The Trouble with Coarsened Exact Matching

Bernard Black
Northwestern University, Pritzker School of Law, Kellogg School of Management, Buehler Center for Health Policy and Economics, Institute for Policy Research

Joshua Y. Lerner
NORC at the University of Chicago, Department of Methodology and Quantitative Social Sciences

Northwestern University Law School

Law and Economics Research Paper 20-09

(draft January 2026)

The data and Stata and R code to generate all results in this project are available at:

[*url to come]

The Trouble with Coarsened Exact Matching

Bernard Black and Joshua Y. Lerner*

Abstract: Balancing methods, which use matching or reweighting to improve the balance between treated and control units, are central tools for causal inference in the social sciences using cross-sectional observational data. We focus here on one method which has attained substantial popularity, especially in political science, Coarsened Exact Matching (CEM) (Iacus, King, and Porro 2012). We report evidence that CEM performs worse than a number of other popular balancing methods (inverse propensity weights, entropy balancing weights, covariate-balancing-propensity score weights, propensity-score matching, and nearest neighbor matching) and explain why it does so. We report evidence both from simulations, and from replicating four recent studies that use CEM. Applied to real datasets, CEM drops substantially more observations than other methods; has larger standard errors; and produces average treatment effect estimates far from both other methods and from CEM itself (applied by subsetting the sample and combining the subset estimates). In simulations based on the real datasets, CEM is sensitive to adding noninformative covariates or varying which covariates one balances on, and can severely over-reject a true null. In simulations with generated data and heterogeneous treatment effects, and thus known truth, CEM has substantial bias and is much less precise than other methods. Our advice: never use CEM as the sole balancing method, and use it, if at all, with sensitivity checks for variable selection, binning choices, and evidence of low treatment heterogeneity. Those checks are rare in current practice.

Keywords: observational studies, balancing, propensity scores, coarsened exact matching, weighting, doubly robust methods

* Black is Nicholas J. Chabraja Professor at Northwestern University, Pritzker School of Law and Kellogg School of Management. Tel. 312-503-2784, email: bblack@northwestern.edu. Lalkiya is at Analysis Group, email: parth.lalkiya@gmail.com. Lerner is Research Methodologist and Data Scientist at NORC at the University of Chicago, email: joshlerner1@gmail.com. Lerner is the corresponding author. We thank Parth Lalkiya for excellent research assistance.

Contents

I. Introduction

Balancing methods, which use matching or reweighting to improve covariate balance between treated and control units, are central methodological tools for causal inference using observational data. Dozens of methods have been proposed. We focus here on one method which is popular in political science, Coarsened Exact Matching (CEM) (Iacus, King, and Porro 2012). Often, balancing methods are used for "preprocessing" a sample to improve balance, and combined with regression analysis. CEM is intended to be used in this way; other balancing methods are often also combined with regression (e.g., Ho et al., 2007; Stuart, 2010).

CEM is a combined coarsening, matching, sample trimming, and reweighting method. When using CEM, the researcher selects a limited number of core variables to balance on. The CEM method divides each continuous variable into bins, requires an exact match between treated and control units on the binned variables, drops unmatched observations, and reweights the remaining observations.

We compare CEM, used to measure the average treatment effect on the treated (ATT), to regression without balancing and to five other well-known balancing approaches, each combined with regression: three reweighting methods (inverse propensity score weighting (IPW); entropy balancing (eBalance, Hainmueller, 2012); and inverse propensity weighting using covariate balance propensity scores (CBPS-weights; Imai and Ratkovic, 2014)), and two matching methods (propensity score matching (PSM), nearest-neighbor matching (nnmatch, Abadie and Imbens, 2011)). We chose well-known methods that appear to perform well from prior research.[1] As a

[1] Busso, DiNardo and McCrary (2014) report good performance for IPW. Zhao and Percival (2017) provide simulation evidence that IPW, eBalance, and CBPS-weights (CBPS used as a reweighting method, which is how we use it here) perform well when used with regression. Chattopadhyay, Hase, and Zubizarreta (2020) report that exact balancing methods, including eBalance and CBPS-weights, perform well relative to IPW. All of these methods are implemented in R in MatchIt or WeightIt (Greifer and Stuart 2021). All are also available in Stata (CBPS is available in Stata through psweight.ado).

basis for comparison, we use both simulated data (with known truth is known) and real data from the four papers using CEM published in the *American Journal of Political Science* (AJPS) over 2012-2016, obtained from a colleague: Black and Owens (2016); Mason (2015); Urban and Niebler (2012); Carpenter et al. (2012).[2] For these four studies, we replicate selected results.

For both the simulated and real data, we compare results across balancing methods, followed by regression on the balanced dataset, and also to ordinary least squares regression without balancing. The three reweighting methods are "doubly robust"; they provide unbiased estimates if either the regression model is correct or the propensity score model is correct.[3] For a dataset with many covariates, CEM users must decide which covariates to balance on and assess the tradeoff between better balance versus loss of sample size as one balances on more covariates. We used the authors' choices of variables to balance on.

We also assessed how CEM is commonly used in political science research by identifying 170 political science papers using CEM over 2010-2020. We chose 100 of these papers at random, and confirmed that the manner in which CEM is used in the four re-examined papers is reasonably typical. For example, the reexamined papers balance on, respectively 7, 10, 6, and 4 covariates; versus a mean of 7.3 across all 100 papers.

For the simulated data, we assessed how far the coefficient estimates are from known truth. For the re-examined papers, we assessed how far coefficient estimates from each method are from the means from the other methods, as a surrogate for the unknown truth.

---

[2] We excluded Broockman (2013), who uses CEM in an idiosyncratic manner. We later conducted our own search of *AJPS* and found several additional papers using CEM during this time period. See Appendix for a full list.

[3] See Bang and Robins (2005) for IPW, Zhao and Percival (2017) for eBalance; Sloczynski, Uysal and Wooldridge (2025) for CBPS-wts. Sloczynski, Uysal and Wooldridge (2025) also show that eBalance, with balancing only on the mean, logit used to estimate the propensity score, and exact balance imposed (the eBalance default option allows limited departure from exact balance), is equivalent to CBPS-wts, and also to inverse propensity tilting (Graham, Pinto and Egel, 2012). We show below that, using the eBalance defaults, there are modest differences between eBalance and CBPS-wts, with eBalance producing somewhat better covariate balance than CBPS.

We also assessed effective sample size and precision. CEM, unlike the comparison methods, drops treated units for which matching controls cannot be found. This reduces sample size; the loss increases rapidly with the number of covariates used for balancing. Thus, CEM suffers strongly from the curse of dimensionality. In contrast, the comparison methods preserve all treated units.

A core question for any balancing method is how well it can handle heterogeneous treatment effects. For the simulated data, we studied three types of treatment effects: a base case of homogeneous treatment effects; treatment effect linear in the true propensity score, and treatment effect quadratic in the true propensity score.

Using CEM, Black-Owens find support for their conjecture that federal appellate judges who are contenders for Supreme Court vacancies write more dissents when a vacancy exists. In contrast, regression and all comparison methods produce coefficients with the opposite (negative) sign and 95% confidence intervals (CIs) that cross zero. We also use a subclassification approach, where we divide the sample into propensity-score deciles, and estimate ATT within deciles. The CEM within-decile estimates are below the full-sample CEM estimate and closer to the estimates from other methods.

For Mason's Thermometer Bias and Like Bias outcomes, both regression and all other methods support her hypothesis that partisan-ideological sorting (the tendency of ideology to align with political party) increases the intensity of partisan preferences. In contrast, the CEM coefficients are small and CIs include zero. All within-decile estimates are higher that the full-sample CEM estimate, and are closer to the estimates from other methods. While we do not know truth, that CEM provides outlier estimates, has much lower precision, and generates very different within-decile versus full-sample estimates, counsels against relying on CEM.

