March 17, 2017

New York State Pay for Success (PFS) Project: "Employment to Break the Cycle of Re-Incarceration"

Phase I Validation Report

Michael J. Puma
MIKE PUMA ASSOCIATES, LLC

This workforce product was funded by a grant awarded by the U.S. Department of Labor's Employment and Training Administration. The product was created by the grantee and does not necessarily reflect the official position of the U.S. Department of Labor.

Table of Contents

"New York State Pay for Success (PFS) Project: Employment to Break the Cycle of Re-Incarceration"

Phase I Validator Report

Section I: Project Overview

Project Purpose

In New York an estimated 22,000 men and women return each year to their communities from the State's prisons. The labor market challenges facing these formerly incarcerated individuals are substantial, which increases their risk of re-incarceration, raises concern about public safety, and imposes costs on the social service system. As stated in the project proposal, *"Not only do formerly incarcerated individuals face the stigma of a criminal conviction when applying for jobs, but they often face other obstacles, including a lack of basic education and occupational skills, limited or no work history, minimal family and community supports, and poverty. Further compounding these limitations are the obligations many people coming home from prison must satisfy that can compete with the demands of full-time employment……. such as drug treatment attendance, curfews or other restrictions on mobility…."*

To help deal with this problem, the New York State PFS project sought to improve post-incarceration outcomes for "high-risk" male parolees[1] who were released to community supervision in Rochester or New York City (NYC) from the State's prison system. Through the use of targeted early intervention services, these individuals were expected to be more likely to obtain gainful employment and less likely to re-offend.

The supportive services were delivered by the Center for Employment Opportunities, Inc. (CEO), an evidence-based service provider currently operating in New York State. The intervention began with an initial orientation meeting, after which individuals could receive five days of life skills training, followed by placement in a subsidized transitional job.[2] These jobs were operated and supervised by CEO staff and generally consisted of providing maintenance, janitorial or grounds-keeping services to local public institutions or to private companies. Such supported employment was intended to give the parolees an opportunity to learn important job skills that would enable them to subsequently join the workforce, provide them with legitimate income during the critical early post-release period, and provide a documented and marketable work history. It also allowed CEO staff to identify and address workplace problems before participants were moved into the traditional labor market.

After an average of nine weeks of engaging in transitional jobs (typically spread over a period of about four months), CEO then attempted to move the parolees into unsubsidized jobs accompanied by post-placement support, including work-related counseling, crisis management,

[1] The identification of "high-risk" individuals was done using the State's existing COMPAS assessment instrument to determine each individual's level of risk of re-incarceration at the time of his release from prison to community supervision. COMPAS assesses risk on a range of factors including age, prior criminal history, identified antisocial attitudes, known criminal associates, the quality of family and marital relationships, and levels of educational and vocational attainment.

[2] These jobs are required to be within New York State.

and help with long-term career planning. CEO also implemented an incentive-based work retention program, which provided monthly bonuses to individuals who met employment milestones.

Overall Research Design

To assess the effectiveness of this intervention, the PFS project implemented a Randomized Controlled Trial (RCT) design[3] in which individual eligible parolees were identified prior to release to community supervision (i.e., participants were selected and randomized while they were still in prison), and assigned at random to one of two groups—a "Treatment" or intervention group that was systematically referred to CEO by their assigned Parole Officers,[4] or a "Control" group that was not specifically referred to CEO. Control Group members were handled using existing post-release procedures in Rochester and the selected Bureaus in New York City, the two sites selected for inclusion in the PFS project.

The strength of this design is that, if implemented properly, there will, on average, be no systematic differences between the two groups. This allows the calculation of the causal impact of the intervention to answer two primary research questions:

- Does the intervention increase high-risk male Treatment Group participants' employment outcomes in the 4th calendar quarter following their release from prison, as compared to Control Group participants? The "Population Outcome"[5] in this domain, measured at the individual level, was whether the individual was employed in a subsidized or unsubsidized job. [6]

- Does the intervention decrease high-risk male Treatment Group participants' recidivism over the Phase I implementation period, as compared to Control Group participants? Population Outcomes in this domain, measured at the aggregate group level,[7] included: (1) total prison "bed days" observed for each group; (2) total jail "bed days" observed for each group; and, (3)

[3] The study plan incorporates separate design and analysis procedures to be used in the event of two possible situations: (1) early termination of the project, or (2) if the RCT failed to achieve a high enough level of random assignment compliance. These options are described in Appendix A. Neither option was, however, invoked in Phase I of the project.

[4] Parolees are assigned to a parole area office based on their expected place of post-release residence. To minimize the risk that individuals in the Control Group were served by CEO, and to ensure that the service slots funded by this initiative were incremental to CEO's existing contracts with the State, parole area offices that referred minimal numbers to CEO were selected as target offices for this project.

[5] The term "Population Outcomes" used in this document is specific to this project and refers to the sum of observed individual participant outcomes calculated across all individuals randomly assigned to the Treatment or Control Groups, respectively (see Appendix A).

[6] Individuals may lack a social security number (SSN), or have an incorrect SSN, making it impossible to successfully match administrative data necessary to measure employment outcomes. At the start of the project, information provided by the Department of Corrections and Community Supervision (DOCCS) indicated that for individuals meeting the eligibility criteria for this project 12% were estimated to lack a valid SSN. According to the study plan, "Because employment-based payments are a small portion of overall Project payments, individuals will not be screened for SSN availability or validity. Instead, employment-related payments will be adjusted to account for the inability to obtain earnings records for individuals for whom a SSN is not available at the time of random assignment."

[7] These outcomes are calculated for the entire Treatment or Control Group regardless of when in Phase I individuals were released from prison, but for at least12 months after release. The estimates are, therefore, an average for each study group covering a range of individual post-release time periods.

remaining sentences for study members who were incarcerated during the Phase I observation period.

The PFS project is being conducted in two distinct phases -- Phase I (the subject of this report) ran from 12/23/2013 through 10/31/2016, and Phase II (currently underway) runs from 10/1/2015 through 10/31/2018. Separate samples, and the subsequent estimation of outcomes, are used in each Phase. [See Appendix A for complete details on the project plans.]

