Improving the Quality of Evaluation Research in Corrections: The Use of Propensity Score Matching

Jason E. Piccone, PhD

Abstract

The effective evaluation of correctional programs is critically important. However, research in corrections rarely allows for the randomization of offenders to conditions of the study. This limitation compromises internal validity, and thus, causal conclusions can rarely be drawn. Increasingly, researchers are employing propensity score matching (PSM) to mitigate this problem. PSM involves matching offenders in an experimental condition with offenders in a control condition on several relevant covariates. Comparisons between these two groups are then more meaningful because we can assume a relative equivalency between groups on the matched characteristics. When utilized appropriately, this procedure can mimic randomization and allow for causal conclusions. Although it is encouraging that this technique is increasingly employed in corrections, it is often misunderstood or misreported. This article provides an overview of the basic methodology and considerations for PSM. In addition, it provides an example of a correctional education program evaluation using PSM strategies.

The inability to maintain experimental control is a fundamental impediment to research in corrections (Sherman et al., 1997). A lack of experimental control is due, primarily, to the fact that random assignment of subjects to conditions of a study is generally not feasible. This limitation makes the determination of causality untenable. For instance, while many correlates to the risk of recidivism are known, without proper experimental control we cannot determine if these variables have a causal connection with recidivism. This methodological issue plagues program evaluations as well. If one wishes to compare the three-year recidivism rate of offenders who enroll in vocational training versus those who do not pursue training, preexisting variability between the groups, as opposed to the program's efficacy, could explain post-treatment differences. However, methodologies can be utilized to maximize the strength of empirical research, given the inherent limitations that render

randomization unfeasible. This paper addresses the use of PSM, as it can mimic true randomization. This is a basic overview, as the nuances of PSM are expansive. The following is meant to motivate a more systematic reporting of such research and to provide a starting point for those interested in utilizing this procedure.

Experimental vs. Quasi-experimental Research

The degree of experimental control dictates the strength of the conclusions that can be drawn from a study. Ultimately, researchers must be able to safely conclude that the independent variable is the only factor that varies across conditions of the study. A study that utilizes randomization is the most effective methodology to accomplish this task. In a randomized experiment, subjects are randomly assigned to conditions of the study, providing assurance that no pre-existing differences exist between the groups. A quasi-experiment, on the other hand, is when random assignment to conditions does not occur (Shadish, Cook, & Campbell, 2002). Randomized experiments allow for the strongest claims of causality.

The ability to make causal conclusions is critical—it is the difference between saying that program X reduces recidivism vs. program X is related to a reduction in recidivism. Gordon and Weldon (2003) investigated the recidivism rates of offenders who participated in educational programming. The authors found substantial differences in recidivism based on whether participants engaged in vocational training (8.75%), GED and vocational training (6.71%), and no educational participation (26%). Although these data suggest an important role of vocational training, one cannot state that educational training had an effect. Because participants were not randomly assigned to a condition of educational participation, other variables could account for the differences in recidivism. For instance, offenders who opt for educational pursuits likely differ on a variety of qualities from offenders who opt out of such pursuits. One might expect that if true experimental methods were used, the "effect" of education would be less substantial because the groups would no longer be confounded by pre-existing conditions (such as motivation) that inflate the positive outcomes for the treatment groups.

A true experimental study would be ideal to evaluate the efficacy of various programs for criminal offenders. Randomly assigning offenders to control and treatment conditions would assure that the groups were equivalent before introduction of the independent variable. Of course, randomly assigning offenders to such conditions or a control group would violate practical and ethical constraints. In this case, the best available alternative is the use of PSM to statistically balance the groups of comparison.

Improving Methodologies in Corrections Research

It is not uncommon for research in corrections to claim causal effects of a program even when quasi-experimental methods are used. For instance, Lockwood, Nally, Ho, and Knutson (2012) make the causal claim clear in the article title, "The Effect of Correctional Education on Postrelease Employment and Recidivism" even though non-experimental methods are employed. Causal claims are often drawn despite an inappropriate level of experimental or statistical control (e.g., Gordon, & Weldon, 2003; Zgoba, Haugebrook, & Jenkins, 2008).

The well-cited Three State Recidivism Study (Steurer, Smith, & Tracy, 2001) employed a quasi-experimental approach. One of the findings of this study was that correctional education participants had lower rates of recidivism than non-participants across the states of Maryland, Minnesota, and Ohio. The authors are rightfully careful to acknowledge that the lack of randomization presents a serious challenge to causal conclusions. They attempted to counter this by using a release cohort design. In such a design, subjects (both from the experimental and control group) are released from incarceration during the same period of time. Steurer et al. (2001) state that such a "design takes advantage of the natural flow of cases through the criminal justice process with an assumption that the treatment group and the comparison group are similar on key variables known to impact recidivism and employment" (p. 12). Such an assumption is not easily defended. Participants could still differ from non-participants on important variables (such as criminal history and achievement motivation) that could have an effect on the outcomes.

