Can PSM strategies replicate RCT results in criminal justice research?: A meta-analytic investigation

Christopher M. Campbell, PhD
Ryan M. Labrecque, PhD
Portland State University
• This project was supported by a National Institute of Justice (NIJ) Award (#2016-R2-CX-0030) Data Resources Program: Funding for Analysis of Existing Data.

• We wish to thank Shenyang Guo, Zachary Hamilton, and Ozcan Tunalilar for their valuable feedback during this process.

• We also extend our gratitude to the staff at the Justice Research and Statistical Association (JRSA) for the opportunity to facilitate this webinar and thank viewers for their interest in our work.
Propensity Score Modeling (PSM)

- Randomized controlled trials (RCTs) are described as the “gold standard” in methodological strength.
  - RCTs, however, are not always possible in many criminal justice settings.

- PSM was designed to simulate the randomization of an RCT (Rosenbaum & Rubin, 1983).
  - Balances/creates similar comparison groups using likelihood of being in the treatment
    - Predicted probability of being treated (i.e., propensity score) is a summary of the relevant observed measures.

- Over the last decade, many criminal justice researchers, agencies, and funders demonstrate a preference for PSM in evaluation research (e.g., Delisi et al., 2009; Gaes et al., 2016).
• While some herald PSM as a panacea for evaluation research, others warn of the inherent dangers in placing too much blind faith in its ability to estimate causal relationships.

• Few cross-validation studies exist that test ability of PSM to replicate the findings of an RCT.

• Previous research suggests mixed findings:
  • It is possible to achieve similar results as an RCT, but
  • PSM may also under- and over-estimate treatment effects (Dong & Lipsey, 2018)
  • and sometimes even yield opposite effect size estimates and statistical significance (Peikes et al., 2008).
Current Study

- There remains a debate whether PSM can accurately simulate the effects of an RCT.

- This investigation tests the ability of PSM techniques to replicate findings from 10 criminal justice RCTs.
  - If PSM simulates RCT findings, it provides support for the continued use of PSM in criminal justice research.
  - If PSM fails to replicate RCT findings, this is evidence for reducing criminal justice's heavy reliance on PSM.
1. Identify appropriate RCTs to test
2. Introduce artificial selection bias (i.e., remove RCT balance)
3. Conduct several PSM analyses
4. Compare balance of RCT and PSM techniques
5. Examine outcomes (i.e., difference in PSM and RCT estimates)
6. Meta-analyze the results
7. Moderator analyses of important PSM traits
We collected eligible studies from the National Archive of Criminal Justice Data (NACJD) using two baseline selection criteria:

1. All cases must be randomly assigned to conditions.
2. Dataset must have at least 130 cases per condition – determined through power analysis.

This process identified 9 eligible studies out of 46 potential RCTs.

- We included one additional quasi-experiment because it prospectively identified a temporal control group.

These studies varied in focus (e.g., policing, courts, corrections).
**ARTIFICIAL SELECTION BIAS**

- Dr. Labrecque recoded these 10 datasets and introduced bias to remove RCT balance.
  1. Determine critical study variables via NACJD report
  2. Identify measures that best predict treatment group placement
  3. Construct an additive scale from the predictive variables
  4. Select only the treatment cases above the additive scale’s mean to create “biased sample”
DESCRIPTION OF SAMPLE

- Total studies included = 10
  - One study involved two separate RCTs, which allowed us to generate 11 effect size (ES) estimates.

- Mean sample size = 573 (Range = 351 to 1,469).
- Mean # of covariates per study = 71 (Range = 33 to 131).
- Mean # of outcomes per study = 9 (Range = 5 to 22).