This report focuses only on Phase I; a second report will be issued at the end of Phase II.

Study Sample Recruitment

As noted above, the target population for the RCT was high-risk males expected to be released to community supervision from facilities operated by DOCCS during the period of Phase I. This identification of study participants was done while the individuals were still incarcerated using available administrative records.

For the purpose of this project, high-risk individuals were defined as those with Supervision Level 1 or 2 as assessed by the COMPAS risk-assessment tool[8] one to five months before release from prison.[9] These two Supervision Levels contained approximately half of those individuals released to community supervision.

In addition to having a Supervision Level of 1 or 2, eligible study participants also had to meet the following criteria: (1) have a predicted release date within 28 days at the time of random assignment; (2) scheduled for release from the Queensboro correctional facility to NYC, or assigned to one of four NYC parole area offices (Queens I, Queens II, Bronx I, or Brooklyn V),[10] or assigned to Rochester; (3) have at least six months of community supervision remaining at the time of release; and, (4) have a projected age at release equal to or greater than 17 years and 11 months. Individuals were also excluded if they were a sex offender or an arsonist, if they had an identified serious mental illness, were an undocumented or "status unknown" foreign-born individual, or if they were categorized as a "Shock Release Hearing Type," or a Harlem Reentry Court case.

Treatment Group participants had 24 hours after their official release to report to their assigned Parole Office, at which time the parolee was assigned to one of the designated PFS Parole

[8] See Brennan, T. & W. Dieterich, (2007), ***New York State Division of Parole COMPAS Reentry Pilot Study Psychometric*** Report: Northpointe Institute for Public Management, Traverse City, Michigan.

[9] Study participants were identified and randomized based on information available while they were still in prison. Per communication from the Project team, individuals who were not subsequently released from prison were not excluded from the overall study sample, i.e., their outcomes were set to zero

[10] The initial design was limited to the four listed NYC parole offices. Because the projected rate of release was not occurring as planned, additional parole offices were added in 2014: Brooklyn II and IV in April, and Brooklyn I and Bronx III and IV in December.

Officers and given a specific date for their first post-release meeting. At least two Parole Officers from each Bureau in NYC and Rochester Metro were designated as PFS Parole Officers.[11]

The parolee's initial meeting typically included both the assigned Parole Officer and a member of the CEO staff.[12] At this meeting, the PFS Parole Officer informed the Treatment Group members that participation in CEO services, although voluntary, was a "Special Condition" of their release from prison (see Appendix A, pages 57-59). The Parole Officers, and the CEO staff member, reportedly encouraged participation in the available services.

Parolees assigned to the Control Group were also required to report to their assigned Parole Office after their release from prison, but these individuals were handled through the existing standard set of operating procedures in the respective project parole offices. They were assigned to any of the active Parole Officers who were not selected to serve as designated PFS Parole Officers. No attempt was made to exclude them from enrolling in CEO services.

Random Assignment to Treatment Conditions

Every other week in New York City, or monthly in Rochester, DOCCS Research identified individuals to be released to community supervision, who met the eligibility criteria listed above, and who were scheduled to be released to one of the selected Parole Bureaus. In addition, eligible individuals were also expected to have completed their community preparation process, i.e., having the necessary release documents and a documented post-release residence.

The identified individuals formed the "Randomization List" that was used to assign study participants. The individual records were matched with other administrative records to append a set of participant characteristics (see Appendix A, page 56), along with the individual's Social Security Number (SSN, if available) used for matching to employment records, and the criminal records ID number ("NYSID") used for matching to incarceration records to measure recidivism outcomes.

The design of the New York PFS project intentionally established unequal probabilities of assignment to either the Treatment or Control Group. First, about 85 percent of the sample was to consist of parolees sent to community supervision in New York City, with the remaining 15 percent consisting of parolees released to Rochester. Within the two locations, assignment to the two treatment conditions groups was based on different randomization probabilities:

- In New York City, DOCCS Research randomly assigned the eligible individuals from the most recent Randomization List using the SPSS randomization command that generates random draws from the uniform distribution (extending from 0 to 1). Individuals with the lowest number were placed into the Control Group until at least 28 percent of the Randomization List individuals were assigned. Once the 28 percent threshold had been reached, the remaining

[11]The designated PFS Parole Officers were responsible for managing the cases of Treatment Group members. Designating and training only about two parole officers per office was intended to increase the consistency and efficiency of operations, and minimize the risk of raising general awareness about CEO and thereby possibly increasing the referral of Control Group members to the intervention. However, because Parole Officers were not assigned randomly, it is not possible to identify the impact of the CEO program apart from the effect of being served by these specific Parole Officers.

[12] CEO staff also conducted "in-reach" at the Queensboro Correctional Facility at least once every two weeks to meet with Treatment Group cases on either on individual or group basis. However, only CEO activities that occurred after release from prison were counted as referrals to, or enrollments in, the intervention.

Randomization List members were placed into the Treatment Group prior to release (72 percent of the total assigned).

- In Rochester, recently identified eligible individuals with the highest random numbers were initially placed in the Treatment Group until at least 31 percent of Randomization List members were assigned. The remaining 69 percent of Randomization List members were placed in the Control Group. In June 2014, the ratios were revised because the actual rate of release did not match the original plans. At that point, 41 percent were assigned to the Treatment Group and 59 percent to the Control Group.

To help achieve balance between the Treatment and Control Groups, random assignment was done within "blocks" by dividing eligible individuals into four groups -- those with COMPAS Supervision Level 1 and those with COMPAS Supervision Level 2 separately for NYC and Rochester Metro. As described in Appendix A, a statistical weighing method was used to account for the different assignment probabilities in the calculation of Population Outcomes.

After each randomization, the randomized list of Treatment Group members was provided to CEO, the Intermediary (Social Finance, Inc.), and DOCCS staff sorted by parole office. The lists included each individual's name, NYSID, Department Identification Number (DIN), assigned parole Bureau and officer, COMPAS Supervision Level, date of birth, and predicted release date. The information in the list was compiled into the Master Data File. Data on the Control Group members was maintained by DOCCS Research but was **not** released to CEO or to the participating Parole Offices. CEO and the Parole Offices were, therefore, blind to the identity of the Control Group members.