PSM is used in corrections research with increasing frequency. For example, Kim and Clark (2013) assessed the effect of prison-based college education on recidivism; Bales and Piquero (2012) assessed the effect of incarceration vs. a prison diversion program on recidivism; Jensen and Kane (2012) tested the effect of in-prison therapeutic communities on recidivism; and there are many others (e.g., Butler, Goodman-Delahunty, & Lulham, 2012; Dirkzwager, Nieuwbeerta, & Blokland, 2012; Johnson & Kurlychek, 2012). The improvement of methodological models is painting a slightly different picture of corrections research. Previously, program evaluations that were assessed through quasi or non-experimental means may have appeared more effective due to uncontrolled confounding variables—such as a self-selection bias. Kim and Clark (2013) point out that while they found a significant effect of a prison-based college education on recidivism when using PSM, they found a stronger effect when the participant group was compared to the control group without the use of matching.

With the increasing prevalence of PSM methodologies in corrections research, it is important that it is conducted and reported correctly. The following provides (a) an overview of PSM, (b) basic components and considerations of PSM, (c) and a brief demonstration.

Treatment-Control Group Matching

Matching designs are generally used when the researcher lacks the ability to randomly assign participants to conditions of the study. They involve selecting a sample of control cases whose distribution of scores on one or more covariates is similar to a treatment group. For example, Stuart and Green (2008) estimated the effect of heavy marijuana use on later outcomes (e.g., mid-adulthood socioeconomic development). Subjects were matched on variables that are known to relate to the dependent measures, such as maternal history of drug use and family income. Thus, the comparison of heavy and non-users of marijuana is no longer confounded by these variables.

Other confounding variables likely still exist (Stuart and Green included others such as teacher ratings of the child's behavior), and further variables can be included to match on. When matching on a single covariate, it is a simple procedure to calculate the relative difference between every potential treatment-control match pair. As covariates are added, this process becomes increasingly complex as several difference scores would need to be compared in determining which cases to match. This problem is handled through the use of propensity scores.

Propensity Score Matching

A propensity score is the probability that each case belongs to the treatment or control condition based on the included covariates: it is "the propensity towards exposure to treatment 1 given the observed covariates x" (Rosenbaum & Rubin, 1983, p. 43). This probability is generally calculated through a logistic regression analysis where the covariates predict membership in the treatment or control group. Thus, groups of treatment and control cases with similar propensity scores are comparable, even if matched pairs differ on specific covariate values. Matching on propensity scores can produce a strong balance between treatment and control groups on the included covariates (Rosenbaum & Rubin, 1985).

While each treatment and control case would ideally possess identical values on each covariate, in reality this is not realistic. One can, however, use exact matching on one (or a few) covariates of critical importance. Another strategy involves calipers where minimum requirements are specified for each match. For instance, Ming and Rosenbaum (2001) set calipers such that the age between treatment and control cases in each matched set must be within five years. Calipers could also be set, for example, by requiring each matched set to be within .20 standard deviations of the logic of the propensity score (Austin, 2010).

The strength of propensity score designs falls on the ability to include meaningful covariates and to find strong matches. Variables that are expected to differ between the treatment and control groups and that are related to the outcome should be included. This selection begins with a review of the relevant scientific literature. For instance, Stuart and Green (2008) understood from past research that maternal drug use is an important variable predicting later socioeconomic development, thus including it as a covariate was crucial in establishing a valid matching model. Several propensity score models should be tested, where different combinations of covariates are included, to determine which provides the strongest overall balance. Interaction terms between covariates could also be used to include higher-order effects (Austin, 2009; 2011a). Although including even weakly associated covariates in the propensity model can reduce bias, this decision should be tempered by sample size considerations. For instance, a large treatment to control case ratio allows for the inclusion of more covariates without a substantial loss of balance between the groups (Ho, Imai, King, & Stuart, 2007).

Matching Procedures

Once propensity scores are calculated, one must determine how to match control and treatment subjects. Traditional "nearest match" or "greedy" strategies link each treatment case, in turn, to a control case. Once a control case is used, it is removed from the list of potential matches and this process is repeated until each treatment case is matched. Thus, each match is made without consideration for the fit of matches in subsequent iterations. For instance, let us say that we have experimental cases A, B, C, and D and control cases W, X, Y, and Z. Case W may be the closest match to case A; however, matching these cases together may not allow for the best fit for the overall model. This might occur if case W is also the closest match for case B. Because B cannot be matched with W, we have to hope that the next best match doesn't generate a greater discrepancy in fit than if B and W were matched with A matched to the next nearest fit.

