NCJRS
National Criminal Justice Reference Service

Office of Justice Programs
BJA BJS NIJ OJJDP OVC SMART

The author(s) shown below used Federal funding provided by the U.S. Department of Justice to prepare the following resource:

Document Title:	**Panacea or Poison: Can Propensity Score Modeling (PSM) Methods Replicate the Results from Randomized Control Trials (RCTs)?**
Author(s):	**Christopher Campbell, Ph.D.; Ryan M. Labrecque, Ph.D.**
Document Number:	**308450**
Date Received:	**January 2024**
Award Number:	**2016-R2-CX-0030**

This resource has not been published by the U.S. Department of Justice. This resource is being made publicly available through the Office of Justice Programs' National Criminal Justice Reference Service.

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

FINAL SUMMARY OVERVIEW REPORT

Submitted to:
National Institute of Justice

Project Title:
Panacea or Poison: Can Propensity Score Modeling (PSM) Methods Replicate the Results from Randomized Control Trials (RCTs)?

Principal Investigator:
Christopher Campbell, Ph.D.

Co-Principal Investigator:
Ryan M. Labrecque, Ph.D.

Department of Criminology and Criminal Justice
Portland State University
506 SW Mill St., P.O. Box 751
Portland, OR 97207-0751
(503) 725-9896
ccampbell@pdx.edu

Recipient Organization:
Portland State University

Award Number: 2016-R2-CX-0030

Project/Grant Period: 01/01/2017 - 12/31/2017

Reporting Period End Date: 09/30/2018

This study was funded by the National Institute of Justice (Award Number 2016-R2-CX-0030). The points of view in this document are those of the author and do not necessarily represent the official position or policies of the U.S. Department of Justice.

Identifying best practices requires ample research with strong methodology. Understood as the "gold standard" in methodological strength, the randomized controlled trial (RCT) has been shown to provide reliable and valid findings (Shadish, Cook, & Campbell, 2002). Practically, however, RCTs are often difficult (and expensive) to implement. In response, statisticians have developed a matching technique known as propensity score modeling (PSM) to simulate the findings of RCT experiments (Rosenbaum & Rubin, 1983; 1985). Over the last several years, criminal justice scholars have become increasingly reliant on the use of PSM in their evaluation research (e.g., Delisi, Barnes, Beaver, & Gibson, 2009; Gaes, Bales, & Scaggs, 2016).

Although prior scholarship generally supports the utility of PSM, there are some important caveats about this statistical technique to acknowledge. For example, tests of PSM's accuracy have been limited to fields quite dissimilar from criminal justice, such as medicine (see Hill, 2008). Some literature also cautions about the underlying dangers (e.g., increased inefficiency and model dependence) in using the propensity score to match cases (King & Nielsen, 2016). Other scholarship highlights examples of many commonly overlooked issues (e.g., high sensitivity to propensity score estimation) in the application of PSM (Smith & Todd, 2001, 2005; Steiner et al., 2010; Wooldridge, 2005). More concerning, however, is the evidence that suggests under some circumstances PSM can yield the opposite results of an RCT (e.g., Peikes, Moreno, & Orzol, 2008), and that scholars appear to use PSM correctly in journal articles only 28% of the time (Austin, 2008a). Another criticism of PSM is that some researchers are simply implementing a "matching" technique with blind acceptance, which may oversimplify a complex process in calculating the propensity score (Caliendo & Kopeinig, 2008; Steiner et al., 2010). These issues are of great concern to the academic and professional communities because researchers may unknowingly be presenting misleading or even incorrect conclusions from their PSM analyses, which could lead to inappropriate policy recommendations and policies.

With the growing popularity of PSM, its technological ease of use, and the numerous concerns regarding its application, there exists a clear need for cross-validation tests to assess the reliability and validity of PSM in criminal justice research. The current project, therefore, addresses a critical question: *Can PSM methods replicate the results from RCTs of criminal justice evaluations*? For this project, we focused our efforts on the following five different PSM techniques (breaking out those with and without a caliper making seven comparable types in total), due to their common utility in criminology and criminal justice research: 1-1 (read, 1 to 1) matching (with

and without a caliper), 1-many matching[1] (with and without a caliper),[2] inverse probability of the treatment weighting (IPTW), stratified weighting scheme,[3] and optimal pairs matching.[4]

Methodology

In order to answer this question, we gathered the datasets of 10 publicly available and restricted RCT studies from the National Archive of Criminal Justice Data (NACJD), introduced an artificial selection bias into the treatment groups of these investigations, and then used each PSM technique to remove this selection bias. We then compared the results generated from the PSM methods to those derived from the original RCT experiments, and meta-analyzed the findings across all studies to reveal the true reliability and validity of PSM in relation to RCTs using criminal justice data.