After each scheduled randomization, DOCCS Research also shared a report with the NYS Department of Labor (NYSDOL), CEO, Social Finance, the Harvard Kennedy School Government Performance Lab (GPL), and the Validator showing the number of individuals who were randomized to each group, their COMPAS Supervision Level, SSN availability, age, assigned Bureau, and information on other characteristics. (See Appendix B for examples of these reports.)

Data Collection Procedures

All data for this project were collected either by the participating New York State agencies, or the service provider CEO, from existing administrative data systems. No special data collection was employed for this project.

Individual Study Participant Data: DOCCS Research routinely matched the project "Master Data File" of all randomized parolees with the DOCCS "Releases File" to determine if a sample member was ultimately released from prison after random assignment, and if so updated the file with the "Event Date," i.e., the individual's date of release to community supervision.

After each randomization, DOCCS Research provided the list of sample participants to DOCCS Information Technology Services which then noted whether an individual was assigned to the Treatment Group in the relevant field of the DOCCS Case Management System ("CMS"); as noted above, this field was visible only to the designated PFS Parole Officers.

Study Outcomes Data: DOCCS Research was responsible for the collection of information on each individual's post-randomization recidivism outcomes (Prison and Jail Bed Days, and where applicable, remaining prison sentences). Individuals could serve time under two

scenarios: (a) while awaiting the completion of a Parole Violation Process while under community supervision, or (b) for a New Court Commitment following discharge from community supervision. Details on the measurement of each of these outcomes, using administrative records, is described in Schedule 1 of Appendix A.

Using the same Master Data File of randomized individuals, NYSDOL separately generated individual-level employment data by matching SSNs provided by DOCCS with the New York State Unemployment Insurance Wage Records. These data were also added to the Master Data File.

Service Participation: CEO tracked participation in project services for all members of the Treatment Group, and reported the information each month directly to DOCCS Research.

Because, as noted above, CEO was unable to identify individual Control Group members, DOCCS staff used their individual NYSIDs to match against service provision information provided separately by CEO to the State under their existing provider contract. This allowed a comparison of the service participation of the Treatment and Control Groups (see the subsequent discussion in Section IV).

Internal Data Audits: Prior to calculating Phase I Population Outcomes, DOCCS Research conducted a manual audit of the Master File with assistance from NYSDOL. This involved an audit of ten randomly selected records during each calendar quarter starting with the first calendar quarter after the start of the project, and an additional audit of 50 randomly selected individual records at the end of Phase I.

Data Analysis

The estimation of the Population Outcomes listed above was conducted at the end of Phase I, and will be separately conducted on a new study sample at the end of Phase II. The following discussion applies only to Phase I. However, no changes to any aspects of the project are expected at this time for Phase II.

General Approach to Estimating Population Outcomes: For each of the designated outcomes, two separate Population Outcome estimates were produced at the end of Phase I:

- Intent to Treat (ITT) average treatment effects (ATE) were calculated using observed outcomes for all of the randomized study participants, regardless of whether an individual received CEO services or not. Weights were incorporated into these analyses to account for varying randomization ratios across time and sites (see Appendix A, Schedule 1, pp 11-16).[13]

  The ITT estimate was calculated as the average weighted individual outcomes for the Treatment Group minus the average weighted individual outcomes for the Control Group:

  $$\text{ITT} = Y^T - Y^C,$$

  where Y^T is the weighted mean of the individual outcomes (employment and recidivism) for the all Phase I parolees randomized to the Treatment Group, and Y^C is the weighted mean of the individual outcomes for the all Phase I parolees randomized to the Control Group.

[13] See Section III for a discussion of the weighting methods used.

- Instrumental Variable (IV) analyses (referred to as Complier Average Treatment Effects) were also conducted to calculate the effect of the intervention on the "treated," i.e., on those randomized individuals who complied with their assigned treatment condition. This separate calculation is intended to account for the fact that some of the Treatment Group members did not participate, as expected, in the CEO services, while some members of the Control Group managed to participate in CEO (see the discussion in Section III below).

  The IV estimate is calculated by dividing each of the ITT outcome estimates by the difference between the weighted percentage of the Treatment Group members that attended the initial CEO orientation meeting (enrollment), "$\hat{p}_T$", and the weighted percentage of the Control Group members that also attended the initial orientation meeting and thereby enrolled with the service provider, "$\hat{p}_C$" (see Appendix A, Schedule 1, pp 16 -19).

It is the IV estimate that will be used to calculate payments as discussed in Appendix A, Schedule I, Article VIII. NOTE: these payment calculations were specifically designated to fall outside the purview of the Independent Validator.

Section II: Role of the Independent Validator

Scope of Work

As the PFS Independent Validator, I have been directly contracted by New York State to verify that the project's Population Outcomes have been calculated in accordance with Schedule 1, Section VII of the PFS project plan (see Appendix A). This determination will then trigger the calculation of Final Outcomes and payments as discussed in Schedule 1, Section VIII.

My role has involved four major activities:

1. Establishing the validation methodology and determination process to be used. This plan was developed, reviewed, and approved by the State and the project team at the start of the project.

2. Conducting an initial assessment of the *planned* project procedures. This early investigation was intended to verify that the study plans matched the procedures described in Schedule 1, and if properly implemented, were capable of producing a rigorous assessment of the impact of the intervention on the selected participant outcomes.

3. Monitoring the *actual implementation* of the project throughout the Phase I period of sample recruitment, random assignment, data collection, and service delivery.

4. At the end of Phase I, and again at the end of Phase II, validating the calculation of estimated Population Outcomes, and providing documentation of the verification process and results.

The plan for accomplishing these objectives is described in Appendix C, *The Validation Methodology Plan.* The information that was made available to me, and how it was used, are discussed below: (1) Section III discusses the assessment of the project as planned (item #2 above); (2) Section IV discusses monitoring of the ongoing project implementation (item #3); and, (3) Section V discusses the final Phase I outcomes and validation determination (item #4).

