Measuring Program Effects in Home Visiting Evaluation: Improving Estimates With Propensity Score Matching

Evaluation Brief

Design Options for Home Visiting Evaluation

March 2021

Introduction

Many home visiting programs want to understand their impact on children and families. To do so, evaluators must establish a suitable group of people to compare against those receiving home visiting services, known as the treatment group. Randomized controlled trials (RCTs) are considered the gold standard for ensuring similar comparison and treatment groups, on average.

When RCTs are not feasible, quasi-experimental designs that employ matching methods like propensity score matching can help evaluators craft a sound comparison group. Rosenbaum and Rubin were first to suggest that propensity score matching can achieve unbiased estimates of program impact or effect when there is balance or equivalence between treatment and comparison groups—i.e., no significant difference between the groups on key variables.[1] Evaluators must take careful steps to ensure a suitable match between the groups.

This brief introduces several approaches to matching before delving into propensity score matching. It describes key steps of propensity score matching, provides approaches for decreasing bias, and suggests resources that support high-quality matching. The brief is geared to evaluators of the Maternal, Infant, and Early Childhood Home Visiting (MIECHV) Program who are already using matching techniques, and it may inform others who are considering propensity score matching. A hypothetical scenario about evaluating home visiting's impact on school readiness illustrates the steps of propensity score matching.

MIECHV funds the development and implementation of evaluations by awardees. MIECHV is administered by the Health Resources and Services Administration in collaboration with the Administration for Children and Families.

Matching

Matching is the process of identifying similar members of treatment and comparison groups to assess outcomes for estimating program effect. Researchers match members based on measurable key characteristics, known as *observed variables*. Ensuring that treatment and comparison groups are *balanced* or *equivalent* on observed variables diminishes *selection bias*. Unlike RCTs, though, matching cannot ensure comparability on unobserved variables. Exhibit 1 outlines several matching approaches that can be used alone or in combination.

DOHVE

OPRE Report #2021-55

Exhibit 1. Sample Matching Methods

Approach	Comparison type
Exact matching	Identical to the treatment case or subject on all matching variables
Coarsened exact matching	Match on variables that have been coarsened
Mahalanobis distance matching	Closest match to the treatment case using the Mahalanobis distance metric
Propensity score matching	Closest match using a propensity score

Note: The Mahalanobis distance metric is the distance of the test point from the center of the mass divided by the width of the ellipsoid in the direction of the test point.[2] Coarsening is a reduction in the number of values for a matching variable to increase the number of matches.[3]

In exact matching, evaluators find a comparison *case* identical to the treatment case on each *matching variable*. Exact matching is more likely with a large pool of candidates and relatively few matching variables. It becomes challenging when a member of the treatment group best matches a comparison candidate on one variable and a different comparison candidate on another.[4] One option is broadening the range of values considered to be an exact match, as occurs in coarsened exact matching. Another option is Mahalanobis distance matching, a form of cluster analysis to identify the closest match.[5]

Propensity score matching, on the other hand, allows for matching on multiple characteristics at once by collapsing matching variables into one value. Propensity score matching is a common method in MIECHV outcome evaluations.

Propensity Score Matching

Propensity implies an inclination to do something, such as participate in a program. Evaluators can estimate a person's likelihood or probability of participation given a variety of personal characteristics, such as age, gender, and race. This probability is called a *propensity score.*

Evaluators estimate propensity scores for all cases—regardless of actual program participation—by combining the predictive ability of all observed variables into one value. This estimation allows evaluators to look for similar propensity scores across treatment and comparison cases,[6] rather than comparing cases by individual variables. Propensity scores decrease the likelihood of selection bias by helping evaluators match on a larger number of variables than some other matching methods. Propensity score matching is only possible when similar data are available for both the treatment and comparison groups, as described later in this brief.

Selected Key Terms

- **Matching:** The process of comparing members of treatment and comparison groups, typically by pairing across groups, using weights, or stratifying data.
- **Observed variables:** Characteristics that are measured and used in the matching analysis.
- **Balance or equivalence:** When there is no meaningful difference between treatment and comparison groups at baseline on variables used to match the groups as assessed using standardized bias testing, variance ratio testing, and/or other accepted approaches.
- **Selection bias:** When the process of selecting people to participate in a study results in a treatment group and/or comparison group not representative of the intended study population; selection bias can lead to an inaccurate estimate of program effect.

Rosenbaum and Rubin argue that bias can be eliminated if propensity score matching meets two conditions:[7]

- **Conditional independence assumption.**[8] A critical assumption is that the evaluator can control for all variables that simultaneously influence both selection into treatment and potential outcomes. When this occurs, then results are assumed to be independent of treatment status and free of selection bias. This is a strong assumption and is challenging to meet in many research settings.
- **Common support condition.**[9] This assumption indicates that each potential value of the variables used for matching could be found in both the treatment and comparison group. In other words, when comparing treatment and comparison groups, there is overlap in the distribution of the matching variables or the propensity scores.

