received. Clients who complete treatment (at the appropriate dosage), for example, should reveal greater risk reduction over time. If that were not the case, the changes observed above in LS/RNR client trajectories would not be particularly meaningful because they would lack an associated cause. It should also be the case that higher risk clients should benefit more notably from SRR because they should receive more (by dosage) and more intensive services.

Yet another goal, and one initially set forth by the current project, remains to be addressed in a future, Phase Two evaluation. As mentioned above, data quality issues prevented an analysis of both CAP timing and how well LS/RNR-identified priorities mapped onto subsequent CAP events (e.g., action steps, classes, etc.). These issues will be addressed in a future report, funding permitting, once UDC has a more fully developed system for tracking CAPs and CAP events.