Eligible RCTs. To be included in this evaluation, studies had to meet two baseline selection criteria: (1) All treatment and comparison cases had to be randomized to their respective condition,[5] and (2) the dataset had to have a minimum number of 130 treatment cases per treatment condition.[6] This number is based on what is needed in preparing the data for PSM. Power analysis indicates that to detect a true difference of a medium effect size (using Cohen's $d = .5$), with approximately .80 power, the study requires 65 cases per group (Cohen, 1988). Given the project design, working backwards from 65 cases and considering the needs of PSM,

[1] Briefly put, PSM aims to summarize covariates in a numeric score that predicts participation in the treatment (selection bias) and then allows a comparison between treatment and control cases with similar propensity scores. In a 1-1 match for every one treatment case, there is only one control case matched. The 1-1 matching for this study employed a greedy matching style, which identifies the first closest match (nearest neighbor) and removes the pair without replacement. In a 1-many match, for every one treatment case, there are multiple control cases matched. The optimal number of matched control cases were determined by using the formula provided by Hansen (2004) and guidelines of Austin (2010). For this study, the 1-many matching was done with replacement meaning control cases could be used multiple times, and then weighted by the number of times they were used.

[2] When calipers were used, they were calculated as the standard deviation of the propensity score multiplied by .25 (Rosenbaum & Rubin, 1985). Once propensity scores were conditioned for each study's biased sample, all PSM approaches for that RCT dataset relied on the same score. Propensity scores were calculated using logistic regression and any covariate that possessed over 10% bias or was significantly different between the treatment and control cases.

[3] IPTW involves a weighting formula that gives more statistical strength to those cases who look most like the treatment group, and less to those who are most different on the propensity score (see Austin & Stuart, 2015). Stratified weighting split the propensity score into equal strata and weighted the cases by their within-strata proportion (see Hong, 2010).

[4] Optimal pairs matching is like 1-1 matching in that it uses pairs, but instead of taking the nearest neighbor, the process selects matched pairs by assessing and optimizing (reducing) the overall difference in the treatment and control groups (see Guo & Fraser, 2014). Optimal pair matching was conducted in the statistical program R, while all other approaches were conducted in Stata using either manually written formulas for the weighting schemes or the *psmatch2* command.

[5] Although seemingly inherent in the term RCT, there are many studies that fall into this category, but may not have entirely random assignment. For instance, some researchers may have been required to stop randomly assigning comparison cases due to ethical issues. Such studies may introduce bias into the treatment effects and therefore are not selected.

[6] In an RCT, 130 treatment cases should ensure roughly the same number of comparison cases.

this translates to 130 cases per group in the initial RCT to act as a guide to identify appropriate studies for 1-1 greedy matching (the most data-hungry technique). A systematic search of NACJD identified 46 potential studies, from which only nine studies of different types (e.g., policing, courts, corrections) met our sample size requirements. To achieve our aim of at least 10 studies we included one quasi-experimental study that prospectively identified comparison group participants prior to the rollout of the program.

Eligibility criteria

- Used a random assignment procedure
- Possessed at least 130 participants in each condition
- Did not examine aggregate rates (because they can inflate effect sizes and contain too many possible unmeasured influences)
- Contained sufficient data to use PSM (i.e., enough measures and those indicative of distinctions between groups)

Artificial selection bias. Inherent in their randomized nature, RCT groups lack selection bias, which is what PSM is designed to remove. In order to test PSM techniques, we first systematically introduce selection bias into the treatment groups. Briefly put, the creation of artificial selection bias involved the following steps: (1) identify the most critical measures from the studies' list of variables according to the study's accompanied report; (2) from this list, identify measures that best predict the likelihood of being in the treatment group using a forward and then backward stepwise logistic regression with relatively relaxed standards of keeping or removing variables in/out of the model (e.g., $p < .5$ rather than .05); (3) construct an additive scale from the variables that successfully predicted treatment group membership (i.e., measures remaining in both forward and backward stepwise models); and (4) selecting only the treatment cases that were above the mean of the additive scale. This process was conducted "blind", meaning it was completed by the second author, and the first author then conditioned the propensity score. The blind component was used to ensure the measures used to create the bias were unknown to the person conditioning the propensity score.[7] Ultimately, the process produced between 30% and 50% of the treatment cases that had a significant bias relative to the comparison cases according to the measures of balance discussed in the next section. These biased treatment cases were merged into a dataset with the full original comparison group.