Limitations

First, my focus was primarily on whether the evaluation was carried out in accordance with the approved project plan. Considerations of alternative operational procedures were not within my purview.

Second, the PFS project, and the role of the Independent Validator, excluded any systematic study of the extent to which the intervention was implemented with fidelity. The only information I was able to collect was related to participants' compliance with treatment assignment, rates of CEO service receipt, and reports by Parole Officers on their adherence to the required parolee referral process.

Third, the scope of work and budget allocated to the Independent Validator did not include the time and resources needed to do an independent replication and analysis of the estimated outcomes using individual-level data. In addition, my role excluded any direct auditing of the project data files, or access to computer programs used for data processing and analysis.

Finally, it is important to note that I am not responsible for any flaws in the information provided to me. Nor am I responsible for reaching an incorrect conclusion because of information that was outside the constraints of my work, and therefore unavailable for my review.

Section III: Review of the Project as Planned

Good scientific practice requires specifying the key elements of how an evaluation, and especially the approach to data analysis, will be conducted in advance of observing the data. This helps guard against "fishing" for favorable results, and thus increases the credibility of the study findings.

Consequently, the validation process began with a careful review of the planned study design and implementation procedures to assess whether they conformed with the original proposal and, if properly executed, could yield reliable and valid estimates of the intervention's effect. At the conclusion of this review, the plan was considered to be "registered," and any subsequent changes were to be examined in terms of the possible implications for the validity of the project's findings.

The following describes the activities conducted to assess the quality of the project plan, and the results of these early activities.

Meetings and Site Visits

At the start of the project in November 2013, I conducted interviews with key project staff, and reviewed available project-related documents, primarily through site visits to Albany and the New York City parole offices selected for inclusion in the PFS project. A detailed discussion of these meetings and document reviews are provided in Appendix D and Appendix E. (Note: Appendix E includes embedded comments from the PFS project team on issues I raised in this report.)

The conclusion I reached at that time was that the project was well designed and in conformance with Schedule 1. If properly implemented, it was my opinion that the project had the capacity to produce valid and reliable estimates of the specified outcomes for the Phase I sample. Two potential concerns were, however, noted in my initial reports:

- The possibility of a low "treatment contrast," i.e., the difference in post-release receipt of services may not be sufficiently large to produce an observable effect on study outcomes given the size of the Phase I sample.

- The possibility of high sample attrition for employment outcomes because of missing or invalid Social Security Numbers (SSN's).

As noted in the embedded comments in Appendix E, both issues were discussed by the PFS team and determined to not warrant any changes to the project's implementation. How these issues have actually played out is discussed below in Section IV.

Initial Review of the Random Assignment Process

To conduct an early check on the integrity of the random assignment process, I requested and received a de-identified data file from DOCCS Research that included baseline variables for all of the parolees who were randomized through May 27, 2014. To test for baseline equivalence, I estimated linear Ordinary Least Squares regression (for continuous variables) or logistic regression (for dichotomous variables) models in which each baseline variable was regressed on the treatment indicator (whether an individual was assigned to the Treatment or Control Group). Because of the varying randomization probabilities, the assignment weight was incorporated into

these analyses. This approach, as appropriate, coincides with the planned estimation of Population Outcomes.

According to these early results, none of the baseline variables showed a statistically significant difference between the two randomization groups. However, the magnitude of the Treatment-Control Group difference on "whites" was large (0.39 SD) so this was noted as something to keep an eye on moving forward. (These checks were also done unweighted and produced the same basic conclusion.)

In conducting these analyses, however, I found myself uncertain about the calculation of the weights used to account for the different random assignment probabilities in NYC and Rochester. In particular, it was the setting of all Treatment Group cases to a constant weight of 1 that was (for me) an atypical method. A more common approach is to compute weights for the Treatment and Control Groups separately by site and cohort (using the planned blocking by supervision level) as the inverse of their selection probabilities, and then re-scaling the weights across randomization periods to account for the different N's in each assignment cohort.

Consequently, I requested some additional information to help me better understand the weighting method, and the actual calculations used to compute the weights. The results of this investigation are discussed below.

Subsequent Review of the Planned Weighting Procedure

As documented in Appendix F, I engaged in a lengthy discussion with the key project members from GPL during the summer of 2014 about the chosen sample weighting methodology, and how it differed from the more typical inverse probability weighting approach. The Harvard team made strong technical arguments in favor of the proposed plan, but I requested an analysis of the two alternatives to determine the implications (if any) of using one method over the other. Testing the sensitivity of the results to the choice of weighting methods was, for me, a necessary step in the validation process.

As shown in the appendix, an analysis was conducted using the actual study data through July 8, 2014 to compare the relative weights assigned to each stratum under the two alternative weighting methods. The results of these analyses demonstrated that the two procedures yielded essentially the same results. On the basis of this finding, along with the GPL's technical literature citations supporting the PFS weighting approach, the team agreed to maintain the use of the planned weighting methodology. However, I requested a change to the weights to include the use of the individual's COMPAS level. This request was made to align the weights with the use of the COMPAS level as random assignment blocks. This change was agreed to and subsequently incorporated into the weighting methodology.

Section IV: Review of the Project as Implemented

Once the project was underway, I monitored and reviewed the ongoing implementation of the study to assess the extent to which the project was implemented with fidelity to the planned design. The results of this activity, as discussed in the *Validation Methodology Plan*, is an important factor in my determination of the validity of the final Phase I Population Outcome estimates.

Was there Continued Adherence to the Planned Parolee Referral Process?

To answer this question, I conducted another site visit in November 2015, this time to meet with the assigned PFS staff in the Rochester, NY Parole Office. As noted in my Site Visit Report (Appendix G), the discussions with local staff allowed me to conclude that the project continued to be implemented in compliance with the written project plans.