A large dataset of variables and cases can help evaluators meet both conditions. Because the conditional independence assumption cannot be tested, evaluators should be prepared to affirm the plausibility of conditional independence (e.g., by using hypothetical sensitivity tests). Evaluators can test the common support condition by comparing the distribution of propensity scores between treatment and comparison groups.

Propensity score matching includes several steps in preparation for the final analysis as described in this brief: (1) estimating the propensity score, (2) matching cases, and (3) assessing balance. Each step provides opportunities to meet the conditions outlined above, decreasing bias in the estimates for program effect.

Estimating the Propensity Score

To calculate a propensity score, evaluators must first specify a propensity score matching model with a thoughtful set of variables used to predict treatment status. Evaluators should then consider the availability of data for each variable. These steps require careful attention to meet the two conditions previously stated to reduce bias.

Using a Well-Conceived Set of Matching Variables

Trustworthy results rely on a good match between treatment and comparison groups, including careful selection of matching variables. Identification of matching variables should not be limited to available data. Instead, evaluators should engage in a strong conceptual process that considers the following to identify matching variables:

- **Treatment predictors.** Evaluators should match on as many variables as possible that predict selection into treatment. These can include demographic characteristics or other factors associated with program participation.[10]
- **Confounders.** Evaluators should also prioritize matching on *confounders,* variables that predict both the likelihood of participating in the program and the outcome of interest.[11]

Selected Key Terms

- **Case:** A study subject; this is typically a parent or child in home visiting evaluation, depending on the research question and dataset.
- **Matching variables:** Characteristics used to match treatment and comparison cases. They are typically predictors of program participation and/or confounders.
- **Propensity score:** The probability that each case (both treatment and comparison) participates in the program, given the set of observed traits used to predict the score.
- **Confounder:** A variable that predicts both program participation and the outcome of interest.

- **Combination.** Many evaluators match on a combination of demographics, treatment predictors, confounders, and a baseline measure of the outcome, if available.

Matching aims to produce unbiased estimates for observed variables only; the greater the number of matching variables, the less likely selection bias will occur.[12] MIECHV evaluators can take several steps to identify variables for matching.

Understand Program Selection

Evaluators need to consider the characteristics of people who are likely to agree to participate in home visiting. Potential matching variables may relate to—

- **Program priorities.** Evaluators can review a home visiting program's selection criteria and priority participants (e.g., teen parents, first-time parents) to understand who is most frequently offered services.
- **Participant characteristics.** Evaluators can compare characteristics of home visiting participants to nonparticipants in a program's catchment area. For example, evaluators with access to a program's management information system data can run the frequency distribution of a wide variety of family characteristics. Data from the U.S. Census, Medicaid, Department of Social Services, or other service providers can provide a comparison.
- **Staff observations.** Evaluators can interview intake workers and other staff about the characteristics of people who sign up for home visiting compared to those who decline to participate.
- **Research literature.** Evaluators can review previous findings about key family characteristics that influence or predict participation in home visiting.

Evaluating Home Visiting's Impact on School Readiness:
Program Selection Example

Evaluators are considering the impact of a home visiting program on school readiness. A review of program priorities and intake data suggests that families participating in the program are typically*—

- First-time parents
- Teen parents
- Lower income
- Parents of a low-birthweight baby
- Recruited from two birthing hospitals

Comparison to data from the American Community Survey and State Department of Health confirms that participating families are more likely to have these characteristics than the catchment area's general population. A review of program practices also reiterates that recruitment focuses on two of the region's six birthing hospitals. Patient demographics in the two hospitals prioritized for recruitment are reflective of the region.

**This list has been truncated for illustrative purposes.*

Identify Confounders

Prior research can also identify potential confounders. Evaluators should review the literature for participant characteristics or conditions found to be linked to outcomes of interest. After developing a list of outcome predictors, evaluators should then consider if program participants are more likely to experience that characteristic or condition compared to people who do not participate in home visiting. Each characteristic that predicts both the outcome and likelihood of home visiting participation is a potential confounder and, therefore, a strong candidate for a matching variable.