To use CEM, researchers need to balance on a limited number of covariates, to reduce loss of sample size. We find that CEM coefficient estimates are highly sensitive to which covariates are used for balancing. For both Black-Owens and Mason, as we increased the number of covariates used for balancing, the CEM estimates for ATT increasingly diverged from those from the other methods. The CEM estimates are also sensitive to adding random covariates (uncorrelated with the outcome or treatment assignment) and balancing on both the actual and added covariates. In contrast, the comparison methods are either unaffected or minimally affected by adding random covariates.

We also compared CEM to other methods using generated data, where truth is known. We varied the overlap between treatment and control groups, and imposed different forms of treatment effect heterogeneity (none, treatment effect linear in the true propensity score, and quadratic in the true propensity score. In the models with treatment heterogeneity, CEM results were highly biased (sometimes more than regression alone). In contrast, the reweighting methods and PSM produced unbiased or nearly unbiased estimates; nnmatch had small bias.

CEM is also much more weakly consistent than other methods. The bias from our generated data shrinks as the sample size increases, but much more slowly than with other methods. This, without more, should be fatal for use of CEM in many practical applications.[4]

Given this array of issues, our advice is to never use CEM as the sole balancing method. If CEM is used, it should be accompanied by sensitivity checks for the issues we highlight. To be sure, we cannot rule out the potential for CEM to outperform the comparison methods for datasets or data structures that we did not study.

[4] See Wooldridge ((2025), § 5.1 (Although not all useful estimators are unbiased, virtually all economists agree that **consistency** is a minimal requirement for an estimator. [Nobel Prize winner] Clive Granger once remarked: 'If you can't get it right as *n* goes to infinity, you shouldn't be in this business.")

Although not a focus of this project, we also find evidence that the reweighting methods generally outperform the matching methods (PSM and nnmatch).[5] The differences between results with different reweighting methods are small.

This project emerged from a broader project using papers drawn principally from the AJPS, in which we are studying the performance of different balancing approaches, applied to real-world datasets. We chose CEM as one comparison method. We did not expect to find dramatic differences between CEM and other methods. We decided to study CEM separately after observing the stark differences between CEM and other methods.

Prior research evaluating CEM performance is limited. Ripollone et al. (2019) and Wan (2025) discuss CEM's loss of sample size as the number of balancing variables increases; Wan also notes CEM model dependence due to need to either limit the number of balancing variables or increase coarsening.

This paper proceeds as follows. Part II provides background: a summary of CEM and our other balancing methods and our evaluation methods, and an overview of the papers we replicate. Part III summarizes the use of CEM across the 100 reviewed papers. Part IV provides an overview of CEM and the other balancing methods. Part V assesses CEM's performance versus other methods for Black-Owens and Mason; Part VI discusses the other reviewed papers. Part VII discusses the main takeaways from our analysis.

II. Background

We provide in the Appendix an overview of each balancing method. We provide a more summary treatment here, focusing on CEM.

[5] A technical note: The CEM native code mishandles binary variables which are not always placed into the bottom or top bins. There are similar issues for categorical variables. The CEM implementation in MatchIt corrects this problem.

A. Summary of Balancing Methods

We consider here estimation of ATT with cross-sectional data. Assume a treatment D is applied to some "treated" other units, but not to other "control" units; let y_{i1} and y_{i0} be the potential outcomes if unit *i* is treated or control, respectively, and $\tau i = y_{i1} - y_{i0}$. The fundamental problem if causal inference is that we observe only one of the two potential outcomes and must impute the missing one. All balancing methods share a common goal: to impute the missing potential outcomes in a way that achieves conditional independence -- assignment to treatment that is as-if randomly assigned, conditional on a set of covariates **X**. For studying the ATT, one needs to impute the missing control outcomes $y_{i0}|D=1$). Conditional independence means that the potential outcomes are independent of treatment assignment, conditioned on the covariates: $((yi0, yi1) \perp\!\!\!\perp D|\mathbf{X})$. If conditional independence holds, all methods should converge to truth in asymptopia, so comparative assessment involvers finite sample performance.

B. Summary of Coarsened Exact Matching

We summarize CEM here and provide technical details in the Appendix. CEM imposes exact matching on a limited set of user-chosen covariates (perhaps drawn from a larger set of available covariates), followed by regression on the matched dataset. CEM divides each selected covariate into bins (the "coarsening" part), requires an exact match between a treated unit and one or more control units on the binned variables, and drops unmatched observations. For a covariate *x* (either continuous or discrete) and a sample of size n. CEM divides the domain of x into $b(n) = \log_2(n) + 1$ bins (rounding up). The number of bins is often substantial; for example, n = 350 leads to 10 bins for each variable. CEM lets researchers choose a different binning structure, but we use the CEM default here, as do most users (see Part IV). Each retained treated unit gets weight

= 1.[6] The control units get varying weights, which sum to the number of retained control observations. Let S denote the multidimensional space which contains the binned variables, s index subspaces that contain at least one treated and at least one control unit, M_T and M_C equal the number of retained treated and control units, respectively, and m^s_T and m^s_C be the number of treated and control units in subspace s. The control weights in subspace s equal the fraction of treated observations divided by the fraction of control observations in this subspace. The weight on a unit *i* in subspace s:[7]

$$w_{i,C} = \frac{m_T^s}{M_T} * \frac{M_C}{m_C^s}$$

These weights are, in effect, inverse propensity weights measured within each subspace. They are used in regression on the matched sample. Unless otherwise specified, we use the default CEM binning structure, as most users will (we confirm in the Appendix that this is how CEM is typically used in political science). CEM is available for both Stata and R; we obtained the same results with both. CEM can be seen as a hybrid between matching and reweighting methods. Treated units are retained if they can be matched exactly to one or more control units, and vice-versa, but control units are also weighted.[8]

C. The Comparison Balancing Methods

We compare CEM estimates of ATT to regression alone and to five other balancing methods which are commonly used in our experience with code available in Stata, R, or both. We compare CEM both to methods that provide balance only in expectation and methods that aim at

[6] The comparison methods also give a weight of 1 to each treated unit.

[7] Our notation loosely follows Iacus, King, and Porro (2012).

[8] Something in the CEM code assigns binary variables to multiple bins, not only the lowest and highest. Categorical variables have a similar problem. The MatchIt package in R fixes this coding error. We use the MatchIt version of CEM.

exact covariate balance. For PSM, we estimate the propensity score with logistic regression and use 1:1 matching with replacement.[9] Nnmatch can be used either with bias correction but without regression on the balanced sample, or without bias correction followed by regression; we use it with regression, using 1:1 matching with replacement and the default Mahalanobis distance measure, using teffects nnmatch in Stata. For IPW we estimate the propensity score with logistic regression. eBalance provides weights that are similar to IPW but ensure exact balance on covariate means between treated and control groups.[10] The CBPS propensity scores provide close, although not exact balance on covariates. They can be used for matching or reweighting; we use reweighting. We combine each method with regression on the balanced dataset.[11]

D. Overview of Performance Measures

For the simulated data and each paper, we report (in the text or the Appendix) estimates of the average treatment effect on the treated (ATT), using regression alone (either OLS, logit, or negative binomial, following the original paper) and balancing using each method followed by regression. CEM does not specify which s.e.'s to use; we use heteroskedasticity robust s.e.'s. We do not use sample trimming, although trimming can often be good practice (e.g., Crump et al., 2009; King, Lucas, and Nielsen, 2017).

For each paper and each method v, we report a z-score, as a measure of how far each estimate is from the average for the other methods, defined relative to the other methods ($-v$) as:

[9] We used psmatch2.ado in Stata, with the "ties" option, which uses all matches if two or more are equally good. Different matching routines, including matching and MatchIt" for R, will produce somewhat different results.

[10] eBalance can be set to provide balance on higher moments, but we use it to balance only on means. By design, in the reweighted sample, the outcome should be orthogonal to the covariates and thus the treatment effect estimate should be the same with or without regression on the covariates used for balancing. Across our replications, this was often but not always true.