Apart from checking on the ongoing project implementation, the staff provided me with some important insights into the challenges faced by the study parolees that could have implications for the likelihood of finding an effect of the CEO services on parolee outcomes. As noted in my report, these included pre-existing conditions that might make the individuals "not ready" to fully engage in the available CEO program services, issues related to needed coordination with various social services agencies, and potential roadblocks to paid employment opportunities. Such obstacles are not, in my view, evidence of an "implementation failure," but rather a reflection of the reality of the challenges faced by the target population in their efforts to achieve a successful re-entry to the community. These factors are, however, an important context for readers to keep in mind when reviewing the final Phase I outcome results.

Was the Sample Recruitment Procedure Implemented as Planned?

Sample recruitment and random assignment for Phase I occurred between December 9, 2013 and September 29, 2015 (see Appendix H). During this time, a total of 2,357 individuals were selected as eligible for the project and randomized during Phase I – 1,502 were assigned to the Treatment Group, and 855 to the Control Group. Of the total sample, 482 were scheduled for release to Rochester (20.4%), and 1,875 to New York City (79.6%).[14]

Information provided by DOCCS Research on August 30, 2016 confirmed that the software programs used to identify parolees eligible for random assignment, and to subsequently randomize them to treatment conditions, were developed and implemented according to the criteria detailed in Schedule 1 of the PFS project plan. To support this assurance, DOCCS Research provided me with additional information to support the validity of the randomization process, as discussed below.

Were all of the Eligible Parolees Randomized? In September 2016, DOCCS Research shared with me the results of an internal audit conducted at the direction of the NYS Office of the State Comptroller.

[14] Throughout the randomization process, DOCCS Research staff provided the entire PFS team with statistical reports that allowed me, and the rest of the group, to track and monitor the ongoing status and integrity of the assignment process (see Appendix I).

The initial step in the audit was a comparison of the list of all PFS randomized individuals with a list all individuals released during Phase I who met the project eligibility criteria. This comparison reportedly identified 1,010 individuals who were released but who did not appear on the randomization file. A random sample of 225 was selected for review, and from these an initial random sample of 50 cases was selected for investigation by DOCCS staff. This involved the review of each individual's records in DOCCS administrative records to determine the reason why the sampled individuals had not been randomized.

The information provided to me shows that **all** of the 50 cases were correctly excluded from random assignment because they did not, in fact, meet all of the required eligibility criteria. For example, 24% of the cases did not have the required six months of remaining Community Supervision at the time of randomization, and 34% had an actual projected release date too far in the future to be included. Because of this, DOCCS staff concluded that no further investigation was necessary beyond the initial 50 cases.

Were all Randomized Individuals Released to Community Supervision? Also as part of the NYS Office of the State Comptroller's audit, a total of 66 cases were identified that were randomized but never appeared on the file of individuals released from prison. DOCCS staff investigated all of these cases and provided me with the results that show proper documentation for the correct exclusion of these individuals. For example, half of the cases were not actually released to one of the targeted parole offices, and another quarter either did not meet the required COMPAS level, or did not have at least six months of remaining Community Supervision.

Did the Random Assignment Process Yield Equivalent Groups?

Although random assignment is designed to create, on average, groups that are not systematically different on measured as well as on unmeasured characteristics, any single execution of a random process can yield groups with some observed statistical differences. To test this possibility, I asked DOCCS Research to run statistical tests of Treatment-Control Group differences on the available parolee characteristics measured at baseline, i.e., at the time random assignment. The results of these analyses are provided in Appendix J.

As shown in the appendix tables, across the 24 tested variables statistically significant differences ($p \leq 0.05$) were found on three variables: (1) the number of previous parole violations ($p=0.03$), (2) the number of previous convictions for escape ($p=0.045$), and (3) the individual's age at time of his first arrest ($p=0.02$). Normally we would recommend controlling for these variables in the estimation of the population outcomes by including them in a regression model used to estimate the effect of the CEO services. However, the design of this project rejected the use of such statistical controls in the analysis. I do not think this is a fatal flaw because the number of differences is not much greater than what would be expected by random chance (slightly more than one difference), the magnitude of the observed differences are relatively small (for example 47% of the Treatment Group had no prior parole violations vs. 40% of the Control Group), and the addition of the statistical controls would only improve the precision of the estimated treatment effect (not the estimated means) which is not, again by design, being used in the estimation of Population Outcomes.[15]

[15] As reported to me, the Project team rejected the use of statistical models to estimate program effects, preferring instead the use of simple group means.

Was there High Compliance with Treatment Assignment? To What Extent was there a Strong Treatment Contrast?

When we conduct a randomized study like the PFS project, we assume that each individual (in this case, the randomized male parolees) has two "potential" outcomes, (1) what he would have experienced had he been assigned to receive the CEO services, or (2) what he would have experienced had the same person NOT been assigned to the CEO program. The difference in these two possible outcomes is the thing we are interested in – the treatment or program effect.

Obviously, we cannot see any one individual in both conditions – each person can only have a single experience. What we can observe, however, is the average outcome for a sample of individuals who are offered the program (the "Treatment" Group), and the average outcome for a sample of individuals who are not offered the program (the "Control" Group – referred to as the counterfactual). If these two groups are, on average, equivalent in all aspects before the offer of the program, then the observed difference in their average outcomes is an unbiased estimate of the effect of offering the intervention to the eligible parolees (referred to as the intent-to-treat, or ITT, impact estimate).

A randomized study is the best way to produce this analytical contrast – if the samples are large enough, we can be very sure that the two groups are initially the same in all ways (or at least not systematically different). But, for there to be an **observable** effect of the intervention there has to be a sufficiently large "treatment contrast." That is, a treatment effect can only occur if there is a difference in the experiences of the two groups, i.e., between those that are offered the program and those that are not. If a treatment contrast does not exist — that is, if the parolees have the same experience in a program as they would have had had they not been in the program — then we cannot expect there be an observable program effect. Thus a treatment contrast is a necessary condition for a program effect to occur.

As I have written elsewhere,[16] variation in this contrast can arise because of differences across individuals and sites on three main dimensions: (1) the intervention may not have been well designed (i.e., the theory about its possible effectiveness may be incorrect) or properly implemented (the quality, and quantity of services provided may be less than intended); (2) the individuals assigned to the Treatment Group may not "take up" the available services to the degree expected (i.e., because participation is voluntary, and subject to individual choice, actual program participation can be low); or, (3) individuals in the Control Group may find ways to obtain the intervention, or have access to very similar alternative services.