Optimal matching is a more sophisticated strategy that considers all possible combinations of match pairings (Hansen, 2004). Consider the following 2 × 2 propensity score distance matrix:

		Treated	Cases
		t1	t2
Control	c1	0	1
Cases	c2	1	10

A greedy match would link t1 with c1 for a distance of 0. The only remaining option would then be to link t2 with c2 for a distance of 10 so that the overall match of this model would have a distance of 10. The optimal match would consider the poor match of t2–c2 and would thus match t1 with c2 for a distance of 1 and t2 with c1 for a distance of 1 for an overall model distance of 2. In some cases, such as when the number of control cases is great relative to treatment cases, the greedy algorithm works well. The optimal algorithm should, however, provide an equal or superior model match (Rosenbaum, 1989).

How Many Controls to Include in Each Match Set

Matches need not be on a one-to-one basis. If control cases greatly outnumber treatment cases, one could apply one or more control cases to each treatment case. While increasing the proportion of control conditions used, it is likely that balance will suffer. In other words, in a 1:1 match, the pair is likely to be more similar than the average of several controls compared to a single treatment. Nevertheless, it is also possible that the imbalance created is overpowered by the benefit of increasing the number of cases included for analysis, as statistical precision is enhanced (Hansen, 2004). Haviland, Nagin, and Rosenbaum (2007) demonstrate the reduction in the error variance term as additional controls are included in each matched set. Ultimately, the number of controls to include depends on the features of the particular data set.

One-to-One/Pair Matching

Each treatment case is matched with one control case. This method is the least flexible, and in cases where a dataset contains many control cases relative to treatment, you will end up excluding a great deal of data. On the other hand, if the data lacks a large pool of control cases, this may be the best option as forcing multiple controls may result in undesirable matches.

Ratio Matching

Each treatment case is matched with one or more controls. One can establish the parameters of this ratio as fixed (e.g., 1:3) or as variable (e.g., 1:1–4).

The ratio of treatment to control subjects in each matched set in variable ratio matching is dependent on the parameters established in the model. For example, strict calipers will result in smaller matched set ratios. Ming and Rosenbaum (2000) found that matching with variable ratios can significantly reduce covariate bias over fixed match models.

Optimal Full Matching

This method combines either a single treatment case with one or more controls, or a single control with one or more treatment cases. This method is flexible. For instance, it handles data sets in which few controls exist in relation to treatment cases. It also takes advantage of the strengths in the data, matching more cases when more cases exist in close proximity and matching fewer cases when the opposite is true (Hansen & Klopfer, 2006).

Matching with Replacement

The same control case can be matched to more than one treatment case. This is desirable when doing so significantly improves the balance of the matching design (Abadie & Imbens, 2006), such as when there are relatively few controls to draw on. However, statistical analysis becomes more complex in order to account for the lack of independence of the matched controls (Stuart, 2010).

Analysis of Covariate Balance

Did the matching procedures sufficiently balance the treatment and control groups on the pre-treatment (covariate) values? While no standard threshold is established, every effort should be made to maximize balance. Imai, King, and Stuart (2008) discuss four reasons that t-tests and other significance tests of mean differences should not be used to assess balance. One of these reasons is that significance tests are easily affected by sample size. A model with a larger sample size will indicate less balance because a larger sample size can more easily indicate statistically significant differences between the control and treatment groups. Recommendations to evaluate balance include a comparison of means, inspection of box plots or quantile–quantile (*Q-Q*) plots, and examining the standardized difference for each covariate before and after matching (Guo & Fraser, 2010; Imai et al. 2008; Rosenbaum & Rubin, 1984). In addition, some authors (e.g., Austin, 2011a; Hill, 2008) suggest the comparison of covariate variance between the matched control and treatment groups. Hill (2008) argues that matched samples could appear balanced in a comparison of means, while a discrepancy in variances can indicate a higher-order bias. In addition, remaining (post-match) differences between the treatment and control groups on the included covariates can be adjusted through regression analysis (Austin 2011b; Hill, 2008; Imbens, 2004; Rubin, 2001; Stuart 2010).

Analysis of the Treatment Outcome

Ultimately, we hope to determine whether the program of interest has an effect on some outcome. As in non-matched designs, a significance test and an estimate for the treatment effect is typically desired. The choice of significance test depends on the nature of the matching design used. For example, a design that employs a 1:1 treatment to control case pairs requires a different test from a design that employs ratio matching. Some authors (e.g., Austin, 2011b) suggest that the analysis should account for the matched nature of the data. In the case where we wish to compare a binary outcome (e.g., recidivism or no-recidivism) between a treatment and a control group, a chi^2 or Fisher's exact test would be appropriate. Once the data is matched, however, McNemar's test would be appropriate because it accounts for the matched pairs. Others (see Stuart, 2010, for example) contend that if covariate balance has been conditioned through a regression adjustment, then the data could be analyzed as if it were unmatched. For example, Stuart (2010) argues that such analysis is appropriate because PSM does not guarantee that each matched set has equivalent scores on the set of covariates. Rather, it ensures that groups of treated and control cases with similar propensity scores have similar covariate distributions. See Guo and Fraser (2010) and Rosenbaum (2002), for further discussion on post-match analyses.