[7] The manner in which we conditioned the propensity score was an a-theoretical approach. Essentially, any covariate in the biased sample that possessed a percent bias more than 20% or yielded a statistically significant difference between the treatment and control groups was selected to be part of the logit model that conditioned the propensity score.

Comparison indices

- Percent of covariates significantly different
- Mean standardized percent bias
- Maximum percent bias
- Percent of covariates over 20% bias
- Percent of covariates over 10% bias
- Ability to reduce the AUC to .5
- Difference in effect size (Cohen's *d*)

Comparing the RCT and PSM techniques. The biased and PSM treatment and control groups were all compared to each other as well as the RCT using six measures of balance. (1) The percentage of covariates with statistically significant differences ($p < .05$). In a true experiment/RCT, we would expect to see approximately 5% of all covariates to be significantly different between the treatment and control groups just due to chance in the random assignment (Shadish et al., 2002). The standardized percent difference/bias was also calculated and compared in four ways. The standardized percent bias is a common method of identifying the degree to which treatment and control groups differ, and is the preferred method over simply using the Neyman-Pearson approach to statistical significance (see Austin, 2008b, 2011). According to Rosenbaum and Rubin (1985), the treatment and control groups should not differ on a covariate more than 20%, with less than 10% being ideal. Thus, the four ways we assessed the groups on the standardized bias included the (2) mean, (3) maximum, and overall percent of covariates that were (4) over 20% and (5) 10% bias. Lastly, we examined the (6) receiver operating characteristic - area under the curve statistic (AUC). The AUC can be used as a sensitivity check to gauge how well the propensity score predicts if a case is in the treatment group (see Austin, 2008b).[8] The closer a PSM technique can get to any of these ideal benchmarks, the closer the technique was at replicating the RCT.

Finally, for each dataset, we calculated a common metric effect size (ES) (i.e., Cohen's *d*) along with a 95% confidence interval (CI) for all of the dependent variables in the original RCT, and for the seven PSM techniques after removing the artificial bias. We then compared the PSM and RCT results on five dimensions. (1) The percentage of ES comparisons that shared the same statistical significance (i.e., $p < .05$, yes or no) and direction (i.e., positive or negative). (2) The percentage of PSM ESs that were larger in magnitude than the RCT ESs. (3) The percentage of PSM ESs that fell within the 95% confidence interval (CI) of the RCT ESs. (4) The Pearson product-moment correlation coefficient (*r*) between the PSM and RCT ESs. (5) Finally, we calculated

[8] High AUC values are expected with the biased sample to note that the propensity score is conditioned well. AUC values close to .5 should be expected after the PSM approach is applied to indicate that the propensity score can no longer identify the difference between the treatment and control cases (i.e., they are balanced).

the difference in the estimated outcomes between the RCT and PSM methodologies and averaged across the total number of outcomes included. For the purposes of this study a point estimate of zero indicates no difference in the ES between the RCT and PSM. A positive value, however, indicates the ES from the PSM method overestimates the magnitude of the treatment effect relative to the RCT, and a negative value indicates the PSM method underestimates the magnitude. The overall ES for the difference between each PSM and RCT approach was calculated using a random-effects model meta-analysis (see Bornstein, Hedges, Higgins, & Rothstein, 2009). The heterogeneity of ESs was determined by the I^2 statistic, which is an intuitive index of the discrepancy of a group of studies results (see Higgins, Thompson, Deeks, & Altman, 2003).[9]

Results

Table 1 provides the descriptive statistics of the 10 RCT studies included in our investigation. One investigation involved two separate RCTs, which allowed us to generate 11 total effect sizes. As can be seen in this table, the mean sample size per study is 573, with a range of 351 to 1,469, and the mean number of comparable covariates per study is 71, with a range of 33 to 131. In addition, the mean number of outcomes per study was 9, with a range of 5 to 22. The average effect size is represented as the mean Cohen's *d* and 95% confidence interval (CI) values across all of the outcomes within each study. Finally, the mean percentage of the original treatment group identified to be in the biased sample is approximately 42%.