[11] Other methods for achieving exact or near-exact balance in a finite sample include Chattopadhyay, Hase and Zubizarreta's (2020) stable balancing weights, Diamond and Sekhon's (2013) genetic matching, and Graham et al.'s (2012) inverse propensity tilting.

$$z_v = \frac{ATT_v - E_{-v}(ATT)}{E_{-v}(s.e.)}$$

For the simulated data, we measure the z-score relative to the known truth (ATT = 1), as the mean across 1,000 simulations.

$$z_{sim} = \frac{E[\hat{\tau}] - 1}{E[s.e.]}$$

For the matching estimators, we report the number of distinct treated and control units used; some control units are used more than once. We also construct and report a measure of the effective number of control units for the reweighting methods (including CEM), that lets us compare estimators in terms of the effective number of control units actually used. We normalize the weights v_i of control units to sum to the number of treated units n_t. We use the v_i to compute the effective number of control units $n_{c,eff}$ (rounded to the nearest whole number) as follows:

> A control unit with $v_i \geq 1$ is counted once (similar to how one counts control units used multiple times in measuring sample size for matching methods)
>
> A control unit with $v_i < 1$ is counted at v_i units.

Consider IPW as an example. Standard IPW weights on control units are p/(1-p). The normalization factor is $F = \frac{n_t}{\sum_j \frac{p_j}{(1-p_j)}}$, so v_j = F * min[1, $\frac{p}{1-p}$].[12]

To measure covariate balance after balancing, we use the normalized difference between treated and controls for method m and covariate g, defined as (Imbens and Rubin, 2015):

$$ND_{mg} = (\bar{x}^g{}_{it} - \bar{x}^g{}_{jc})/[(s_{tg}^2 + s_{cg}^2)/2]^{1/2}$$

Here s_{tg} and s_{cg} are the standard deviations of the treated and control observations. We then

[12] In the Appendix, we obtain similar results for effective sample size using Kish's (1965) measure, which was developed for survey sampling with survey weights, adapted for matching methods, which Kish does not address.

compute the means of the absolute values of the ND's across all covariates $\overline{|ND_m|}$.

E. Comparison Papers

We summarize here the four papers using CEM that we reassess. We discuss Black and Owens and Mason in the text and the other two papers in the Appendix.

1. Black-Owens (2016)

Black-Owens assess whether federal appellate judges who are plausible candidates for the U.S. Supreme Court change their voting behavior to curry favor with the President, at times when Supreme Court vacancies exist. CEM plus logistic regression is their sole method. We study the measures for which they report evidence of a change in voting: is a candidate judge more likely to write a dissent (Judge Writes Dissent); to write a pro-US decision (Pro-US Opinion), or to support the President's position (Decision Consistent with President's Ideology). They hypothesize that candidate judges write more dissents and make decisions closer to the President's political preferences. Their sample is 11,787 decisions by panels with at least one candidate judge over 1946 to 2010.

2. Mason (2015)

Mason studies voter polarization. She hypothesizes that polarization will increase if voters are "sorted" -- hold party identification consistent with their ideological views. She uses regression and CEM without regression to study whether sorting, controlling for the strength of party identity, predicts four partisanship measures, termed thermometer bias, like bias, activism, and anger (her Figure 5). She finds strong support with regression but weak support with CEM. Her sample is 9,858 survey respondents, whom she considers to be treated if a sorting measure is roughly above the sample median.

3. Urban and Niebler (2014)

Urban and Niebler study the effect on campaign contributions of "spillover" Presidential campaign ads, which reach residents in noncontested states who live in the same TV-reception area as residents of a neighboring contested state. Their primary balancing method is PSM; they also use genetic matching (Sekhon, 2009) and CEM in robustness checks, without regression. In Black and Lerner (2022), we show that Urban-Niebler misestimated propensity scores, correct this error, apply a zero-inflated negative-binomial model, and find that spillover ads predict higher contribution amounts, but not higher likelihood of contributing.

4. Carpenter et al. (2012)

Carpenter and coauthors examine whether Federal Drug Administration (FDA) drug approvals, issued close to a time deadline for FDA action, are more likely to lead to involve drugs that later turn out to have important side effects. They use CEM and optmatch (Hansen, 2004) in robustness checks. We study the outcomes (black box warnings, safety-based withdrawals, safety alerts) for which they report statistical significance for their base model.

F. Simulated Data

We used a synthetic sample size of N = 5,000 (half treated, half control) and 6 covariates indexed by j. This choice was intended to model a moderate-sized real-world dataset, with both a large enough sample and few enough covariates, so that the effects of the curse of dimensionality on CEM would be manageable. We used the six covariates both to define the propensity to be treated and the heterogeneous treatment effect. The covariates were:

$$cov_{\{i1\}} \sim N(0,1),$$
$$cov_{\{i2\}} \sim N(0,1)$$
$$cov_{\{i3\}} \sim N(0,1)$$
$$cov_{\{i4\}} \sim Unif(0,1)$$
$$cov_{\{i5\}} \sim Unif(0,1)$$

$cov_{\{i6\}} \sim N(0,1)$

We use the covariates to generate a true propensity score for each unit *i* using a logistic model, either with or without additional noise in treatment assignment, defined by σ (we used values of 0 or 0.6). Define:

$$\eta_i = \alpha + \sum_{j=1}^{6} \beta_j \, cov_{ij} + u_i, \quad u_i \sim N(0, \sigma^2)$$

The "true" propensity score is:

$$p_i^{\{true\}} = \frac{1}{\{1 + e^{\{-\eta_i\}}\}}$$

Treatment is assigned as a draw from a Bernoulli distribution with probability $p_i^{\{true\}}$:

$$treat_i \sim Bernoulli(p_i^{true})$$

We fix $(\boldsymbol{\beta_1}, \ldots, \boldsymbol{\beta_5}) = (\boldsymbol{0.5}, \boldsymbol{0.3}, \boldsymbol{-0.4}, \boldsymbol{0.2}, \boldsymbol{-0.2})$ and we let β_6 vary. Larger β_6 induces more extreme values of the true propensity score, and therefore weaker overlap between the propensity score distributions for treated and control units. We report results for a high overlap case ($\beta_6 = 0.2$) and a low overlap case ($\beta_6 = 1.2$).

For methods that are based on the estimated propensity score (PSM and IPW), we fit a standard logistic regression of treatment on the covariates (excluding the unobserved noise term u_i). Let $\widehat{\boldsymbol{\beta}}$ denote the fitted coefficients. The estimated propensity score is:

$$\widehat{p_i} = \frac{1}{1 + \mathrm{e}^{\left(-\sum_{j=1}^{6} \widehat{\beta_j} \, cov_{ij}\right)}}$$

We generate outcomes using a baseline outcome surface (for both control and treated units), using the same β values used to estimate the propensity score, plus an additive treatment effect for the treated units:

$$y_i = \mu_{0i} + \tau_i \, treat_i + \varepsilon_i, \quad \varepsilon_i \sim N(0,1)$$

$$\mu_{0i} = \frac{1}{2}\sum_{j=1}^{6} \beta_j \; cov_{ij}$$

The treatment effect is defined in three ways; homogeneous, linear in the true propensity score, and quadratic in the true propensity score. The outcome for each is defined as:

Homogenous: $\tau_i^{raw} = 1$
Linear: $\tau_i^{raw} = p_i^{true}$
Quadratic: $\tau_i^{raw} = 2(p_i^{true})^2$

To make results comparable across heterogeneity and overlap/noise conditions, we normalize treatment effects within each design condition (defined by β_6 and σ) so that ATT = 1. Concretely, within each condition we rescale τ_i^{raw} so that:

$$E[\,\tau_i \mid treat_i = 1\,] = 1$$

In practice, this is implemented by setting:

$$\tau_i = \tau_i^{raw} \cdot \frac{1}{E[\,\tau^{raw} \mid treat = 1\,]},$$

For each model, we run 1,000 simulations, record the measured treatment effects, robust s.e.'s, and *t*-statistics for each draw, and compute $E[\widehat{\tau}]$, $E[\widehat{s.e.}]$, and $E[\widehat{t}]$.