Sensitivity Analysis

Even if the treatment and control groups are balanced perfectly on the included covariates, we still do not know if relevant, non-measured covariates are balanced. We can, however, determine how sensitive the model is to the presence of a hypothetical variable. "A sensitivity analysis is a specific statement about the *magnitude* of hidden bias that would need to be present to explain the associations actually observed" (Rosenbaum, 2005, p. 7). A sensitivity analysis is based on the parameter Γ (Gamma), which indicates the extent of departure from random assignment. In randomized studies, or observational studies with no hidden bias, Γ should equal 1 (no departure from random). If Γ =2, then a case in a matched pair with similar scores on the covariates could be twice as likely to receive treatment due to an unobserved covariate (Rosenbaum, 2002).

As Γ deviates from 1, we have less confidence in the model as it becomes more sensitive to a hypothetical unmeasured covariate. On the other hand, if Γ needed to be of an enormous magnitude to significantly alter the model, then we would have greater confidence. Such a test is necessary to demonstrate the robustness of the model and thus confidence that randomization is adequately mimicked. While evaluations employing PSM methodologies should include a sensitivity analysis, very few studies applying PSM techniques in corrections do.

Demonstration of Propensity Score Matching

In addition to providing a basic understanding, it is our hope to show the relative accessibility of PSM. The following demonstration utilizes a generated data set meant to mimic a typical corrections program evaluation. The sample consists of 300 male offenders, 60 of whom completed a vocational training program. The treatment subjects represent a group of individuals who self-selected to participate in this program. It is an obvious concern in this study, as in most observational studies, that the treatment group differs in systematic ways from the control group. We are attempting to determine the efficacy of this training in reducing three-year post-release recidivism rates. In this example, the covariates consist of ethnicity (Black and Caucasian), age (18–65), number of prior arrests (0–7), sentence length (2–60 months), last grade completed (6–14), reading level (5–14), and math level (5–14). While interaction terms among covariates can be included, none of the available interactions contribute meaningfully in this case and are thus omitted. In addition, a dichotomous recidivism score (0, 1) was generated. Table 1 presents the unmatched participant characteristics. An examination of the standardized difference scores indicates substantial discrepancies between the two groups on several of the covariates. In particular, the treatment group had significantly fewer prior arrests ($p < .05$), a significantly higher last grade completed ($p < .05$), and a significantly higher reading level ($p < .01$) than the control group. It is apparent that the two groups are not equivalent, as is the concern in many program evaluations in corrections that employ observational designs.

It is not uncommon in corrections research to see non-randomized data and unmatched data analyzed with brave conclusions drawn. We argue that such conclusions are largely meaningless as significant confounding variables cannot be ruled out. In analyzing this unmatched data, we find that Fisher's exact (one-tailed) test is statistically significant ($p = .033$), indicating that offenders who complete the vocational training are less likely to recidivate within three years 10/60 (16.7%) than offenders not completing such training 70/240 (29.2%).

Table 1. Unmatched Participant Characteristics

Variable	Treatment ($n = 60$)	Control ($n = 240$)	Standardized Difference
Ethnicity	45% Black	55% Black	0.1993
Age	38.20 (11.74)	38.04 (13.01)	0.0135
Prior Arrests	1.58 (1.31)	2.06 (1.46)	–0.3670*
Sentence Length	28.10 (16.71)	31.23 (17.36)	–0.1875
Last Grade Completed	10.45 (2.36)	9.66 (2.48)	0.3336*
Reading Level	10.07 (2.15)	8.71 (2.31)	0.6306**
Math Level	8.88 (2.34)	8.33 (2.08)	0.2389

Continuous variables are presented with means and standard deviations.
* $p < .05$
** $p < .01$

The relative risk ratio = .5714 (95% confidence interval .314 to 1.040), which indicates that exposure to treatment is associated with almost half the likelihood of recidivism compared to the control condition.

Comparison of Matching Procedures

PSM can be computed using STATA, SAS, and R. This example is completed using the R platform. R was chosen because it is available for free, it is highly flexible, and it is relatively easy to work with. A data file was generated and imported into R. For the sake of simplicity, we will compare only the balance of a nearest neighbor (greedy) match, a full match, and an optimal match. Each model is computed using Ho, Imai, King, and Stuart's (2011) MatchIt package in R. Propensity scores were calculated by regressing treatment status (treatment or control) on the seven covariates listed in Table 1. The matching procedures utilize calipers of .2 of the standard deviation of the logit of the propensity score, per Austin (2011a) and Lunt (2014).