Table 1. ***Descriptive statistics of RCT studies***

Study (ICPSR #)	Treatment Group *n*	Control Group *n*	# Covariates	# Outcomes	Mean Cohen's *d*	[95% CI]
1 (6358)	239	219	99	15	1.34	[1.12, 1.56]
2 (32901)	632	837	69	12	0.17	[0.07, 0.27]
3 (3201)	209	194	33	5	0.25	[0.06, 0.45]
4 (2686)	264	236	97	5	0.06	[-0.12, 0.24]
5 (8496)	176	175	82	12	-0.12	[-0.33, 0.09]
6 (34975)	214	204	131	22	0.01	[-0.18, 0.20]
7 (3353)	187	208	69	8	-0.13	[-0.33, 0.07]
8a (2890)	361	379	51	7	-0.05	[-0.19, 0.10]
8b (2890)	393	367	49	6	-0.16	[-0.30, -0.01]
9 (25261)	237	229	47	6	-0.19	[-0.37, -0.01]
10 (4307)	190	186	50	6	-0.18	[-0.38, 0.02]

[9] The interpretation of I^2 is the proportion of total variation in the estimates of treatment effect that is due to heterogeneity between studies, which is presented in percentage terms. Higgins and Thompson (2002) proposed that I^2 percentages of around 25%, 50%, and 75% indicate low, medium, and high heterogeneity amongst the ESs.

Table 2 summarizes the model fit for the original RCTs, the biased samples, and the PSM approaches. In the original RCTs, 9% of the covariate comparisons between the treatment and control group are statistically significant ($p < .05$). This estimate is much higher in the artificially biased sample (37%), and ranges within the PSM types between 2% for the 1-1 (with caliper) method to a high of 14% for the IPTW method. There is also a large increase in the mean percent bias in the biased sample compared to the original RCT (21% compared to 8%, respectively). The seven PSM methods are all able to substantively reduce the mean percent bias compared to the biased sample (range = 11% to 14%), but none fall below the estimate for the original RCT. Finally, the AUC value is much larger in the biased sample than in the original RCT (.847 compared to .661, respectively). The seven PSM methods are all also able to reduce the AUC values compared to the biased sample, and all but one, the 1-1 (no caliper) method, even reduced the AUC value below the original RCT.

Table 2. ***Model balance summary, by PSM type***

Method	% Stat. sig. diff.	Mean % bias	Max. % bias	% Bias > 20	% Bias > 10	AUC
Original RCT	9.3	7.8	27.5	5.6	28.1	.661
Biased Sample	36.9	20.7	69.5	41.6	71.5	.847
1-1 (no caliper)	3.1	10.8	38.4	13.0	45.1	.664
1-1 (with caliper)	2.4	10.5	36.9	13.0	40.5	.539
1-many (no caliper)	4.7	11.3	36.9	15.1	45.4	.575
1-many (with caliper)	4.6	11.4	38.0	15.9	45.9	.551
Stratified weighting	9.8	12.6	42.6	20.2	48.9	.526
IPTW	14.3	13.7	48.9	24.1	51.3	.562
Optimal Pairs	4.0	11.1	38.0	13.7	46.2	.650

Table 3 compares the results generated by each of the seven PSM types with those from the original RCT. Across these various methods, 55% to 68% of the PSM effect sizes are of the same statistical significance and direction as those of the RCT. Furthermore, between 70% and 75% of the PSM effect size estimates are larger than those from the RCT, and between 67% and 89% of the PSM estimates fall within the 95% confidence intervals of the RCT effect sizes. Finally, the correlations between the PSM effect size estimates and the RCT effect size estimates are all large ($r = .937$ to $.974$).[10]

[10] These estimates include the total number of outcome comparisons across all studies ($n = 104$). We also conducted the same set of analyses by averaging the ESs within each study first and then making comparisons ($n = 11$). This yielded similar findings to those presented here.