- Mean % of original treatment group identified in biased sample = 42%.
Dr. Campbell then conditioned the propensity score using biased sample and applied 7 PSM variants:

- 1-1 matching (with and without a caliper)
- 1-many matching (with and without a caliper)
- Inverse probability of the treatment weighting (IPTW)
- Stratified weighting scheme
- Optimal pairs matching
We assessed group balance in the RCT, biased, and PSM samples using six comparison indices:

<table>
<thead>
<tr>
<th>Comparison Indices</th>
<th>Aim</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Statistical significance ($p &lt; .05$)</td>
<td>Less than 5% of covariates significant</td>
</tr>
<tr>
<td>2. Mean standardized percent bias (% bias)</td>
<td>At least less than 20%</td>
</tr>
<tr>
<td>3. Maximum % bias</td>
<td>Minimize to less than 20%</td>
</tr>
<tr>
<td>4. Percent of covariates &gt; 20% bias</td>
<td>Minimized to similar as RCT</td>
</tr>
<tr>
<td>5. Percent of covariates &gt; 10% bias</td>
<td>Minimized to similar as RCT</td>
</tr>
<tr>
<td>6. Area under the receiver operating characteristic (AUC)</td>
<td>Reduced to .500</td>
</tr>
</tbody>
</table>
## Model Balance Summary

<table>
<thead>
<tr>
<th>Comparison Indices</th>
<th>Original RCT</th>
<th>Biased Sample</th>
<th>PSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent significant differences</td>
<td>9.3</td>
<td>36.9</td>
<td>6.1</td>
</tr>
<tr>
<td>Mean standardized bias</td>
<td>7.8</td>
<td>20.7</td>
<td>11.6</td>
</tr>
<tr>
<td>Maximum percent bias</td>
<td>27.5</td>
<td>69.5</td>
<td>40.0</td>
</tr>
<tr>
<td>Percent bias over 20</td>
<td>5.6</td>
<td>41.6</td>
<td>16.4</td>
</tr>
<tr>
<td>Percent bias over 10</td>
<td>28.1</td>
<td>71.5</td>
<td>46.2</td>
</tr>
<tr>
<td>Area under the curve (AUC)</td>
<td>.661</td>
<td>.847</td>
<td>.581</td>
</tr>
</tbody>
</table>
ES Calculation and Comparative Tests

- Cohen’s $d$ was selected as the common metric with 95% confidence intervals (CIs) to estimate the magnitude of the ES for each study outcome ($n = 104$).

- Compared PSM effects size estimates to those from the original RCT on a number of dimensions:
  - Statistical significance
  - Direction
  - Magnitude
  - Overlap
## Comparative Results, by PSM Type

<table>
<thead>
<tr>
<th>PSM method</th>
<th>% Same stat. sig. and direction</th>
<th>% PSM ES &gt; RCT ES</th>
<th>% PSM ES within 95% CI of RCT ES</th>
<th>Correlation ($r$ ES)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1 (no caliper)</td>
<td>68.3</td>
<td>71.2</td>
<td>89.4</td>
<td>.974</td>
</tr>
<tr>
<td>1-1 (w/caliper)</td>
<td>67.3</td>
<td>70.2</td>
<td>81.7</td>
<td>.937</td>
</tr>
<tr>
<td>1-many (no caliper)</td>
<td>56.7</td>
<td>72.1</td>
<td>67.3</td>
<td>.944</td>
</tr>
<tr>
<td>1-many (w/caliper)</td>
<td>56.7</td>
<td>69.2</td>
<td>68.3</td>
<td>.944</td>
</tr>
<tr>
<td>Stratified weighting</td>
<td>54.8</td>
<td>69.2</td>
<td>74.0</td>
<td>.967</td>
</tr>
<tr>
<td>IPTW</td>
<td>48.1</td>
<td>72.1</td>
<td>63.5</td>
<td>.942</td>
</tr>
<tr>
<td>Optimal pairs</td>
<td>65.4</td>
<td>75.0</td>
<td>86.5</td>
<td>.981</td>
</tr>
</tbody>
</table>
The Cohen’s $d$ estimates were averaged across all outcome measures within each study.
  - One for each RCT and one for each PSM method were examined.

We calculated the ES for the meta-analysis as the difference in $d$ between the PSM and RCT results.
  - An ES of 0 indicates PSM perfectly replicates RCT.
  - A positive ES indicates PSM overestimates RCT.
  - A negative ES indicates PSM underestimates RCT.