The greedy match procedure resulted in 56 matched pairs. Four treatment cases were excluded because they did not meet the caliper restriction. All absolute values of the standardized difference scores decreased, with the exception of age, in comparison to the pre-matched samples (see Table 2). The full match successfully matched all 60 treatment and all 240 control cases. Again, all standardized difference values moved toward zero, with the exception of age (Table 3). Finally, the optimal match (Table 4) also indicates improvements

Table 2. Participant Characteristics for Greedy Match

Variable	Treatment (*n* = 56)	Control (*n* = 56)	Standardized Difference	Variance Ratio
Ethnicity	51.79% Black	48.21% Black	0.0712	1.000
Age	38.54 (11.88)	39.46 (12.78)	–0.0791	1.157
Prior Arrests	1.70 (1.28)	1.79 (1.32)	–0.0684	1.070
Sentence Length	28.27 (16.84)	28.79 (17.04)	–0.0310	1.024
Last Grade Completed	10.39 (2.43)	10.36 (2.41)	0.0151	0.985
Reading Level	9.86 (2.02)	9.98 (1.93)	–0.0580	0.907
Math Level	8.89 (2.40)	8.96 (2.22)	–0.0306	0.858

Continuous variables are presented with mean and standard deviation.

Table 3. Participant Characteristics for Full Match

Variable	Treatment (*n* = 60)	Control (*n* = 240)	Standardized Difference	Variance Ratio
Ethnicity	45.00% Black	49.79% Black	0.0956	0.988
Age	38.20 (11.74)	38.54 (13.01)	–0.0286	1.227
Prior Arrests	1.58 (1.31)	1.71 (1.46)	–0.0950	1.252
Sentence Length	28.10 (16.71)	28.94 (17.36)	–0.0504	1.079
Last Grade Completed	10.45 (2.36)	10.72 (2.48)	–0.1130	1.106
Reading Level	10.07 (2.15)	10.12 (2.32)	–0.0266	1.156
Math Level	8.88 (2.34)	8.96 (2.08)	–0.0313	0.793

Continuous variables are presented with mean and standard deviation.

Table 4. Participant Characteristics for Optimal Match

Variable	Treatment (*n* = 60)	Control (*n* = 60)	Standardized Difference	Variance Ratio
Ethnicity	45.00% Black	46.67% Black	0.0332	1.004
Age	38.20 (11.74)	38.52 (13.38)	–0.0270	1.298
Prior Arrests	1.58 (1.31)	1.68 (1.23)	–0.0766	0.885
Sentence Length	28.10 (16.71)	28.63 (16.76)	–0.0319	1.005
Last Grade Completed	10.45 (2.36)	10.40 (2.27)	0.0212	0.926
Reading Level	10.07 (2.15)	10.15 (2.17)	–0.0387	1.104
Math Level	8.88 (2.34)	8.72 (2.25)	–0.0713	0.925

Continuous variables are presented with mean and standard deviation.

on the standardized difference values across all covariates except for age. Examination of the variance ratios indicates very similar distributions between the greedy and the optimal models, while the full match model exhibits, on average, greater variance ratios. Table 5 compares the standardized difference scores between the unmatched data and the three matching models.

Ideally, the final matched model would exhibit no differences on the covariates between the treatment and control groups. Of course this is unlikely, so we must find the best model possible. All three matching models indicate substantial improvement from the unmatched model. While the optimal and full matched models are very similar in terms of mean differences, the optimal model displays variance ratios closer to one and is deemed the superior model. It should be noted that while in this case the optimal match appears superior, each data set will present unique qualities that might make one matching strategy more effective than another. Multiple models should always be assessed. In addition, it is important to experiment with many combinations of covariates and covariate interactions with several matching strategies to determine which combination produces the strongest model (Stuart, 2010).

Inspection of graphical representations of the data is also helpful in determining covariate balance. Many authors (e.g., Ho et al., 2011; Imai et al., 2008; Stuart, 2010) suggest the use of Q-Q plots as a means to visually compare distributions of continuous variables between pre- and post-matched groups. The closer the data points fall on the 45-degree angle, the closer the treated and control cases are balanced. Figure 1 presents the Q-Q plots for four of the covariates of the optimally matched data. The plots indicate greater balance on the covariates after the optimal match compared with the baseline values. Although all of the Q-Q plots were examined, only four were included for space considerations.