Table 3. ***Comparative results, by PSM type***

Method	% Same stat. sig. and direction	% PSM ES > RCT ES	% PSM ES within 95% CI of RCT ES	Correlation of PSM ES and RCT ES
1-1 (no caliper)	68.3	71.2	89.4	.974
1-1 (with caliper)	67.3	70.2	81.7	.937
1-many (no caliper)	56.7	72.1	67.3	.944
1-many (with caliper)	56.7	69.2	68.3	.944
Stratified weighting	54.8	69.2	74.0	.967
IPTW	48.1	72.1	63.5	.942
Optimal Pairs	65.4	75.0	86.5	.981

Table 4 presents the main meta-analytic findings of the difference in Cohen's *d* values between the PSM methods and the original RCT results. To reiterate, a difference in *d* that is zero indicates a perfect replication of the RCT, and a positive point estimate indicates PSM effect sizes are larger than (or overestimate) those generated from the original RCTs. Overall, the magnitude of observed differences are rather small (range = .03 to .09). In five of the seven comparisons, for example, the differences in the PSM and RCT estimates are not statistically significant, and in six of the seven methods the level of heterogeneity in the effect sizes is low (I^2 < 25%). Finally, both Table 4 and Figure 1 highlight the large amount of overlap between the 95% confidence intervals of the estimates across all seven of the PSM methods.

Table 4. ***Meta-analysis of difference in Cohen's d between PSM methods and RCT results***

Method	Difference in Cohen's *d*	[95% CI]	*p*	*n*	I^2
1-1 (no caliper)	.03	[-0.05 to 1.05]	.489	2,612	0%
1-1 (with caliper)	.05	[-0.04 to 0.13]	.281	2,282	5%
1-many (no caliper)	.09	[0.00 to 0.18]	.049	2,571	19%
1-many (with caliper)	.08	[-0.01 to 0.16]	.077	2,565	10%
Stratified weighting	.07	[0.00 to 0.13]	.041	3,791	1%
IPTW	.08	[-0.03 to 0.18]	.143	4,561	63%
Optimal Pairs	.06	[-0.02 to 0.14]	.143	2,602	0%

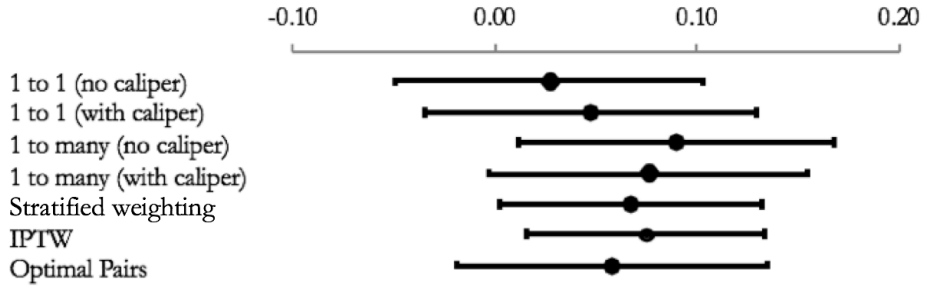

Figure 1. ***Meta-analysis of difference in Cohen's d between PSM methods and RCT results.***

Discussion

Given the mixed-findings and theoretical debate regarding the ability of PSM to establish causal inferences in social science research, we set out to assess the reliability and validity of seven PSM methods in replicating the results of 10 criminal justice RCT experiments. This investigation provides support for the use of PSM in criminal justice research. Specifically, our results suggest that these seven PSM methods can be an effective means for estimating reliable and valid simulation of RCT experiments. The strong correlations found between the RCT and PSM estimates (i.e., all *r* values > .90), indicate that the seven PSM approaches examined here were able to replicate the direction and magnitude of the effect sizes from original RCT experiments to a high degree. In addition, the average difference in Cohen's *d* between the PSM and RCT estimates were relatively small, ranging from .03 to .09 across the seven types. While the practical significance of *d* should be determined by the individual study context (Cohen, 1988), these differences fall within general allowable differences described in other cross-validation studies (i.e., $d = .24$; Dong & Lipsey, 2018)[11]. Furthermore, our meta-analyses of the differences in *d* show that only two of the seven PSM variations compared were significantly different ($p < .05$) from the original RCT. These results support an optimistic view of PSM, and further coincide with the findings from cross-validation studies analyzing other social science data (Dehejia & Wahba, 1999; Dong & Lipsey, 2018).