Evaluating Home Visiting's Impact on School Readiness: *Confounders Example*

Evaluators conduct a literature review and find the following characteristics* predict school readiness:

- **Teen parenthood.** Children born to mothers less than 20 years old tended to have less-developed math skills when starting kindergarten than children born to mothers age 20 or older.[13]
- **Child's birthweight.** Children born at very low birthweight (< 1,500 grams) showed lower readiness in reading and math at kindergarten entry than other children. Lower math readiness persisted for children born at low birthweight (≥ 1,500 grams and < 2,500 grams).[14]
- **Family income.** Children in families with less income and fewer resources than other children scored lower on all domains of kindergarten readiness.[15]
- **High-quality early learning.** Children experiencing high-quality preschool and childcare early-learning environments showed greater skills in language, reading, and math than other children. They also had greater social competence and fewer behavior issues.[16]

Next, the evaluators consider whether program participants are also more likely to experience these predictors of school readiness. Based on their program selection review, the evaluators know that each characteristic (except for high-quality preschool) also relates to program participation. The overlapping characteristics are likely confounders and, therefore, strong candidates for matching.

**This list has been truncated for illustrative purposes.*

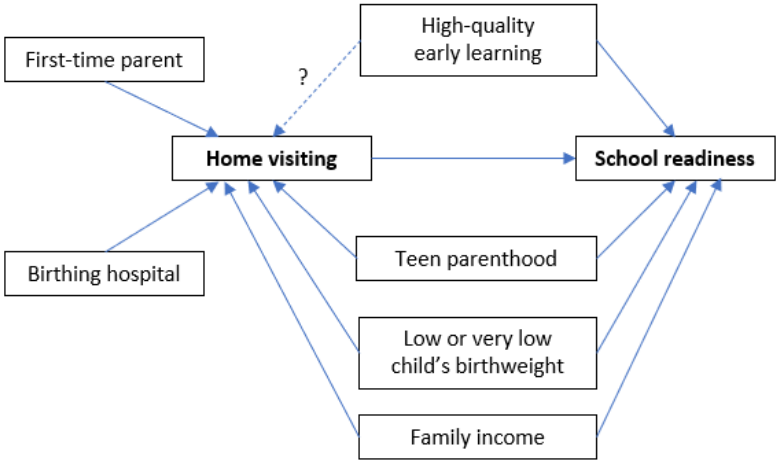
Use a Directed Acyclic Graph

Epidemiologists commonly use a directed acyclic graph (DAG) to think through and depict the relationships among the treatment, outcome, confounders, and other variables.[17] Evaluators can also use a DAG to identify matching variables. A DAG should reflect literature findings and strong theory when research results are not available. Exhibit 2 provides a DAG for the hypothetical example presented throughout this brief. For example, the DAG illustrates earlier findings that teen parenthood, child birthweight, and parent income relate to both home visiting participation and school readiness. All three characteristics are confounders and key matching variables.

The DAG also shows the association between the birthing hospital and a family's selection into home visiting. The birthing hospital is likely not predictive of school readiness, however, because patients in the targeted hospitals are representative of the region. Likewise, being a first-time parent is predictive of home visiting participation—but not necessarily kindergarten readiness—based on the literature review. Despite not being confounders, both characteristics are candidates for matching variables as predictors of program participation.

Lastly, the DAG highlights the uncertain role of high-quality early learning as a potential confounder and matching variable candidate. While the research finds that participation in high-quality learning environments predicts school readiness, the initial review of program selection did not indicate if home visiting families are more or less likely to participate in high-quality early learning than similar families. Evaluators should further consider the theoretical relationship between high-quality early learning and home visiting to clarify its role as a potential confounder and matching variable.

Exhibit 2. Partial Example of a Directed Acyclic Graph



Note: The example has been truncated for illustrative purposes. A more complete DAG will reflect all key confounders, selection predictors, and other related variables. DAGs also help evaluators identify collider variables and backdoor pathways,[a] which are fundamental to covariate selection but beyond the scope of this brief. See the resources section for more information.

Justify Matching Variables

Careful selection of the matching variables is essential for decreasing bias in propensity score matching. Evaluators should explain why they selected each matching variable in a study report or paper to demonstrate efforts to decrease bias in estimates of program effect and establish credibility for their findings. They can include study citations for variables identified in prior research and/or a completed DAG. Lack of explanation leaves the study open to criticism that key variables were omitted.

[a] A collider occurs when two variables share the same effect. A backdoor pathway results when two variables share the same cause. Williams, T. C., Bach, C. C., Matthiesen, N. B., Henriksen, T. B., & Gagliardi, L. (2018). Directed acyclic graphs: A tool for causal studies in paediatrics. *Pediatric Research*, *84*(4), 487–493.

Accessing Data for Key Variables

Once evaluators have developed a robust list of matching variables, they should work to locate data on each variable for both the treatment and comparison groups. When data are not available for a key variable, estimates of program impact could be biased.