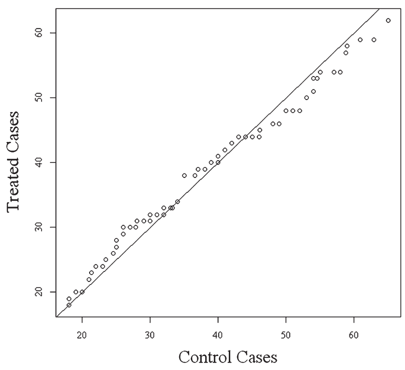
Table 5. Comparison of Standardized Differences Across Matching Models

Variable	Unmatched	Greedy	Full	Optimal
Ethnicity	0.1993	0.0712	0.0956	0.0332
Age	0.0135	–0.0791	–0.0286	–0.0270
Prior Arrests	–0.3670	–0.0684	–0.0950	–0.0766
Sentence Length	–0.1875	–0.0310	–0.0504	–0.0319
Last Grade Completed	0.3336	0.0151	–0.1130	0.0212
Reading Level	0.6306	–0.0580	–0.0266	–0.0387
Math Level	0.2389	–0.0306	–0.0313	–0.0713

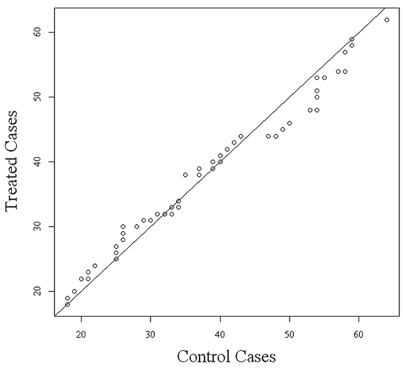
Post-match Analysis

The optimal match model results in 60 matched pairs. In this sample, the probability of re-incarceration within three years was 28.3% (17/60) and 16.7% (10/60) for the control and treated groups, respectively. This is compared to 29.2% (70/240) and 16.7% (10/60) in the unmatched data. The discrepancy in the recidivism rate between the treated and control groups declined slightly in the matched sample. McNemar's test of statistical significance is appropriate for this data since it is a 2 × 2 contingency table with dichotomous variables (treatment and recidivism) and the matched sample is considered paired (Rosenbaum, 2002). This difference, as analyzed through a one-tailed McNemar's test without the Yates adjustment produced a statistically significant result ($p = .046$). The relative risk ratio = .5882 (95% confidence interval