PSM methods ranked by percent of ESs within RCT confidence intervals

1. 1-1 (no caliper)	89.4%
2. Optimal pairs	86.5%
3. 1-1 (with caliper)	81.7%
4. Stratified weighting	74.0%
5. 1-many (with caliper)	68.3%
6. 1-many (no caliper)	67.3%
7. IPTW	63.5%

When comparing the relative abilities of the various PSM techniques, the 1-1 matching (both with and without a caliper) and optimal pairs approaches, appear to perform the best in our cross-validation studies. These three methods provided the closest replication to the original RCTs in virtually all areas, including reducing the standardized bias, the number of covariates still significantly different after the match, and producing statistically similar effect sizes. The 1-many (with and without a caliper) and weighting (stratified and

[11] It is important to note that Cohen's *d* is interpreted in relation to the discipline-specific measurement of the outcome variables (Cohen, 1988). The .24 difference in *d* was established in the education literature and specific to educational performance in relation to various intervening practices. That said, it is related to many, if not all, of the outcomes we used in that the educational measures are typically behavioral and/or scaled tests.

inverse probability) techniques were also moderately reliable and accurate, but had notable drawbacks. For instance, less than 75% of their ESs fell within the 95% confidence intervals of the RCT ESs, and less than 60% of the ESs from these techniques were of the same statistical significance and direction as the RCT ESs. Moreover, the average difference in Cohen's *d* in relation to the RCT ESs were statistically different ($p < .05$) in two of these four methods (i.e. 1-many with caliper and stratified weighting). With that said, there were instances in which the 1-many and weighting approaches provided a closer replication than the 1-1 approaches. This suggests that the reliability and validity of PSM may be dependent on situational factors that have yet to be identified. Some scholars have attempted to address these situational components when it comes to estimating differences that may be sensitive in the modeling process (e.g., Austin, 2009, 2010; Hill, 2008). However, more work is needed to isolate such boundaries of reliability and validity within criminal justice data.

The findings of this investigation also raise some concerns about using PSM techniques for determining causal effects. For example, while the majority of the PSM methods examined here yielded a statistically similar ES as the original RCT, none of these techniques had more than 89% of their ESs fall within the 95% confidence intervals of the RCT estimates. This means that at least 11% of the time, PSM provided an estimate that was significantly outside the boundaries of the original RCT estimate. In addition, the majority of the ESs generated across these seven PSM techniques appear to overestimate the findings of the original RCT. This overestimation of the treatment effect has important research and policy implications. Finally, within the average ESs generated through meta-analysis are smaller failures that exist for each of the PSM techniques. In other words, while most of the estimated ESs were comparable to the RCT, at some point each of the techniques provided a significantly different estimate for one or more of the outcome measures. The inherent problems with this, is that the one dissimilar effect size may be masked by another, more accurate estimate. These somewhat pessimistic results coincide with the mixed-findings reported in other PSM cross-validation studies (e.g., LaLonde, 1986; Smith & Todd, 2005).

Based on these findings, we would like to emphasize two critical points of caution for anyone using PSM in evaluation research. First, *test multiple techniques to gain confidence in the estimated effect sizes.* The precise number and types of techniques depend on the study's situation (e.g., number of control and treatment cases). Should the results of multiple types show that there is a significant difference in the outcome between techniques, then

considerable attention should be given to review what measures are used in conditioning the propensity score, the technique limitations, and degree of common support (comparable propensity scores) that exists between the treatment and control groups. Second, *assess balance using all appropriate methods and disseminate those assessments.* When considering the wide differences in the model fit indices, a note should be made as to what assessment indices ought to be used and reported when using PSM techniques. We recommend using all of the comparison indices we used in this study, especially for 1-1 techniques (1-1 with and without a caliper or optimal pairs) whenever possible. These indices will indicate how close the study is to the expected values of a field experimental design (see Table 2), as well as provide the researcher, and reviewers, more information to assess the performance of the PSM technique. Other indices are available for PSM weighting techniques that were not included here (e.g., assessing the ability of measures to predict the treatment pre- and post-weighting). We recommend those using weighting techniques to also include such indices offered by the literature (e.g., Guo & Fraser, 2014).