Home visiting evaluators often use administrative records and programmatic data in MIECHV outcome evaluations. Potential sources of administrative data for matching variables and study outcomes include birth records, Medicaid or insurance data, and school records, among others. An added benefit is that data in administrative records from outside the home visiting program are typically collected in a consistent manner for both treatment and comparison families.

Evaluators should access data and then determine what variables are available in each data source, how they are measured, and the degree of missing data within each variable. Cases missing a variable will drop from the propensity score matching analysis, so evaluators should consider the appropriateness of options to address missing data such as multiple imputation. Evaluators look for baseline or pre-test variables measured in the same way and ideally from the same source for both the treatment and comparison groups when possible. They should not use variables potentially impacted by treatment if measures are only available after treatment has begun.

When a matching variable is not available in the initial data sources, evaluators need to consider alternatives to fill the gap and decrease bias such as—

- **Primary data collection.** Evaluators can collect data directly from treatment and comparison families, although locating families may be time intensive, require substantial resources, and need additional institutional review board approval. A less-burdensome option for evaluators is contacting service providers, such as home visitors, health providers, and childcare staff, to supply data on missing variables.
- **Administrative data.** Evaluators can search for administrative data sources not previously considered or pursued. Sources may be available at the local, state, or federal levels, or from university colleagues.
- **Proxy variables.** Evaluators can consider proxy variables similar to and/or highly associated with variables of interest. Proxy variables may be available in administrative, government, or other data sources, such as the U.S. Census.
- **Acknowledgement of missing variables.** Evaluators who cannot find a suitable data source should note the key variables left out of the analysis and discuss the potential impact on estimates as part of a thorough discussion of study limitations. A variable that is only marginally associated with receiving program treatment will introduce less bias in the estimate of program effect if omitted. A variable that is highly predictive of program participation will yield greater bias if not included.

Evaluating Home Visiting's Impact on School Readiness: *Addressing Missing Variables Example*

Evaluators find four matching variables in birth record data: birthing hospital, first-time parent, teen parenthood, and child's birthweight. They do not find family income at time of child's birth.

Matching variable	Birth record data
Birthing hospital	✓
First-time parent	✓
Teen parent at time of birth	✓
Birthweight	✓
Family income	-

Literature shows that the missing variable is a key predictor of school readiness; not including family income would likely bias the study findings. After much discussion, the evaluators secure state Medicaid enrollment data to use enrollment status as a justifiable proxy for family income. They also plan to use maternal education level from the birth record as a second proxy for income.

Calculating the Estimated Propensity Score

Evaluators who have selected variables, identified data sources, and developed a dataset can then estimate the propensity score. This process involves predicting the probability of a case being in the treatment group given its baseline values for the matching variables. This is typically done with a logistic regression of treatment status on baseline characteristics.

Put simply, the process begins by looking at the common characteristics of those participating in the program or receiving treatment. It then compares each case (both treatment and comparison) to these common characteristics to estimate the likelihood that someone participated in the program given their characteristics. Those most resembling the typical treatment case have a propensity score close to 1. Once propensity scores are developed for all treatment and comparison cases, the evaluator should then assess common support or the degree of overlap among the scores as described earlier. Ensuring common support helps establish a better match and decrease bias in estimates of program impact.

Widely used statistical packages (e.g., Stata, R, SAS, SPSS) have code to estimate propensity scores. See software user guides for more information.

Matching Cases

Once the propensity score is estimated, evaluators should choose an approach for using the score to match treatment and comparison cases:[18]

- **Pair matching.** Evaluators match a treatment case with a comparison case based on similar propensity scores. Evaluators should consider the following parameters for pair matching:
    - *Number of matches.* One-to-one matching creates a pair with one treatment and one comparison case. One-to-many matching links multiple comparisons to one treatment.

  - *Replacement of cases.* Comparison cases can be matched once or placed back in the pool to be matched again.
  - *Order of matching.* In greedy matching, a treatment case is selected at random and matched with a comparison case that has the closest propensity score. Once a pair is made, the match is not reassessed. Matching continues until all cases are matched. Optimal matching seeks to minimize the difference in propensity scores in each pair. Pairs are not finalized until all cases have been matched.
  - *Distance of scores.* Nearest neighbor matching pairs the treatment case with the comparison that has the closest propensity score. Evaluators may choose to set the maximum allowable distance (i.e., a caliper) between the two cases' propensity scores to help ensure better matches.
- **Weights.** Evaluators use weights based on the propensity score, typically the inverse of the probability of treatment, and functions similar to a survey sampling weight. Weights will be large and potentially unstable for those with a very low probability of treatment. Evaluators should explore methods to stabilize weights in these cases.
- **Stratification.** Evaluators separately sort treatment and comparison groups from highest to lowest propensity score. Cases are then assigned to subgroups based on predetermined propensity score cutoffs. Each treatment subgroup is matched with a similar comparison subgroup. Evaluators need to determine the maximum number of subgroups (typically five) and the propensity score cutoffs for each.
- **Covariate adjustment.** Evaluators use the propensity score as a covariate in the outcome model, regressing the outcome on treatment status and the propensity score.