To the extent that these sorts of things occur, the treatment contrast can be reduced. Why is this important? Because for a particular sample size, it becomes more difficult to detect a difference in outcomes – a "treatment effect" – if it exists. In technical language, the power to "detect" a true difference is reduced.

The treatment compliance data observed for Phase I are provided in Appendix K (page 1).[17] Using this information, Exhibit 1 shows the observed Phase I treatment contrasts. What I see

[16] Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods* (NCEE 2014–4017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. Retrieved from http://ies.ed.gov/ncee/edlabs.

[17] Compiled from data provided by DOCCS Research.

is, by typical research standards, a relatively small treatment contrast. Understandably, the project team didn't have the ability to either force the Treatment Group members to participate, especially given the challenges these men face after their release from prison, nor to embargo the Control Group members from finding their way into the CEO program. The implication, however, is that this makes it less likely that the study will observe an impact of the selected study outcomes, if one does, in fact, exist.

The same table also shows the rather stark differences between the two locations, New York City and Rochester. For example, the treatment contrast for attending the initial CEO orientation meeting (the Project's defined indicator of participation) is 39% vs. 61% for New York compared to Rochester, respectively. Similar differences are noted for the other steps in the CEO service process: 32% vs.57% for completion of the Life Skills training program, 24% vs. 44% for participation in transitional employment, and 9% vs. 25% for placement in an unsubsidized job. Clearly, these are very different contexts but something does seem to be happening that should be further investigated to see what can be learned from the Phase I experience.

Exhibit 1: Phase I Treatment Contrasts (Treatment-Control Mean Differences), Overall and by Site (Weighted Data)

Service Participation	Overall Sample	New York City	Rochester
Referred To CEO	43.6%	39.4%	70.1%
Attended Orientation (*PFS participation indicator)*	42.4%	39.3%	61.3%
Completed Life Skills	35.8%	32.4%	57.1%
Worked Transitional Job	26.5%	23.7%	43.8%
Placed in Unsubsidized Employment	11.5%	9.4%	24.8%
Average Days Between Release and Referral	-92.3 Days	-87.6 Days	-143.2 Days
Average Days Between Release and Orientation Attendance	-92.3 Days	-88.2 days	-122.9 Days

Source: Data provided by NY DOCCS Research (see Appendix K).

The second page of Appendix K provides another piece of information about the potential inability of the CEO services to produce impacts on individual employment or recidivism outcomes during the Phase I study period. Overall, 45% of the Treatment Group members, and 41% of the Control Group members, had their parole revoked during the Phase I observation period, returning them to prison.[18] This event certainly limited the ability of the Treatment Group participants to take full advantage of the CEO services available to them, and likely contributed to the small treatment contrast. That is, had the Treatment Group participants experienced a lower rate of parole violations, they may have been able to increase their engagement with CEO, thereby increasing the magnitude of the treatment contrast.

[18] A revocation did not exclude an individual from accumulating bed days before or after the revocation.

Were the Data Measurement and Collection Procedures Implemented as Planned?

There are two relevant questions about the measurement and data collection process: (1) Were the outcome data measured and collected as specified in Schedule I? and, (2) Was there consistent measurement and data collection between the Treatment and Control Groups (i.e., is there evidence of a possible treatment-related confound)?

The data collection procedures are well described in Schedule I, and I have accepted the assurances of the project staff that the programs used to define and select the data were developed and implemented to correctly match the planned process. As an additional check, I requested, and received, written documentation of the QC procedures that were used by DOCCS Research to check the data extracts (see Appendix L). I am comfortable that the data measurement and collection procedures were implemented as required under Schedule 1.

With regard to the issue of consistent measure/collection between the Treatment and Control Groups, I do not see any evidence of possible systematic differences between the two groups which could confound the outcomes estimation. The study uses existing, and well established, State administrative record systems which have high validity and reliability, and have "high stakes" uses well beyond this project. In my view, the likelihood of any systematic data errors or manipulation is extremely low.

Finally, as discussed in Section I the recidivism outcomes (e.g., prison and jail bed days) are calculated as aggregate outcomes for each study group summed across the full Phase I time period. This means that the outcome measure includes individuals who were paroled early in the process and those who were paroled near the end of Phase I. The result is an outcome measure that could have been measured differently across the Treatment and Control Groups if there were any systematic differences in the distributions of the number of observed months.

These data, provided in Appendix J, show no differences between the Treatment and Control Groups, and the analyses conducted by DOCCS Research confirm the fact that there are no statistical differences between the two distributions.

Was the Integrity of Random Assignment Maintained at the Point of Outcome Analysis

Even well-designed RCTs can lose individual study members between the time of initial random assignment, and the point at which study outcomes and treatment effects are calculated. Such study "attrition" can arise for many reasons including the inability to collect data on particular individuals, or the inability to locate particular individuals at the time of data collection.

This sort of sample loss can compromise the comparability of the Treatment and Control Groups at the point of estimating treatment effects. In particular, if overall attrition is high, and systematically different across groups, this can lead to biased estimates of the intervention's effectiveness, particularly if the cause of missingness is related to the outcomes of interest. For example, if parolees least likely to have successful outcomes were systematically more likely to have missing SSNs, thereby causing them to not have available employment data, the estimated employment outcomes could be biased upwards.

Recidivism Outcomes Analysis Sample: According to DOCCS Research, only nine randomly assigned individuals (five Treatment Group and four Control Group members) were excluded from the calculation of recidivism outcomes because they were not released from prison during Phase I. Consequently, for these estimates attrition is an ignorable factor.

Employment Outcomes Analysis Sample: As expected, the situation is different for the calculation of employment outcomes because of the lack of available, or valid, SSNs needed to match the list of randomized individuals with State employment records. Information created by NYDOL, provided in Exhibit 2, shows an overall attrition rate of about 25% (both using weighted and unweighted data), and the following levels of attrition by study group: (1) Treatment Group – 26.8% (weighted and unweighted), and, (2) Control Group – 22.3% unweighted and 24.2% weighted.