Random effects model results are reported.
META-ANALYSIS OF DIFFERENCE IN ES

<table>
<thead>
<tr>
<th>$p$</th>
<th>$n$</th>
<th>$I^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.489</td>
<td>2,612</td>
<td>0%</td>
</tr>
<tr>
<td>0.281</td>
<td>2,282</td>
<td>5%</td>
</tr>
<tr>
<td>0.049</td>
<td>2,571</td>
<td>19%</td>
</tr>
<tr>
<td>0.077</td>
<td>2,565</td>
<td>10%</td>
</tr>
<tr>
<td>0.041</td>
<td>3,791</td>
<td>1%</td>
</tr>
<tr>
<td>0.143</td>
<td>4,561</td>
<td>63%</td>
</tr>
<tr>
<td>0.143</td>
<td>2,602</td>
<td>0%</td>
</tr>
</tbody>
</table>
M O D E R A T O R  A N A L Y S E S

• 1-1 approaches performed the best (ES = .05), followed by weighting (.08) and 1-many (.09).

• We found little difference in the effectiveness of PSM methods using or not using a caliper:
  • Average difference in $d$ of .02 for 1-1 approach and .01 for 1-many approach.

• Studies with fewer covariates available to match produced more valid results than those with more covariates available (average difference in $d$ of .04 compared to .08, respectively).

• Studies with greater post-match bias reduction averaged much better results than those with less bias reduction (average difference in $d$ of .07 compared to .11, respectfully).
This study provides support for the use of PSM in criminal justice research.

Our results suggest that PSM can be an effective means for simulating an RCT experiment:

- PSM replicated the direction and magnitude of RCT findings to a high degree.
- Magnitude of ES differences across PSM types were rather small (difference in $d$ range = .03 to .09).
- In 5 of the 7 PSM comparisons, differences in $d$ were not statistically different than 0.
However, this investigation also raised some concerns about using PSM to estimate causal effects:

- None of the PSM techniques examined had more than 89% of its ESs fall within the 95% CI of the RCT.
- The majority of PSM ESs overestimated the treatment effects.
- At some point, each PSM technique provided an estimate that was significantly different from the RCT on one or more of the outcome measures.
Recommendations for PSM Application

1. Assess balance using appropriate methods and disseminate those assessments.
   • We recommend at a minimum to include the six indices described here.

2. Strive to collect quality over quantity measures.
   • Efforts to increase the number of covariates (vs. theoretical, quality covariates) may be counterproductive.

3. Check the findings of PSM analyses against other techniques to gain confidence in the results.
   • If results differ, attention should be given to…
     • the measures used to condition the propensity score (theoretical or not),
     • the amount of common support between treatment and control (compatibility), and
     • the limitations of the specific PSM techniques used (if others had problems with it).
   • If doubt exists about which technique to use/report, we recommend erring on greatest bias reduction.
Conclusion

• Researchers should approach the use and interpretation of PSM with a cautious optimism.

• Our cross-validation meta-analysis provides strong evidence that PSM can be an effective approach for estimating causal inferences most of the time.

• Nevertheless, our study did not replicate the experimental findings with absolute accuracy, and in some cases, PSM overestimated the treatment effects quite concerningly.

• We encourage more research on this topic to help unpack what conditions PSM is more/less effective.
Contact Information

Christopher M. Campbell, Ph.D.
Assistant Professor, Department of Criminology and Criminal Justice
Portland State University

Phone: 503-725-9896
E-mail: ccampbell@pdx.edu

Ryan M. Labrecque, Ph.D.
Assistant Professor, Department of Criminology and Criminal Justice
Portland State University

Phone: 503-725-5164
E-mail: rml@pdx.edu
Web: www.ryanmlabrecque.com
Included Studies

1. Intensive Supervision Probation (ICPSR 6358)
   - Petersilia & Turner (1987-1990)
2. Youth Dating Violence Prevention (ICPSR 32901)
   - Taylor et al. (2009-2010)
3. Response to Elder Abuse (ICPSR 3130)
   - Davis et al. (1996-1997)
4. Children at Risk Program (ICPSR 2686)
5. Reducing Fear of Crime (ICPSR 8496)
   - Pate & Annan (1983-1984)
6. Police Use of Lethality Assessments (ICPSR 34975)
   - Messing et al. (2009-2013)
7. Portland DV Experiment (ICPSR 3353)
   - Annette et al. (1996-1997)
8. Reporting of Drug Use (ICPSR 2890)
   - Wish et al. (1997)
9. Enforcement No Contact Order (ICPSR 25261)
   - Brame et al. (2005-2008)
10. King County DV Experiment (ICPSR 4307)
    - Davis et al. (1995-1997)