The matching approach is limited to cases with an acceptable match, as defined by the evaluators' parameters. Depending on the approach, some treatment and comparison cases may not receive a match. Cases without a match are excluded (i.e., pruned) from the outcome analysis. Evaluators should review how many treatment cases are pruned and consider the characteristics of each omission to determine the potential impact on the outcome analysis and potential bias.

Various studies comparing these approaches have found that pair matching is better at addressing baseline differences between treatment and comparison groups than stratification and covariate adjustment. Weighting achieved similar results to pair matching in some scenarios.[19] As outlined in the appendix, choosing an approach

Evaluating Home Visiting's Impact on School Readiness: *Matching Cases Example*

Evaluators start with pair matching, the most common use of propensity score matching, to assess data on 50 children whose families participated in the home visiting program. Since this is a relatively small number, the evaluators choose a one-to-many approach, matching each treatment case with five comparisons. The pool of potential comparison cases gathered from statewide school data is quite large, so the evaluators opt for matching without replacement. These choices boost the study *n* to 300. The evaluators also opt for optimal matching and set a caliper of .1 after testing various caliper widths for best fit based on mean square error, as discussed in Wang et al. found in the resources section.

commonly involves a tradeoff between reducing bias and increasing efficiency. See the appendix and resources section for more information.

Assessing Balance

Matching can approximate results from an RCT only to the extent that treatment and comparison groups are well matched. Evaluators should confirm groups are equivalent after matching to help ensure an unbiased estimate of program effect.

Testing for Balance

Once the treatment and comparison groups are identified through matching, evaluators must test for equivalence or balance between the groups. Common forms of testing include—

- **Significance testing.** One common approach is comparing treatment and comparison groups on each matching variable—one at a time—by running a chi square or *t*-test. If there are no statistically significant differences between the treatment and control groups on any of the matching variables, they are considered balanced, and baseline equivalence is established. Some experts indicate significance testing is not appropriate to test balance and should be presented with a more appropriate test, if used.[20]
- **Standardized bias testing.** A more rigorous approach is to test for standardized bias on each matching variable. Similar to effect size, standardized bias is the difference between the group means (or proportions) divided by the standard deviation.[21] Groups are commonly considered to be balanced on a variable when the standardized bias is less than .1.[22]
- **Variance ratio testing.** Beyond balancing the means on the covariates, propensity score matching should also balance variance. The variance ratio is calculated for each variable by dividing treatment group variance by comparison group variance. Evaluators use an *F*-distribution to interpret the ratio.[23]
- **Plotting.** Evaluators can use a variety of graphical methods for visual comparison, such as side-by-side boxplots and cumulative distribution functions.[24]

Evaluators should test for balance using the same approach to matching they plan to use for the outcome analysis such as pair matching, weights, stratification, or covariance adjustment. For instance, if matching the outcomes will be assessed using propensity score weights, then the same weights should be applied to the treatment and comparison groups when assessing balance.

Adjusting to Achieve Balance

Testing for balance may show that treatment and comparison groups still differ on one or more variables. Evaluators should make adjustments to ensure baseline equivalence between groups to reduce bias in estimates of program effect. There are several options to help establish balance:

- **Modify the matching model.** Propensity score models typically use logistic regression that assumes a linear relationship between matching variables and the likelihood of program participation. Evaluators can adjust the matching model to explore nonlinear relationships. Options include adding an exponent to the variable that is not balanced, introducing an interaction term that uses the unbalanced variable, or adopting another statistical approach. Evaluators may also adjust the number of matches and decide whether to use replacement or calipers.

- **Adjust the list of matching variables.** Evaluators can weigh the need for variables on which the groups are significantly different. For example, can a variable be removed from the matching model without introducing much bias? Removal may be justified if research shows only a weak association between the variable and the treatment status. Evaluators can also consider adding a variable overlooked in the program selection review, previous literature, or theory.
- **Prune cases.** Evaluators can review the frequency distribution of the unbalanced variable across cases. Perhaps there is a small number of cases quite different from the others. Graphing can help evaluators identify outliers to remove from the analysis before retesting for balance. If treatment and comparison groups are balanced once these cases are removed, the analysis may be conducted without them. Case pruning should be done with caution and noted in evaluation reports or papers. See the resources section to learn more about pruning and related concerns.