We view the high-noise case as more realistic in a real-world setting where the propensity score may depend on both observed and unobserved covariates; below, we often present results limited to this case.

III. Use of CEM in Published Political Science Papers

We explored typical use of CEM in published political science papers over 2010-2020, by identifying 170 papers using CEM across 47 political science journals and reviewing 100, chosen

at random. See Appendix for the search strategy, and Appendix Table App-1 for summary statistics on the reviewed papers.

A slight majority of papers (58/100) balance on fewer than all covariates, presumably to preserve sample size. Of these, 19 report results for alternative choices of which variables to balance. However, it is rare for authors to explain why they balance on only some covariates or how they decided which covariates to balance on. Most papers (69/100) use the CEM defaults to bin continuous variables. The other 31 use fewer bins, often only 2-5, likely to preserve sample size. Only 5 papers explain the binning choices.

IV. Overview of Differences Between CEM and Other Methods

This part provides an overview of the performance of each method in proving covariate balance; preserving sample size; precision; and whether results vary between CEM and other methods.

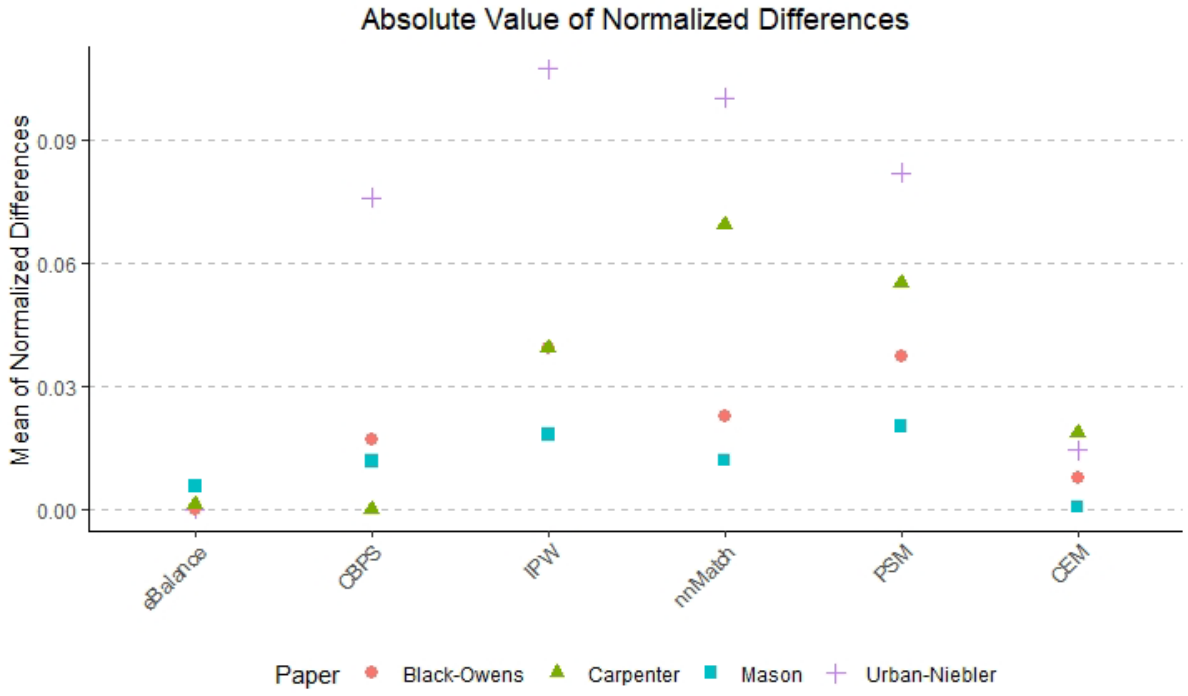
A. Covariate Balance

In Figure 1, we plot the mean of the normalized differences (n.d.'s) for each method. The methods are arrayed along the x-axis with the weighting methods first, then other matching methods, then CEM. The average for each paper is shown as a data point in the "column" for each method. For the simulations, we use both low- and high-overlap scenarios, with high noise in estimating the propensity score. The average (across papers and simulations) of the mean of the n.d.'s (across covariates) is shown in a small table underneath each panel.

For eBalance, covariate balance is exact or very close across comparison papers, with both measures. CEM also creates excellent balance for the observations it retains, with all ND means less than 0.03. The other methods show greater variation across papers.

Figure 1: Covariate Balance Across Methods

Means (across covariates) of |normalized difference (n.d.)| between treated and control units, for each balancing method and each model. "Sim-high" and "sim-low" are mean n.d.'s for simulated data with high and low overlap between treated and control units. Small table below figure shows mean of |n.d.|'s for OLS (no balancing) and each method.



Mean of means	OLS	eBalance	CBPS	IPW	nnmatch	PSM	CEM
NDs	0.265	0.002	0.026	0.051	0.051	0.049	0.011

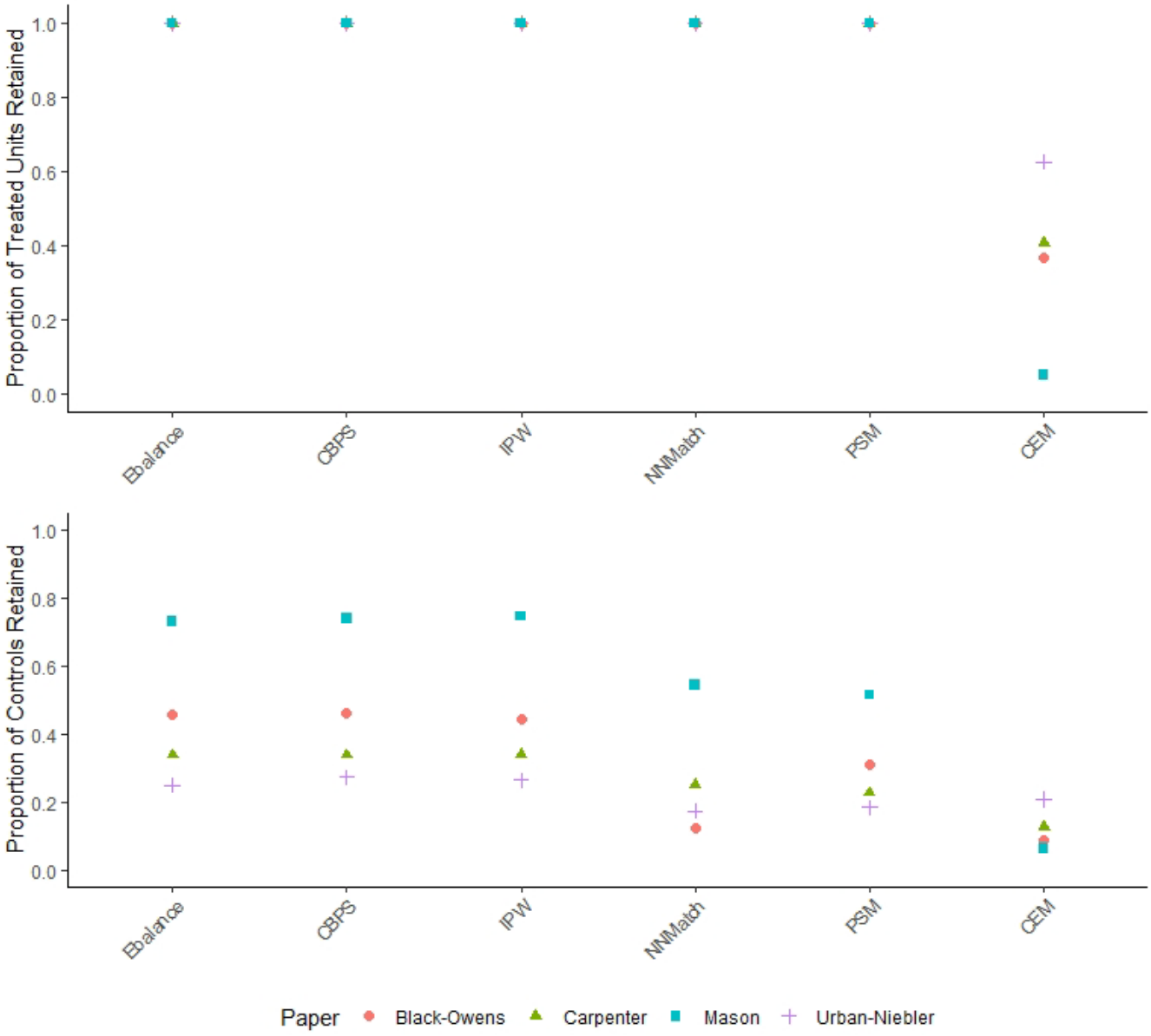
B. Effect of Different Balancing Methods on Sample Size