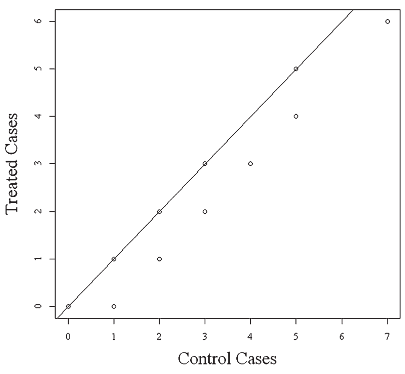
Q-Q Plot of Age—Unmatched Sample



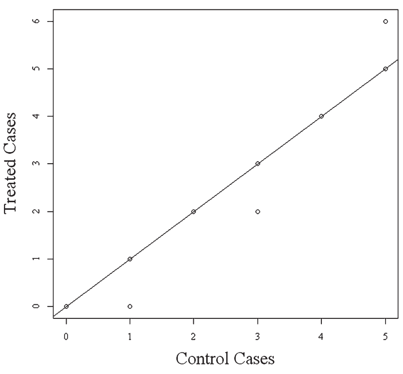
Q-Q Plot of Age—Matched Sample

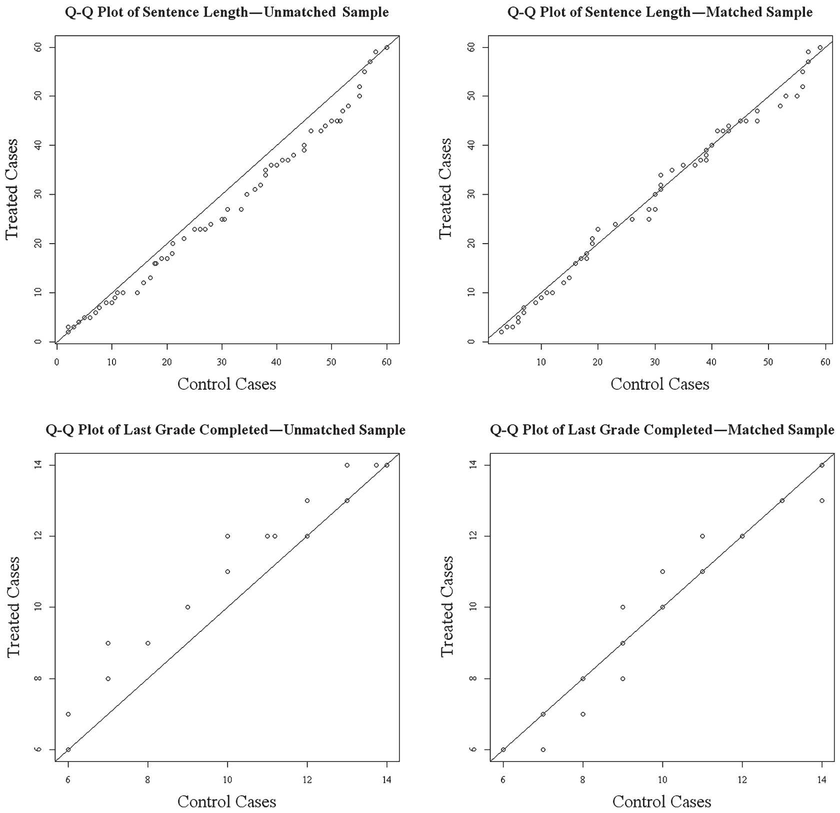


Q-Q Plot of Prior Arrests—Unmatched Sample



Q-Q Plot of Prior Arrests—Matched Sample




Q-Q Plot of Sentence Length—Unmatched Sample

Q-Q Plot of Sentence Length—Matched Sample

Q-Q Plot of Last Grade Completed—Unmatched Sample

Q-Q Plot of Last Grade Completed—Matched Sample

Figure 1. Q-Q Plots for Unmatched and Optimally Matched Data

.2938 – 1.1779), indicating that after matching, exposure to treatment is again associated with almost half the likelihood of recidivism compared to the control condition.

Sensitivity Analysis

After applying an optimal matching model to the data, we conclude that vocational training does have a statistically significant effect on three-year recidivism rates. However, a sensitivity analysis must be conducted to assess how sensitive our model is to unmeasured confounders. If it is determined that the study is highly robust, then causal conclusions are appropriate. A sensitivity analysis asks how large Γ would have to be to change the conclusions of the study. Table 6 indicates model sensitivity by presenting *p* values for upper and

Table 6. Sensitivity Analysis of Final Matched Data

Gamma	Lower Bound	Upper Bound
1.0	0.046	0.046
1.2	0.022	0.088
1.4	0.011	0.140
1.6	0.006	0.199
1.8	0.003	0.261
2.0	0.001	0.322

lower bounds for Γ values from 1 to 2 in increments of .2. This analysis was computed using the rbounds package in R (Keele, n.d.), which employs the procedures set forth by Rosenbaum (2002) for binary outcomes. The model is very sensitive to unmeasured covariates as even when $\Gamma = 1.2$, the model is no longer statistically significant. In fact, the highest value of Γ in which the model achieves statistical significance is 1.02. Since this study is sensitive to potential confounding variables, it would be prudent to omit claims of causality.

Conclusions

Since randomized studies are typically infeasible in the fields of corrections and correctional education, the utility of PSM is apparent. PSM could accurately determine the efficacy of various programs on offender outcomes, even when experimental control is lacking. With the inclusion of an appropriate sensitivity analysis, it can also provide a measure of certainty for these conclusions. Although the present example examined only three-year recidivism rates, obviously multiple outcomes can be assessed. For instance, survival analyses can be computed to time to re-incarceration. While PSM is increasingly being used in corrections research (e.g., Bales & Piquero, 2012; Bohmert & Duwe, 2012; Mears, Cochran, Siennick, & Bales, 2012), it is still underutilized and when it is used, it should include basic foundational components (such as the often neglected sensitivity analysis).

This is by no means an exhaustive review of PSM. Many details and nuances exist, for which we have only touched the surface. In addition, PSM is becoming increasingly popular and new techniques are introduced frequently. Books by Guo and Fraser (2010) and Rosenbaum (2010) are detailed resources, and both present guides in the use of relevant statistical software. PSM is processed using the R platform (available free for download:

http://www.r-project.org/), with the use of add-on packages such as MatchIt by Ho et al. (2011), optimal matching by Hansen (2004), and rbounds sensitivity analysis by Keele. PSM can also be computed using the Stata and SAS packages (see for example, Ming & Rosenbaum, 2001; Stuart, 2010).

As mentioned previously, PSM is not without limitations. It is important to keep in mind that the main issue with studies lacking randomization is that we cannot know if an outcome is due to pre-existing differences in the treatment and control groups or to the independent variable. Matching can successfully control for the pre-existing differences, but only those that are observed or those that are unobserved insofar as they are correlated with the observed covariates (Imai et al., 2008; Rosenbaum, 1989). Further, the use of PSM requires that you have access to relatively large data sets and that the treatment and control groups have strong overlap (Shadish, et al., 2002). If the data lack these qualities, then PSM may not be the best approach.

State and federal budget shortfalls and an increased awareness of the need for rigorous evaluations research in corrections (MacKenzie, 2000) highlight the need for new and innovative methodologies. While it may be necessary to rely on quasi-experimental research, simply depending on research that compares the recidivism rates of offenders who complete some program vs. those who do not is not adequate. PSM provides a strong fit given the needs and limitations of corrections research.

References

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica, 74* (1), 235–267.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine, 28*, 3083–3107.

Austin, P. C. (2010). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*.

Austin, P. C. (2011a). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics, 10*(2), 150–161.

Austin, P. C. (2011b). A tutorial and case study in propensity score analysis: An application to estimating the effect of in-hospital smoking cessation counseling on mortality. *Multivariate Behavioral Research, 46*, 119–151.