Evaluating Home Visiting's Impact on School Readiness: *Adjustments to Establish Equivalence Between Groups Example*

Evaluators find no standardized differences between treatment and comparison groups on all variables, except for Medicaid enrollment. They consider three options to establish balance.

- **Adjust the model.** Bivariate analysis suggests that Medicaid enrollment is associated with another matching variable, teen parenthood. The evaluators introduce an interaction term between these two variables into the propensity score model, rerun the scores, and reassign matches. Further testing establishes that the groups are now balanced on all variables.
- **Remove the unbalanced variable.** Research establishes family income as a confounder, so evaluators cannot remove it from the model without introducing bias.
- **Prune cases.** This last resort is not needed since model adjustments achieved balance.

Considering Changes in Circumstances

Evaluators should account for any events or changes in circumstance that could affect balance between groups. This could be a targeted policy change, abrupt change to the program, or other scenario that disrupts group equivalence. Evaluators ideally should monitor variables that could change over time due to reasons beyond the intervention (e.g., changes in income, exposure to new services) and retest for equivalence, or note when such threats to equivalence cannot be tracked. If disruptions are case by case at the individual level, evaluators should also apply a conservative intent-to-treat approach that analyzes cases in their original groups, regardless of how much or little treatment they receive.

Indicating Balance Is Achieved

Once balance is established, evaluators need to describe the steps taken to achieve balance or equivalence on observed variables and related test results. This includes describing which matching variables, if any, were removed from the propensity score model to attain balance or did not reach the desired level of equivalence. Evaluators should also discuss potential impact on bias. More bias is likely when balance is not achieved for variables that prior research shows are stronger predictors of program participation. Including these descriptions confirms the evaluator’s understanding that matching without testing is insufficient to establish unbiased estimates of program effect.

Analyzing Outcomes

Evaluators who carefully estimate the propensity score using a well-conceived set of matching variables, match cases with need for little or no pruning, and establish balance between the treatment and comparison groups on baseline characteristics are then ready to test the impact of the treatment on the outcome. Following these steps helps ensure these estimates of program impact are trustworthy with minimal bias.

Key Takeaways

Propensity score matching can achieve unbiased estimates of program impact when it results in balance or equivalence between the treatment and comparison groups—i.e., when there is no significant difference between the groups on variables associated with both treatment and outcomes. Trustworthy results rely on a good match between treatment and comparison groups, including careful selection of matching variables. Evaluators should match on a combination of demographics, treatment predictors, confounders, and a baseline measure of the outcome, if available. Evaluators can use programmatic data, staff observations, research literature, and DAGs to identify matching variables.

Evaluators should work to ensure that data are available for all key matching variables. This may include collecting additional primary data, seeking new administrative data sources, or identifying a proxy variable. If data are not available for an important matching variable, the evaluator should note this as a limitation in the evaluation report and discuss its potential impact on estimates.

After estimating the propensity score, evaluators should choose an approach for using the score to match treatment and comparison cases. There are four approaches to using propensity scores for matching: (1) pair matching, (2) weights, (3) stratification, and (4) covariate adjustment.

Once the treatment and comparison groups are identified through matching, evaluators must test for equivalence or balance between the groups. Common forms of testing include standardized bias testing, variance ratio testing, and plotting. Evaluators can adjust the list of matching variables, modify the matching algorithm, or prune cases to achieve balance. Any steps taken to achieve balance or equivalence on observed variables should be described by evaluators to demonstrate their understanding that matching without testing is not enough to establish unbiased estimates of program effect.

Resources

Matching

Matching Methods for Causal Inference: A Review and a Look Forward – Elizabeth A. Stuart

An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies – Peter C. Austin

Propensity Score Analysis: Statistical Methods and Applications – Shenyang Guo and Mark W. Fraser

Directed Acyclic Graphs

An Introduction to Directed Acyclic Graphs – Malcom Barrett

Directed Acyclic Graphs: A Tool for Causal Studies in Paediatrics – Thomas C. Williams, Cathrine C. Bach, Niels B. Matthiesen, Tine B. Henriksen, and Luigi Gagliardi

Calipers

Optimal Caliper Width for Propensity Score Matching of Three Treatment Groups: A Monte Carlo Study – Yongji Wang, Hongwei Cai, Chanjuan Li, Zhiwei Jiang, Ling Wang , Jiugang Song, Jielai Xia

Balance and Pruning

Balance Diagnostics for Comparing the Distribution of Baseline Covariates Between Treatment Groups in Propensity-Score Matched Samples – Peter C. Austin

Visual Pruner: Visually Guided Cohort Selection in Observational Studies – Lauren Samuels and Robert Greevy