Limitations. The findings of this study should be contextualized in its key limitations. First is the fact that RCTs in criminology and criminal justice research are almost always field experiments. As indicated by Table 2, the RCTs used here were not quite to the level of expectation that a laboratory experiment might produce (e.g., less than 5% statistically significant differences among covariates). Unfortunately, this is an inherent issue in any field RCT experiment, and policy/evidence-based decisions are often made at this level based on such methodology. Subsequently, our analyses adhere to that which is most common in criminal justice studies. Second, we used multiple different types of criminal justice studies (e.g., policing, courts, corrections). While this makes the interpretation of effect sizes difficult, it allows for criminal justice researchers to view how PSM can perform in relation to the topical areas and in the justice system/criminological settings as a whole. Subsequently, our analysis provides a more methodological examination rather than a substantive/theoretical one. Third, on a methodological note, our baseline conditioning of the propensity score was purposefully a-theoretical and involved no interactions. This was to ensure the possibility of replication by removing as much subjectivity as possible. Plus, this cross-validation process provides an absolute baseline of how to condition the propensity score. It stands for future research to show how theoretical selection of the conditioning covariates would improve the match. Lastly, we did not test other techniques like matching estimators,

treatment effect estimators, Mahalanobis matching or kernel matching, thus we cannot speak to some of the criticisms about PSM (e.g., King & Nielsen, 2016). These are areas of future research, along with the use of non-inferiority/equivalence tests between the PSM and RCT effect sizes (e.g., van Wormer & Campbell, 2016), examinations of the unmatched cases, and identifying the appropriate type and number of covariates that should be used to condition the propensity score.

Conclusion. In conclusion, our cross-validation meta-analysis demonstrates clear evidence that PSM can be a reliable and valid approach to estimating causal inference in the absence of the ability to conduct an RCT experiment. However, there are also some instances in which PSM appears to fall short of the RCT. While these issues may be in relation to some limitations to our approach, others may be more situational (data and setting) related, and still others may be related to the specific technique employed. As a result, researchers and policy makers should approach the use and interpretation of PSM with cautious optimism as it will provide a reliable and valid estimate of the treatment effect *most* of the time.

References

Austin, P. C. (2008a). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, *27*(12), 2037–2049. https://doi.org/10.1002/sim.3150

Austin, P. C. (2008b). Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. *Pharmacoepidemiology and Drug Safety*, *17*(12), 1202–1217. https://doi.org/10.1002/pds.1673

Austin, P. C. (2009). Some Methods of Propensity-Score Matching had Superior Performance to Others: Results of an Empirical Investigation and Monte Carlo simulations. *Biometrical Journal*, *51*(1), 171–184. https://doi.org/10.1002/bimj.200810488

Austin, P. C. (2010). Statistical Criteria for Selecting the Optimal Number of Untreated Subjects Matched to Each Treated Subject When Using Many-to-One Matching on the Propensity Score. *American Journal of Epidemiology*, *172*(9), 1092–1097. https://doi.org/10.1093/aje/kwq224

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, *46*(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

Austin, P. C., & Stuart, E. A. (2015). The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Statistical Methods in Medical Research*. https://doi.org/10.1177/0962280215584401

Caliendo, M., & Kopeinig, S. (2008). Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys*, 22(1), 31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.x

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2 edition). Hillsdale, N.J: Routledge.

Dehejia, R. H., & Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*, 94(448), 1053–1062. https://doi.org/10.1080/01621459.1999.10473858

Delisi, M., Barnes, J. C., Beaver, K. M., & Gibson, C. L. (2009). Delinquent Gangs and Adolescent Victimization Revisited: A Propensity Score Matching Approach. *Criminal Justice and Behavior*, 36, 808–824.

Dong, N., & Lipsey, M. W. (2018). Can Propensity Score Analysis Approximate Randomized Experiments Using Pretest and Demographic Information in Pre-K Intervention Research? *Evaluation Review*, Online ahead of print. https://doi.org/10.1177/0193841X17749824

Gaes, G. G., Bales, W. D., & Scaggs, S. J. A. (2016). The effect of imprisonment on recommitment: an analysis using exact, coarsened exact, and radius matching with the propensity score. *Journal of Experimental Criminology*, 1–16. https://doi.org/10.1007/s11292-015-9251-x

Guo, S., & Fraser, M. W. (2014). *Propensity Score Analysis: Statistical Methods and Applications* (2 edition). Los Angeles: SAGE Publications, Inc.

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.

Hill, J. (2008). Discussion of research using propensity-score matching: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. *Statistics in Medicine*, 27(12), 2055–2061. https://doi.org/10.1002/sim.3245

Hong, G. (2010). Marginal Mean Weighting Through Stratification: Adjustment for Selection Bias in Multilevel Data. *Journal of Educational and Behavioral Statistics*, *35*(5), 499–531. https://doi.org/10.3102/1076998609359785

King, G., & Nielsen, R. (2016). *Why Propensity Scores Should Not Be Used For Matching*. Cambridge, MA. Retrieved from j.mp/PScore

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review*, *76*(4), 604–620.