Exhibit 2: Phase I Sample Attrition: Employment Population Outcomes, Overall and by Assignment Group (Weighted and Unweighted Data)

Study Status	Treatment Group				Control Group				Total Analysis Sample			
	Unweighted		Weighted		Unweighted		Weighted		Unweighted		Weighted	
	N	%	N	%	N	%	N	%	N	%	N	%
Total Randomized	1502		1502		855		1502		2357		3004	
Not Paroled	5	0.3%	5	0.3%	4	0.5%	6	0.4%	9	0.4%	11	0.4%
Event Date after 9-30-15	71	4.7%	71	4.7%	37	4.3%	67	4.5%	108	4.6%	138	4.6%
Missing/Invalid SSN	326	21.7%	326	21.7%	150	17.5%	290	19.3%	476	20.2%	616	20.5%
Total Missing Data	402	26.8%	402	26.8%	191	22.3%	363	24.2%	593	25.2%	765	25.5%
Total Analysis Sample	1100	73.2%	1100	73.2%	664	77.7%	1139	75.8%	1764	74.8%	2239	74.5%

Source: Data provided by NYDOL.

Are these rates unacceptably "high?" To answer this question, I referred to standards for high quality education research studies[19] developed by the Institute of Education Sciences, US Department of Education for the highly-regarded ***What Works Clearinghouse (WWC)***.[20] The WWC's attrition standard is based on a model for attrition bias that combines the rates of overall and differential attrition and the relationship between attrition and outcomes. To determine reasonable values to use in assessing the extent of potential attrition bias in a particular study, the WWC makes assumptions about the relationship between attrition and outcomes based on findings from several randomized trials in education. Based on the WWC's more "conservative" standards for attrition, the observed PFS employment outcomes attrition rates – 25% overall and a differential rate of 4.5% unweighted and 2.6% weighted – would fall within a range that would not be considered to be worrisome in terms of potential bias.

As a further check, I asked DOCCS Research to compare the Treatment and Control Group individuals included in the sample used to compute employment Population Outcomes on

[19] This is a relevant basis for comparison because most of the CEO services are essentially adult education. This includes the early Life Skills training and the subsequent job skills training.

[20] ***What Works Clearinghouse Procedures and Standards Handbook*** (Version 3.0), US Department of Education, Washington, DC. See https://ies.ed.gov/ncee/wwc.

characteristics measured at the time of random assignment. The purpose of this analysis was to assess the extent to which the observed baseline equivalence measured for the full randomized sample was maintained for the Phase I analysis sample.

As shown in Appendix M, across the 24 tested baseline variables only one shows a statistically significant difference between the two assignment groups – Treatment Group participants were younger than Control Group members at the time of their first arrest. Because a single difference is what one would expect from random chance alone, I concluded that the two analysis groups demonstrated baseline equivalence on measured characteristics at the point of estimating the Population Outcomes. In other words, no attrition bias is observed due to the loss of sample due to invalid or missing SSN's.

Section V: Determination of Phase I Population Outcomes

Reported PFS Findings

Exhibit 3 provides the estimated ITT and IV Population Outcomes (see Appendix N for details). As shown, the Phase I results do not provide evidence of a positive effect of CEO services on the targeted PFS outcomes. As the Project's Independent Validator, I certify that these calculations were conducted in accordance with the agreed upon methodology as described in Schedule I of the PFS Agreement.

In terms of recidivism, on average there was only a 3-day difference in the rates of incarceration while under Community Supervision, and a two percentage point difference in the estimated fraction with remaining sentences. In both cases the results were "worse" for the Treatment Group with participants serving a larger number of bed days over the period of Phase I.[21] The IV estimates – adjusting for observed rates of CEO service enrollment – are, as expected, slightly higher but not qualitatively different from the ITT estimates.

With regard to employment, the Phase I results show a less than one percentage point difference in the rate of employment in the 4th quarter after parole. The IV estimate are again larger with an estimated employment difference of two percentage points.

Exhibit 3: Phase I Recidivism and Employment Population Outcomes, ITT and IV Estimates (Weighted Data)

	ITT Estimates			IV Estimates
Population Outcomes	Treatment Group	Control Group	Difference	Difference
Recidivism Outcomes				
Average Bed Days[22]	232.5	229.5	3.0	7.0
Remaining Sentences	19.4%	17.4%	2.0%	4.7%
Employment Outcomes				
Employed in 4th Quarter	17.6%	16.8%	0.8%	2.0%

Source: Data provided by NY DOCCS Research and NYDOL (Appendix N).

Additional Employment Information: Although not part of the official Population Outcome measures, Appendix N also shows the estimated average 4th quarter wages for the Treatment and Control Groups. As shown, the Control Group had, on average, slightly higher average wages than the Treatment Group, $603 vs. $582, respectively.

Also shown in Appendix N are the distributions of average 4th quarter wages for the Treatment and Control Group, for those with non-zero wages. As indicated by the overall results favoring the Control Group, the more detailed data show that the Treatment Group participants were more likely to have lower quarterly wages (under $1,000), and less likely to have higher wages (over $3,500). Average wages among those who were employed were $3,595 for the

[21] As noted in previous sections, the PFS project opted to not base estimated outcomes on the results of a statistical model to test the statistical significance of any observed group differences.

[22] Combined prison and jail bed days.

Control Group and $3,298 for the Treatment Group, a difference of $297. (Median wages were $2,208 and $2,107 for the Control and Treatment Groups, respectively.)

As noted by NYSDOL in their comments on a draft of this report, these wage comparisons may be affected by the fact that a greater number of Treatment Group participants were in subsidized employment in the 4th quarter after release than were Control Group participants (75 vs. 37, weighted). Subsidized employment is typically paid at the minimum wage.

Comment on the Results

The Phase I results are certainly less than what was hoped for at the start of the PFS project. The intervention, and the evaluation, were both carried out as originally planned, i.e., the execution of the project did not diverge from the plan in any significant way that would invalidate the reported results. So what happened? Although this is outside my responsibility on this project, I would like to offer my thoughts on the possible reasons for the disappointing outcomes.