The balance achieved by CEM comes at a cost in sample size. Figure 2 shows the fraction of treated units (top panel) and the fraction of control units (bottom panel) retained by each method. The methods are again arrayed along the x-axis; the y-axis shows the fraction of units retained. As the top panel illustrates, all methods except CEM retain all treated units. The fraction of treated units retained by CEM ranges from 61% for Urban-Niebler to only 5% for Mason. The large sample loss for Mason reflects the curse of dimensionality: Mason balances on 10 variables, of which 3 are continuous. CEM also retains fewer effective control units than other methods (bottom panel). This loss of control units is driven mainly by CEM's loss of treated units.

In the Appendix, we simulate CEM's loss of sample size for continuous covariates, as a function of sample size, using the CEM defaults for number of bins. For covariates drawn from a uniform distribution, loss of sample is large for 4 covariates, even with a sample of 10,000 (half treated, half control), and nearly complete for 5 covariates. For covariates drawn from a normal distribution, loss of sample is large for 5 covariates, even with a sample of 10,000, and nearly complete for 7 covariates.

Figure 2: Percent of Treated and Control Units Retained

Proportion of treated units (**top panel**) and control units (**bottom panel**) retained after balancing. "Sim-high" and "sim-low" are means for simulated data with high and low overlap between treated and control units.



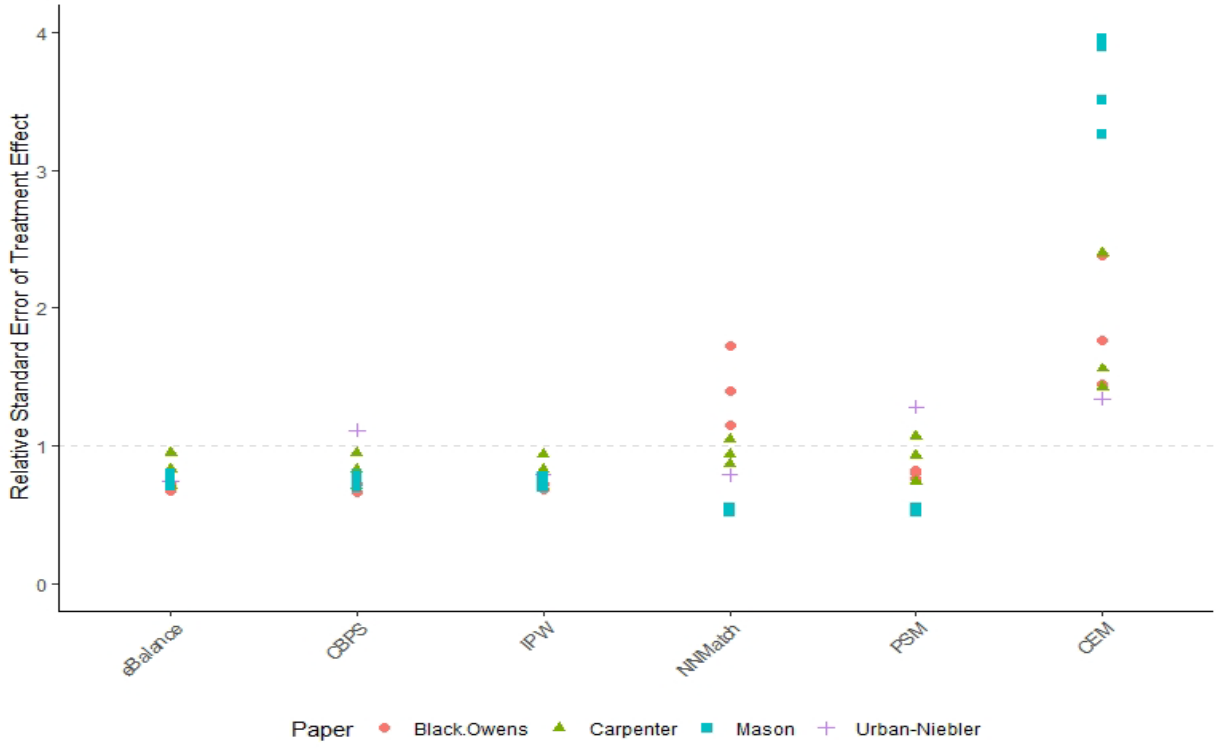
C. Relative Precision

The smaller samples retained by CEM reduces precision (see also Ripollone et al., 2020). We illustrate this in Figure 3. The figure shows the ratio of the robust s.e. using a particular method to the average for the other methods. For papers for which we examine more than one outcome (Black-Owens, Carpenter, Mason), the figure shows one data point for each outcome. For the

simulated data, we jitter the data points so they appear to the right of those for the re-examined papers. A small table underneath the figure shows the mean of the relative s.e.'s for each method. CEM has much larger s.e.'s than any other method.

Figure 3: Relative Precision

Comparison of s.e.'s across methods. Y-axis shows ratio of robust s.e. using indicated method to average for other methods. For papers with multiple outcomes and for simulated data, graph shows one data point for each outcome. Simulated data results are jittered to appear slightly to the right of other results. Small table below figure shows mean of relative s.e.'s for each method.



	eBalance	CBPS	IPW	nnmatch	PSM	CEM
Mean (4 papers)	0.821	0.835	0.836	0.891	0.779	2.22

Figures 2 and 3 illustrate a tradeoff for CEM, not present for the other methods: One must either limit the number of matching variables and accept imbalance on other variables, or use more matching variables, leading to smaller samples and larger s.e.'s.