Bales, W. D., & Piquero, A. R. (2012). Assessing the impact of imprisonment on recidivism. *The Journal of Experimental Criminology, 8*, 71–101.

Bohmert, M. N., & Duwe, G. (2012). Minnesota's affordable homes program: Evaluating the effects of a prison work program on recidivism, employment and cost avoidance. *Criminal Justice Policy Review, 23*(2), 327–351.

Butler, L., Goodman-Delahunty, J., & Lulham, R. (2012). Effectiveness of pretrial community-based diversion in reducing reoffending by adult intrafamilial child sex offenders. *Criminal Justice and Behavior, 39*(4), 493–513.

Dirkzwager, A., Nieuwbeerta, P., & Blokland, A. (2012). Effects of first-time imprisonment on postprison mortality: A 25-year follow-up study with a matched control group. *Journal of Research in Crime and Delinquency, 49*(3), 383–419.

Gordon, R. D., & Weldon, B. (2003). The impact of career and technical education programs on adult offenders: Learning behind bars. *The Journal of Correctional Education, 54*(4), 200–209.

Guo, S., & Fraser, M. W. (2010). *Propensity score analysis: Statistical methods and applications*. Thousand Oaks, CA: Sage Publications.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association, 99*(467), 609–618.

Hansen, B. B., & Klopfson, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics, 15*(3), 1–19.

Haviland, A., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods, 12*(3), 247–267.

Hill, J. (2008). Discussion of research using propensity-score matching: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, *Statistics in Medicine. Statistics in Medicine, 27*, 2055–2061.

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

Ho, D., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, 42*(8), 1–28.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings among experimentalists and observationalists: Balance test fallacies in causal inference. *Journal of the Royal Statistical Society, Series A*, 171 (part 2), 481–502.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics, 86*, 4–30.

Jensen, E. L., & Kane, S. L. (2012). The effects of therapeutic community on recidivism up to four years after release: A multi-site study. *Criminal Justice and Behavior, 39*(8), 1075–1087.

Johnson, B. D., & Kurlychek, M. C. (2012). Transferred juveniles in the era of sentencing guidelines: Examining judicial departures for juvenile offenders in adult criminal court. *Criminology, 50*(2), 525–564.

Keele, L. J. (n.d.). Rbounds: An R package for sensitivity analysis with matched data.

Kim, R. H., & Clark, D. (2013). The effect of prison-based college education programs on recidivism: Propensity score matching approach. *Journal of Criminal Justice, 41*, 196–204.

Lockwood, S., Nally, J. M., Ho, T., & Knutson, K. (2012). The effect of correctional education on postrelease employment and recidivism: A 5-year follow-up study in the state of Indiana. *Crime & Delinquency, 58*(3), 380–396.

Lunt, M. (2014). Selecting an appropriate caliper can be essential for achieving good balance with propensity score matching. *American Journal of Epidemiology, 179*(2), 226–235.

MacKenzie, D. L. (2000). Evidence-based corrections: Identifying what works. *Crime & Delinquency, 46*(4), 457–471.

Mears, D. P., Cochran, J. C., Siennick, S. E., & Bales, W. D. (2012). Prison visitation and recidivism. *Justice Quarterly, 29(*6), 888–918.

Ming, K., & Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics, 56*, 118–124.

Ming, K., & Rosenbaum, P. R. (2001). A note on optimal matching with variable controls using the assignment algorithm. *Journal of Computational and Graphical Statistics, 10*, 455–463.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association, 84*(408), 1024–1032.

Rosenbaum, P. R. (2002). *Observational studies (2nd ed.).* New York, NY: Springer.

Rosenbaum, P. R. (2005). Observational study. In B. S. Everitt and D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1451–1462). New York, NY: John Wiley and Sons.

Rosenbaum, P. R. (2010). *Design of observational studies*. New York, NY: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516–524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33–38.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology, 2*, 169–188.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Boston, MA: Houghton-Mifflin.

Sherman, L. W., Gottfredson, D., MacKenzie, D. L., Eck, J., Reuter, P., & Bushway, S. (1997). *Preventing crime: what works, what doesn't, what's promising*. Washington, D.C.: National Institute of Justice.

Steurer, S. J., Smith, L., & Tracy, A. (2001). *Education reduces crime: Three-state recidivism study.* Lanham, MD: Correctional Education Association.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1–21.

Stuart, E. A., & Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology, 44*(2), 395–406.

Zgoba, K. M., Haugebrook, S., & Jenkins, K. (2008). The influence of GED obtainment on inmate release outcome. *Criminal Justice and Behavior, 35*(3), 375–387.

Biographical Sketch

JASON E. PICCONE has served as an assistant professor at Nova Southeastern University for the past eight years. He previously conducted research for the Virginia Department of Correctional Education.