Peikes, D. N., Moreno, L., & Orzol, S. M. (2008). Propensity Score Matching: A note of caution for evaluators of social programs. *The American Statistician*, *62*(3), 222–231. https://doi.org/10.1198/000313008X332016

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. https://doi.org/10.1093/biomet/70.1.41

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. https://doi.org/10.2307/2683903

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton Mifflin.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, *125*(1–2), 305–353. https://doi.org/10.1016/j.jeconom.2004.04.011

Smith, J., & Todd, P. (2001). Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods. *American Economic Review*, *91*(2), 112–118.

Steiner, P. M., Cook, T. D., Shadish, W. R., & H, M. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, *15*(3), 250–267. https://doi.org/10.1037/a0018719

van Wormer, J. G., & Campbell, C. (2016). Developing an Alternative Juvenile Programming Effort to Reduce Detention Overreliance. *Journal of Juvenile Justice*, *5*(2), 12.

Wooldridge, J. M. (2005). Violating ignorability of treatment by controlling for too many factors. *Econometric Theory*, *21*, 1026–1028. https://doi.org/10.1017/S0266466605050516

Studies used

1. Petersilia, Joan, Susan Turner, and RAND Corporation. Intensive Supervision for High-Risk Offenders in 14 Sites in the United States, 1987-1990. ICPSR 06358-v2. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2009-05-15. http://doi.org/10.3886/ICPSR06358
2. Taylor, Bruce, Nan D. Stein, Dan Woods, and Elizabeth Mumford. Experimental Evaluation of a Youth Dating Violence Prevention Program in New York City Middle Schools, 2009-2010. ICPSR 32901-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2012-05-18. http://doi.org/10.3886/ICPSR32901.v1
3. Davis, Robert C., Juan Medina, and Nancy Avitabile. Effectiveness of a joint police and social services response to elder abuse in Manhattan [New York City], New York, 1996-1997. ICPSR version. New York, NY: Victim Services Research [producer], 2000. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2001.
4. Harrell, Adele V., Cavanagh, Shannon, Sridharan, Sanjeev. Impact of the Children at Risk Program: Comprehensive Final Report II, Executive Summary. NCJ 193897, Washington, DC: United States Department of Justice, National Institute of Justice, 1998. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor].
5. Pate, Antony, and Sampson Annan. Reducing fear of crime: Program evaluation surveys in Newark and Houston, 1983-1984 [Computer file]. 2nd ICPSR version. Washington, DC: The Police Foundation [producer], 1985. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 1998.
6. Messing, Jill T., Campbell, Jacquelyn, Wilson, Janet S., Brown, Sheryll, Patchell, Beverly, Shall, Christine Police Departments' Use of the Lethality Assessment Program: A Quasi-Experimental Evaluation. *Final Report*. NCJ 247456, Washington, DC: United States Department of Justice, National Institute of Justice. 2014
7. Jolin, Annette, Robert Fountain, William Feyerherm, and Sharon Friedman. Portland [Oregon] domestic violence experiment, 1996-1997 [Computer file]. ICPSR version. Portland, OR: Portland State University [producer], 1998. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2002.
8. Wish, Eric D., Thomas Gray, and Jonathan Sushinsky. Experiment to enhance the reporting of drug use by arrestees in Cleveland, Detroit, and Houston, 1997. ICPSR version. College Park, MD: University of Maryland, Center for Substance Abuse Research (CESAR) [producer], 2000. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2001. http://doi.org/10.3886/ICPSR02890.v1
9. Brame, Robert, Catherine Kaukinen, Angela R. Gover, and Pamela Lattimore. Impact of Proactive Enforcement of No-Contact Orders on Victim Safety and Repeat Victimization in Lexington County, South Carolina, 2005-2008. ICPSR 25261-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2013-07-31. http://doi.org/10.3886/ICPSR25261.v1
10. Davis, Robert C., Bruce G. Taylor, and Christopher D. Maxwell. Domestic Violence Experiment in King's County (Brooklyn), New York, 1995-1997. ICPSR 04307-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2006-08-01. http://doi.org/10.3886/ICPSR04307.v1