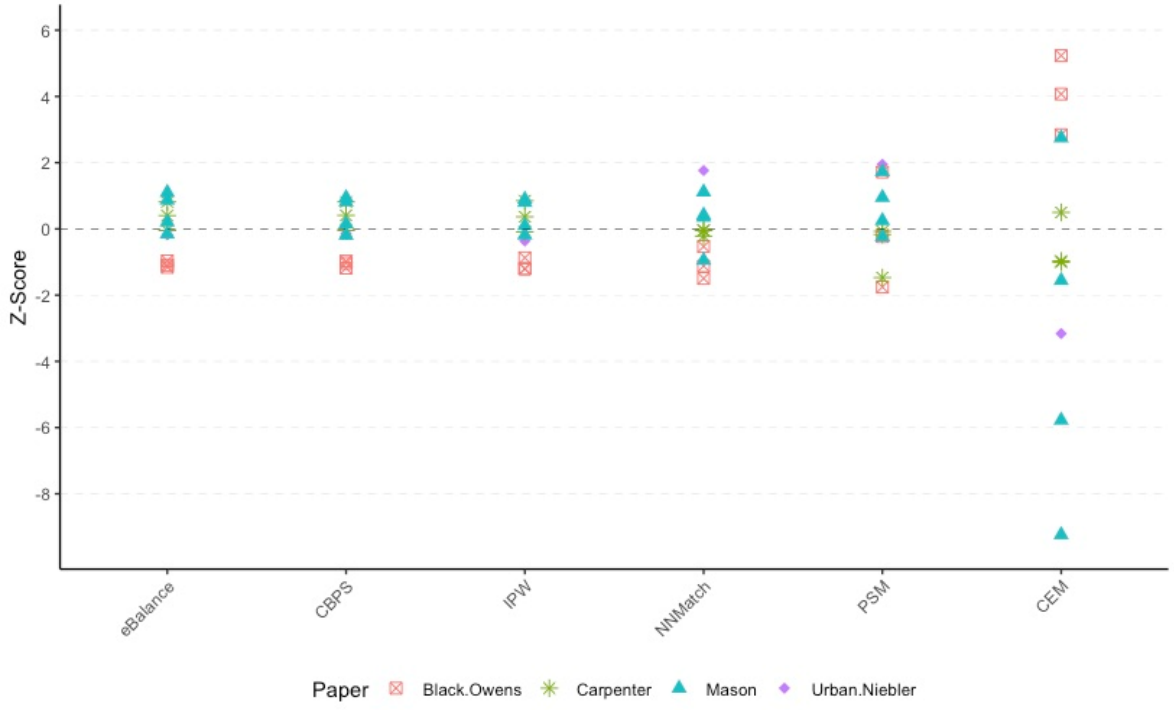
D. Point Estimates: CEM Versus Other Methods

In Figure 4, we plot the z-scores for the treatment effect estimates from each method. We do not know truth, but there is reason for concern about the accuracy of an estimate that is far from the others. The format of Figure 4 is similar to Figure 3. Each column shows the z-scores for a particular method, with one data point for each outcome. A small table underneath the figure shows the mean of the absolute values of the z-scores for each method.

CEM has much higher z-scores than the other methods. Many CEM estimates are outside the CIs from the other methods; with z-scores ranging from -9.23 to +5.19.

Figure 4: Z-Scores for Treatment Effects

Comparison of z-scores for treatment effect estimates by method. For papers with multiple outcomes and for simulated data, graph shows one data point for each outcome. Simulated data results are jittered to appear slightly to the right of other results. Small table below figure shows mean of relative s.e.'s for each method.



	eBalance	CBPS	IPW	nnmatch	PSM	CEM
Mean (|z|) (4 papers)	0.636	0.610	0.631	0.732	0.958	3.37
Mean (simulations)						

V. Analysis of Black- Owens, Mason, and Simulated Data

In this part, we compare CEM to the other balancing methods, for the Black-Owens and Mason datasets, and then for the simulated data.

A. Black-Owens

Table 1 presents Black-Owens results using logistic regression alone; comparison methods plus regression, and CEM plus regression. Column (7) replicates the Black-Owens results. In Panel A (Judge Writes Dissent) (Panel A). The CEM coefficient is 0.907 and far from the other estimates (z=5.19). All other methods provide *negative*, coefficients with CIs that cross zero. The CEM results for the other two outcomes are also outliers, with large z-scores (3.74, 2.84).

Table 1. Black-Owens Results with Different Balancing Methods

Last column replicates Black-Owens (CEM plus logistic regression). Coefficients on covariates are suppressed. Robust s.e.'s in parentheses.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Balancing Method	**None**	**PSM**	**nnmatch**	**IPW**	**eBalance**	**CBPS-wts**	**CEM**
Panel A. Judge Writes Dissent							
Vacancy Exists	-0.155	-0.167	-0.062	-0.042	-0.056	-0.057	0.907
s.e.	(0.154)	(0.186)	(0.255)	(0.168)	(0.166)	(0.165)	(0.115)
z-score	--	-1.75	-1.12	-0.87	-0.97	-0.97	5.23
treated (control)	4901(6886)	4901(2147)	4901(847)	4901 (6886)	4901 (6886)	4901 (6886)	1803 (687)
effective controls		1244	690	3051	3146	3191	611
Panel B. Judge Writes Pro-US Decision							
Vacancy Exists	0.668	0.679	0.652	0.574	0.579	0.576	1.161
s.e.	(0.104)	(0.131)	(0.184)	(0.120)	(0.120)	(0.119)	(0.091)
z-score	--	-0.23	-0.53	-1.20	-1.16	-1.18	4.07
treated (control)	4901(6886)	4901(2147)	4901(847)	4901 (6886)	4901 (6886)	4901 (6886)	1803 (687)
effective controls		1244	690	3051	3146	3191	611
Panel C. Judge's Decision Consistent w President's Ideology							
Vacancy Exists	0.291	0.283	0.126	0.131	0.137	0.140	0.350
s.e.	(0.048)	(0.063)	(0.097)	(0.056)	(0.056)	(0.056)	(0.044)
z-score	--	1.72	-1.49	-1.21	-1.09	-1.03	2.84
treated (control)	4901(6886)	4901(2147)	4901(847)	4901 (6886)	4901 (6886)	4901 (6886)	1803 (687)
effective controls		1244	690	3051	3146	3191	611

B. Mason

Table 2 provides a similar comparison, for Mason. All other methods support Mason's hypothesis. The CEM estimates are far below other methods for thermometer bias and like bias

(z=-5.77, -9.23), somewhat lower for activism (z=-1.56), but higher for anger (z=2.75). CEM is also much less precise. Across the other methods, thermometer bias, like bias, and activism have CIs that cross zero for CEM but are bounded well away from zero with other methods.

Table 2. Mason Results with Different Balancing Methods

Last column provides CEM-plus-OLS-regression results, similar to her figure 5 (which reports CEM results without regression). Sorting dummy = 1 if idcomplexity ≥ 0.15. Coefficients on covariates are suppressed. Robust s.e.'s in parentheses.

	(0)	(1)	(2)	(3)	(4)	(5)	(6)
Balancing method	**none**	**PSM**	**nnmatch**	**IPW**	**eBalance**	**CBPS-wts**	**CEM**
Panel A. Thermometer Bias							
Sorting dummy	0.0563	0.0601	0.0568	0.0590	0.0594	0.0591	0.0309
s.e.	(0.00493)	(0.00407)	(0.00406)	(0.00536)	(0.00540)	(0.00533)	(0.0170)
z-score	--	0.95	0.42	0.80	0.87	0.82	-5.78
Panel B. Like Bias							
Sorting dummy	0.0483	0.0610	0.0573	0.0558	0.0570	0.0561	0.0166
s.e.	(0.00441)	(0.00373)	(0.00375)	(0.00485)	(0.00492)	(0.00486)	(0.0175)
z-score	--	1.73	1.12	0.89	1.10	0.94	-9.24
Panel C. Activism							
Sorting dummy	0.0283	0.0323	0.0329	0.0314	0.0320	0.0317	0.0248
s.e.	(0.00436)	(0.00381)	(0.00374)	(0.00521)	(0.00533)	(0.00523)	(0.0152)
z-score	--	0.25	0.35	0.10	0.21	0.15	-1.56
Panel D. Anger							
Sorting dummy	0.0812	0.0814	0.0717	0.0822	0.0826	0.0821	0.108
s.e.	(0.0103)	(0.00850)	(0.00853)	(0.0112)	(0.0114)	(0.0113)	(0.0397)
z-score	--	-0.24	-0.95	-0.19	-0.16	-0.19	2.75
observations	9,858	7,890	8,009	8,970	9,858	9,858	572
treated (control)	5802 (4056)	5802 (2088)	5802 (2207)	5802 (3168)	5802 (4056)	5802 (4056)	294 (278)
Effective controls				3023	2966	3000	254

VI. Simulations with Known Data Generating Process

Could the CEM estimates reflect truth, despite being far from the other estimates? We investigate that question as follows. In this Part, we use simulated data, with a known data generating process and known truth. In the next part, we develop a number of pathologies to which CEM is subject, that can help to explain why CEM can go wrong.