Was the Targeted Population Too Challenging? I heard from Parole Officers about the serious challenges faced by the participating parolees in their efforts to fully engage with the CEO services and to subsequently find gainful employment. But the design of this project was based on cited research evidence indicating that offenders at the highest risk of recidivism were most likely to benefit from such an intervention.

What's worth checking, however, is whether the individuals involved in the PFS project had characteristics similar to those examined in previous studies, and in particular, if they faced similar contextual factors that may have consequences for their success. For example, staff I met with in Rochester pointed to obstacles such as a misalignment between participants' job skills and the needs of local employers, and access to transportation, as potential inhibiting factors. It would, therefore, be worthwhile to examine information that may be available from Phase I, and certainly from the ongoing implementation of Phase II, to explore the extent to which such issues may have reduced the potential effectiveness of the CEO services.

Was the Theory Underlying the Intervention Incorrect? That is, were the assumptions underlying the design and implementation of the tested intervention incorrect? Based on available information I don't see a reason to question the basic conceptual model. The decision to use CEO as the provider of the service intervention was based on the positive results of an RCT evaluation of CEO in which significant impacts were found for employment outcomes for high-risk recently released offenders.[23] This, of course, doesn't guarantee success with this study's particular sample, context, location, and timing but it does provide some important evidence of the program's ability to produce positive results, at least in the instance previously studied.

Was the Intervention Implemented with Fidelity to the Model? Although the logic model for the intervention may be theoretically supported, the actual delivery of the CEO services may have fallen short of the desired level of implementation fidelity, i.e., the quality or intensity of the actual services may have differed from the planned or desired level of implementation. Because the PFS project did not include a systematic study of implementation there is no way for

[23] *More Than a Job: Final Results from the Evaluation of the Center for Employment Opportunities (CEO) Transitional Employment Program,* OPRE Report 2011-18. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, DHHS, 2012, http://www.mdrc.org/sites/default/files/full_451.pdf.

me to know if this was, in fact, a possible factor in the observed lower than expected impact on participant outcomes. If it is possible to examine this retrospectively, I would suggest conducting an investigation of actual service delivery. And, because Phase II is currently underway, conducting a systematic study of each of the CEO service delivery components would be a useful addition to our deeper understanding of the project's results.

Was There a Large Enough Treatment Contrast? As discussed in Section IV, the observed "treatment contrast" was relatively small, i.e., a 42 percentage point difference in initial CEO enrollment and markedly lower differences for the provided service components (e.g., a 26 percentage point difference for placement in a subsidized job).[24] In my view, the observed group participation differences are unlikely to drive substantively important changes in participant's rates of recidivism or employment. The calculation of estimated intent-to-treat (ITT) effects from an RCT is based on the assumption that only one factor distinguishes the Treatment Group from the Control Group, i.e., one is allowed access to the program services and one is not. To the extent that this difference in the offer of an intervention is reduced – exactly what we have seen in Phase I – the likelihood of observing a treatment effect is diminished.[25]

One set of analyses that I would suggest doing, however, is to examine the data for the Treatment Group participants to see if higher levels of engagement with the CEO services is associated with "better" outcomes. In other words, does a greater "dosage" correlate with lower recidivism and better employment outcomes. Such an analysis could help us understand whether the intervention is sufficiently potent to affect the targeted outcomes when participants are highly engaged with the program. If so, this could indicate if achieving a high level of assignment compliance (i.e., a large treatment contrast) would be more likely to show stronger outcome differences.

Was the Study Adequately "Powered?" At the start of the PFS project, the team calculated the study's statistical power based on the planned size of the study sample, the unequal assignment probabilities (which reduce statistical power from equal allocation), and for different levels of treatment compliance. At the observed level of treatment contrast, the minimum detectable impact (the difference in outcomes between the Treatment and Control Groups) would have had to be about 30 percent of the observed Population Outcome means. For the bed day outcome this would require a minimum observed difference of about 70 bed days, far above the observed level of three bed days. For the employment outcome, this would have required a minimum observed difference of about five percentage points, again much higher than the observed level of under one percentage point. What this suggests, is that the Phase I study sample

[24] As noted by staff of the Harvard GPL, the project team anticipated a 40 percent point difference in take up rates with 60% of the treatment group taking up CEO and 20% of the control group taking up CEO as part of the evaluation plan (as documented in responses to USDOL questions in June 2013). These estimates were based on: 1) the operational challenge in referring all treatment group members to CEO (given other challenges the parole officer may seek to address at the time of release) and converting referrals into enrollees (the anticipated conversion rate was 87%); and, 2) the study design, which allowed control group members to enroll in services.

[25] As noted above, there may also have been a reduced opportunity to fully benefit from the CEO services resulting from the fact that about 40% of both groups (Treatment and Control) had their parole's revoked during the study observation period.

was inadequate to detect a small treatment effect – if a true difference does, in fact, exist –at the observed level of treatment compliance.

One analysis worth doing, is to calculate the realized statistical power of the Phase I sample using actual observed sample characteristics (instead of the assumed values incorporated into the study plan). And, to use these data to estimate minimum detectable effects under different levels of treatment compliance. This would help us to understand how much larger the sample, and CEO participation differences, needed to be to generate substantively important outcome differences.

Suggestions have also been recently advanced to pool the Phase I and Phase II samples to increase the overall study sample and associated statistical power. This is certainly worth doing as it would deepen our understanding of PFS project experience. Before moving ahead with this plan, however, the larger sample assumptions should be incorporated into the estimates of realized statistical power to see how much of difference could be expected from the pooled analysis. In addition, at this point in Phase I we're not seeing a significant difference in the magnitude of the treatment contrast. Consequently, efforts to improve the rate of service take-up by Treatment Group members would also be worthwhile. (Note: I assume that little can be done to lower participation by Control Group members.)

Final Note. None of my comments here should be interpreted as a statement about whether the observed results would have changed under different circumstances. There is simply no way to know this with certainty from the available data.