References

[1] Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://pdfs.semanticscholar.org/360c/957fe8f823d58e0b0e2e975ff08c503126c2.pdf?_ga=2.28351484.1757894109.1596550670-505513729.1596550670

[2] Mathematics Stack Exchange. (2016). *Distance of a test point from the center of an ellipsoid.* https://math.stackexchange.com/questions/428064/distance-of-a-test-point-from-the-center-of-an-ellipsoid

[3] Ripollone, J. E., Huybrechts, K. F., Rothman, K. J., Ferguson, R. E., & Franklin, J. M. (2020). Evaluating the utility of coarsened exact matching for pharmacoepidemiology using real and simulated claims data. *American Journal of Epidemiology*, *189*(6), 613–622. https://academic.oup.com/aje/article/189/6/613/5679490

[4] Heinrich, C., Maffioli, A., & Vazquez, G. (2010). *A primer for applying propensity-score matching.* Inter-American Development Bank. https://publications.iadb.org/publications/english/document/A-Primer-for-Applying-Propensity-Score-Matching.pdf

[5] Baltar, V. T., Sousa, C. A. D., & Westphal, M. F. (2014). Mahalanobis' distance and propensity score to construct a controlled matched group in a Brazilian study of health promotion and social determinants. *Revista Brasileira de Epidemiologia*, *17*, 668-679. https://www.scielo.br/pdf/rbepid/v17n3/1415-790X-rbepid-17-03-00668.pdf

[6] 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://www.ncbi.nlm.nih.gov/pmc/articles/PMC3144483/

[7] Rosenbaum & Rubin, 1983.

[8] Heinrich et al., 2010.

[9] Ibid.

[10] Ibid.

[11] Foster, E. M. (2010). Causal inference and developmental psychology. *Developmental Psychology, 46*(6), 1454–1480. https://osf.io/76ga4/download

[12] Heinrich et al., 2010.

[13] Mollborn, S. (2016). Young children's developmental ecologies and kindergarten readiness. *Demography*, *53*(6), 1853–1882. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5137473/pdf/nihms831598.pdf

[14] Ibid.

[15] Ibid.

[16] Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly*, *25*(2), 166–176. http://www.xqjyyj.com/xqzzs/upload/100169/upload/file/2017-12-26/1514252465522059558.pdf

[17] Foster, 2010.

[18] Austin, 2011.

[19] Ibid.

[20] 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. https://dash.harvard.edu/bitstream/handle/1/4214880/King_MatchingNonparametric.pdf;sequence=2

[21] Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, *15*(3), 234–249. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2936698/pdf/nihms-192966.pdf

[22] Austin, 2011.

[23] 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*(25), 3083–3107. https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.3697

[24] Austin, 2011.

Appendix

This table provides many of the dimensions and approaches that can be used in a propensity score models. These are related to how the data are analyzed with the intent of identifying the best possible matches between treatment and control. While the table lays these out as separate and competing methods, they are not necessarily mutually exclusive. Evaluators can often combine various methods such as greedy matching (as opposed to optimal matching) with or without calipers. Weighting methods are separated in the table as they tend to be a more distinct approach to propensity score matching. There is no single correct method, with the decision about which method to be used based largely on the data available.

Exhibit A-1. Concepts within Propensity Score Matching

Approach	Brief description	Benefits	Potential issues
Matching			
Greedy pair matching	• Matches cases in treatment and comparison groups based on closest matching propensity scores • Creates a pair with treatment and comparison cases (can be 1:1, 1:many or many:1) • Usually completed with one iteration/pass through the data	• Usually completed as a one-pass method with no constraints (in the form of calipers, which is the maximum allowable distance between propensity scores) • Considered fairly quick and efficient at pulling pairs together, especially for larger datasets • Generates lower variance than use of propensity score weights • Is highly intuitive	• Does not build in a check and recalibration to ensure the matched pairs produce homogeneous populations • Differs from optimal matching by making a match and fixing the match (compared to reconsidering all other matches before making the next match)

Approach	Brief description	Benefits	Potential issues
*K*th nearest neighbor matching	• Used for 1:1 or many:1 matching, where up to a specified number (*K*) of comparison cases are assigned to a treatment case based on close proximity of propensity scores • Is in simplest form when used for 1:1 matching	• Takes advantage of average similarities amongst a pool of "neighbors" • Uses more of the data to create a more complete picture of likely program impact • Generates lower variance than use of propensity score weights • Is highly intuitive	• Requires specification of *K*, though the best value of *K* is unknown • Can cause extreme weights and high variance if the nearest neighbor includes a case with a propensity score that is not close to the treatment case, leading to worse bias than if nothing was done at all
Caliper matching	• Uses a nearest neighbor matching approach that relies on specified caliper widths (the maximum distance allowed for matching between neighbors on a given factor) to determine what *propensity score values* will be matched	• Has been shown to be among the least biased methods • Generates lower variance than use of propensity score weights • Is highly intuitive	• Is discretionary and without clear, universally accepted guidelines • Can cause extreme weights and high variance if extreme values are used, leading to worse bias than if nothing was done at all • Can lead to overpruning if techniques like matching pairs only if they fall within a certain radius of the control (i.e., radius matching) are used to improve the quality of matches