A. Relative Bias in Simple Settings with Known Truth

We compare the balancing methods using: (i) homogeneous treatment effects; (ii) treatment effects that are linear in the true propensity score, which simulates a simple form of treatment effect heterogeneity; and (iii) treatment effects that are quadratic in the true propensity score. We consider propensity score distributions with high overlap ($\beta_6 = 0.2$) and low overlap (($\beta_6 = 1.2$) between treated and control units in the true propensity score. We also simulate low noise ($\sigma = 0$) and high noise ($\sigma = 0.6$) in estimating the propensity score.[13]

Figure 5 shows the overlap in the estimated propensity score distributions in the high-overlap and low-overlap cases, with high estimation noise. The Appendix provides similar plots in the low-noise case, and also using the true propensity score.[14]

Figure 5. Overlap in Simulated Propensity Score Distributions

Figure shows kernel density plots for estimated propensity score for treated and control units, assuming high noise in estimating the propensity score. **Panel A**. High-overlap case. **Panel B**. Low-overlap case.

Panel A **Panel B**

[13] The simulation model was developed, in the spirit of a pre-analysis plan, before assessing balancing method performance.

[14] A concern with IPW and possibly other reweighting methods is that estimates can be sensitive to observations with very high weights (Kang and Schaefer, 2007; in particular, IPW estimates of ATT can be sensitive to control outcomes with p close to 1. A common response is to trim the sample to avoid very high weights. We assessed whether trimming was important for the simulated data in the low-overlap case, and concluded that was not.

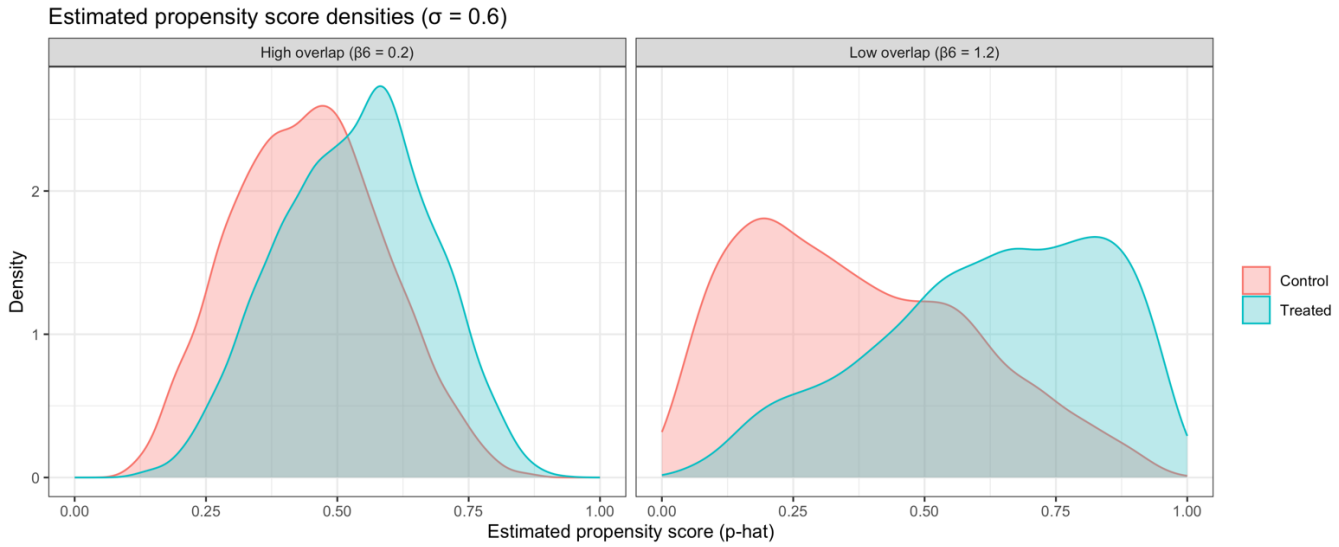


Figure 6 summarizes the simulation results; Appendix Table App-**[*xx]** provides the corresponding numerical estimates. The first "column" shows bias for the simple, uninteresting case of homogeneous treatment effects. In this simple setting, OLS recovers truth; so necessarily do the reweighting methods, which are doubly robust. The matching methods and CEM also recover truth; it is hard to imagine how a balancing method could fail to do so. CEM is much less precise than other methods (compare Figure 3).

The second column of Figure 6 shows bias for the more interesting case of a treatment effect linear in the true propensity score. This is an important practical case, which corresponds to the common situation where persons select into treatment based in part, on expected benefit, so that the treatment effect is positively correlated with treatment probability. OLS can be biased in this setting, and indeed is strongly biased in our simulations. The magnitude of the OLS bias is around 0.08 with high overlap, and much larger, around -0.20, with low overlap; the OLS bias is slightly smaller with high versus low propensity score estimation noise.

The core task for a balancing method can be seen as correcting potential bias from regression alone. The reweighting methods and PSM succeed in eliminate meaningful bias.

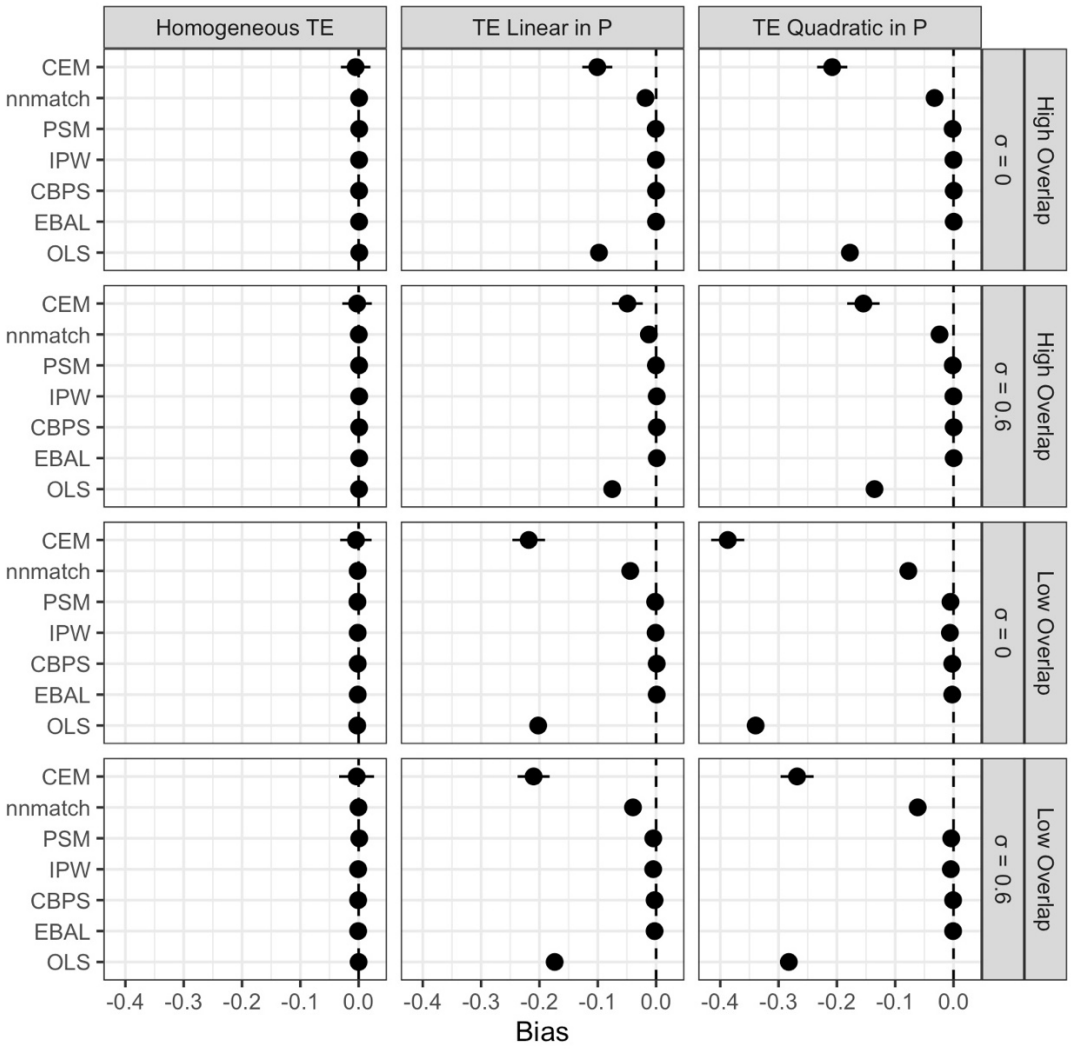
Nnmatch removes around **[*xx%]** of the OLS bias with high overlap, but somewhat less, around **[*75%]**, with low overlap. CEM performs terribly. With low overlap, CEM *increases* the OLS bias; with high overlap and low noise, OLS and CEM bias is similar. Only with high overlap and high noise does CEM modestly reduce the OLS bias, by around **[*xx%].**