Approach	Brief description	Benefits	Potential issues
Weighting			
Inverse probability of treatment weighting	• Includes the inverse of the propensity score value as a weight on the regression model, similar to survey weights	• Is simple to implement. • Keeps all cases in the analysis	• Can lead to higher variance estimates • Is more susceptible to bias from underlying model assumptions and nonintuitive data representation • Can cause extreme weights and high variance if extreme values are used, leading to worse bias than if nothing was done at all
Kernel optimal matching	• Weights data by estimating the worst-case scenario conditional mean square error of the weighted estimator for all weighting possibilities in the model	• Tends to minimize the model specification biases • Can be used to estimate treatment effects in a number of different ways, including the Kernel Optimal Weighted Average Treatment Effect, shown to be more accurate and less biased than some matching approaches	• Is prone to error and bias when there is significant model misspecification or moderate-to-strong practical positivity violations

Approach	Brief description	Benefits	Potential issues
Covariate adjustment			
Propensity score regression correction	• Includes the propensity scores as a covariate in the final model • Is the simplest method originally proposed by Rosenbaum and Rubin in 1983	• Is simple to implement • Does not typically add additional bias to the data as there are no restrictions on the weighting, no pruning, etc.	• Does not adequately correct for the initial bias between treatment and control, according to some findings • Has been shown to fail when the discriminant—a function that represents class membership—does not have a uniform effect on the propensity score, or when variance is unequal between treatment and control
Stratification			
Propensity score stratification	• Sorts all cases by propensity score and then group cases based on score cutoffs • Calculates treatment effect by comparing treatment and comparison cases within each grouping • Pools results across the groups to estimate an overall treatment effect	• Has been found to reduce approximately 90 percent of the bias related to the matching variables • Uses the entire sample when implement full matching, where each strata consists of at least one treatment and one or more control or the reverse	• Does not reduce bias to the same degree as matching, according to some findings • Requires reducing the sample into a set of subclasses which must be determined

Appendix References

Austin, P. C., & Stuart, E. A. (2015). Optimal full matching for survival outcomes: A method that merits more widespread use. *Statistics in Medicine, 34*(30), 3949–3967. https://doi.org/10.1002/sim.6602

Elzem, M. C., Gregson, J., Baber, G., Williamson, E., Sartori, S., Mehran, R., Nichols, M., Stone, G. W., & Pocock, S. J. (2017). Comparison of propensity score methods and covariate adjustment. *Journal of the American College of Cardiology, 69*(3) 345–357. https://www.onlinejacc.org/content/69/3/345.abstract

Govindasamy, P. (2016). *A comparison between propensity score matching, weighting, and stratification in multiple treatment groups: A simulation study* (Publication No. 1173) [Doctoral dissertation, University of Denver]. https://digitalcommons.du.edu/etd/1173

Myers, J. A., & Louis, T. A. (2007). *Optimal propensity score stratification*. (Johns Hopkins University Dept. of Biostatistics Working Papers No. 155). https://core.ac.uk/reader/61318500

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://www.semanticscholar.org/paper/The-central-role-of-the-propensity-score-in-studies-Rosenbaum-Rubin/f0b25b16bdcb7b6418e284255b9e2ba32a7585d4?p2df

Schneider, B., & McDonald, S. K. (2010). Methods for approximating random assignment. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (pp. 97–103). DOI: 10.1016/B978-0-08-044894-7.01689-4

Submitted to:
Pooja Gupta Curtin, Project Officer
Office of Planning, Research, and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services
www.acf.hhs.gov/opre

Suggested citation: McCombs-Thornton, K., & Poes, M. (2021). *Measuring program effects in home visiting evaluation: Improving estimates with propensity score matching.* OPRE Report #2021-55. Office of Planning, Research, and Evaluation, U.S. Department of Health and Human Services.

This resource was developed by James Bell Associates under Contract No. HHSP233201500133I. It does not necessarily reflect the views or policies of the Office of Planning, Research, and Evaluation; the Administration for Children and Families; the Health Resources and Services Administration; or the U.S. Department of Health and Human Services. For more information, see https://www.jbassoc.com/project/design-options-home-visiting-evaluation-dohve/.