The third column of Figure 6 shows bias for a treatment effect quadratic in the true propensity score. OLS bias increases in magnitude, and is especially large in the low overlap case , at 0.34/0.29 with low/high noise. The reweighting methods and PSM again eliminate meaningful bias. Nnmatch removes around **[*xx%]** of the OLS bias with high overlap, but somewhat less, around **[*yy%]**, with low overlap. CEM again performs terribly; with higher bias than OLS in three of the four scenarios, and slightly lower, but still very large bias with low overlap and high noise.

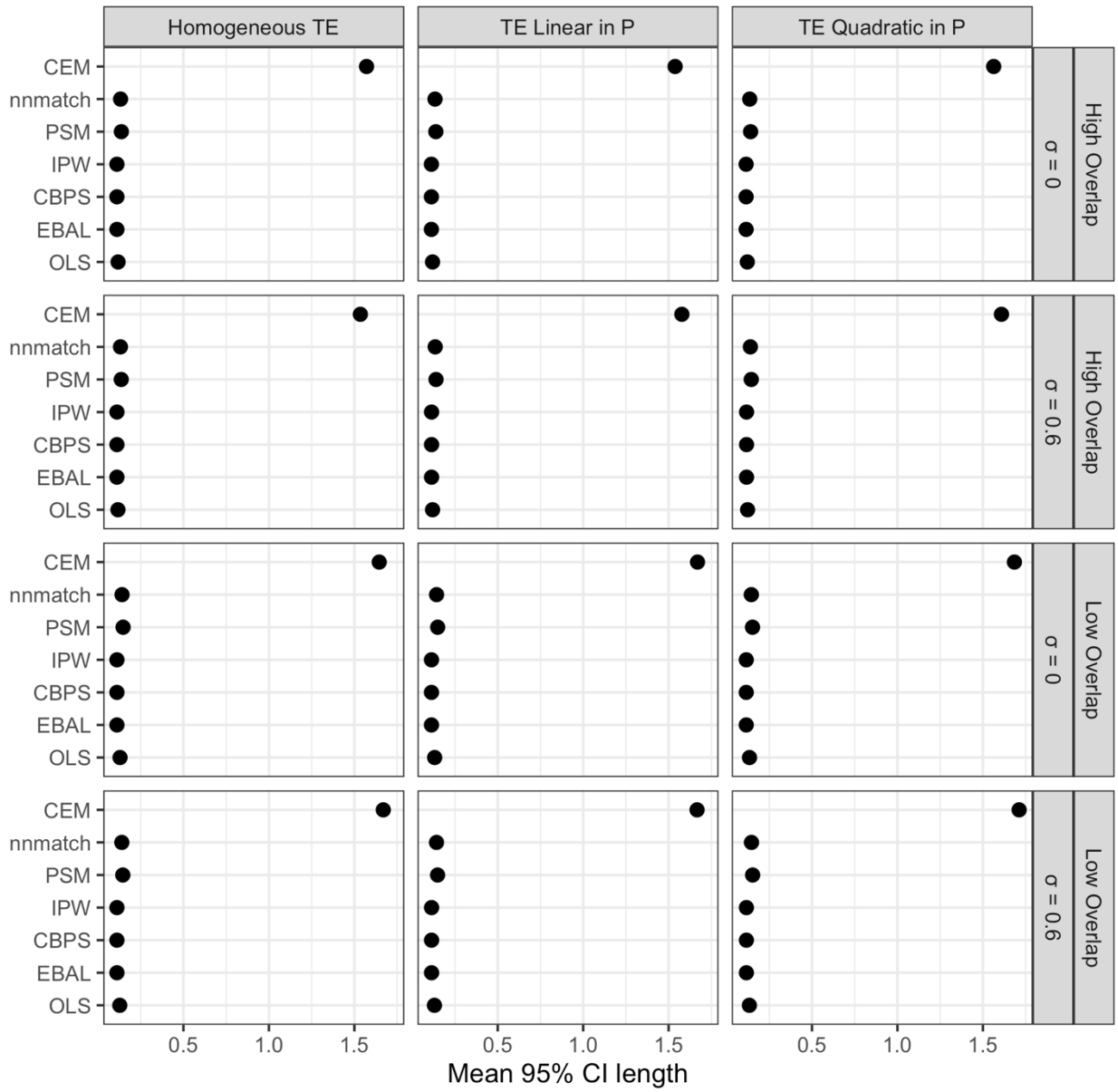
One can, of course, imagine many variations on the simulation setup. For example, nnmatch might perform better if the treatment effect heterogeneity was less closely tied to the propensity score. But the simulation is sufficient to show that CEM can perform badly in a setting when truth is known. When CEM provides outlier results (as in the Black-Owens and Mason studies), relative to other methods, CEM could well be wrong.

Figure 6. Bias with Linear and Quadratic Treatment Effect Heterogeneity

Figure presents estimated bias (relative to know truth of ATT = 1). Point estimates (circles) show mean bias across 1,000 simulations; horizontal lines show 95% CIs (based on mean s.e.'s across simulations). Models presented are (i) homogenous treatment effect (TE) (=1 for all units), TE linear, and TE quadratic in true propensity score. Models are computed for high- and low-overlap between propensity score distributions, and for low and high noise in estimating the propensity score.



Extra Figures – using CIs as informative



B. Understanding the Sources of CEM Bias

We can use the simulation to understand both why CEM is biased, and the sign of the bias. Consider treatment effects linear in p. In a typical sample, one will have sparse treated units for low p, relative to controls, and sparse controls for high p, relative to treated units. The ATT estimate is driven by high-p observations. CEM will differentially drop those observations, because it will often fail to find a control unit within an exact-matching cell. This will produce

negative bias, relative to the true ATT. This effect will be stronger with low versus high overlap, and stronger with treatment effects quadratic versus linear in p.

Bias can also arise for low p values, if most propensity scores are not close to 0 or 1 (as in our simulation, in the high-overlap case). There will be relatively few low-p control observations; CEM will differentially drop treated observations, dur to failure to find a matching control.

We confirm this intuition in Figure 7. We subclassify the sample into propensity score ventiles (5-percentiles), each with an equal number of units (treated + control). The figure shows the fraction of treated units retained in each ventile, for the four scenarios in Figure 6: high-versus-low overlap, and high-versus-no estimation noise. **[*discuss results]**

Figure 7. Treated Units Retained in Different Scenarios

Figure shows fraction of treated units retained for ventiles of the propensity score for the simulated data, in four scenarios: (high-overlap, low estimation noise), (high-overlap, high noise), (low-overlap, low noise), and (low-overlap, high noise).

[*Figure to come]

The tendency for control units to be scarce for high p-values is a central feature of the propensity score. So, therefore, is the CEM tendency to drop high-p than low-p treated units. This will create bias whenever the treatment effect is correlated with p; the stronger the correlation, the stronger the bias. The bias will be negative if the is positive, as in our simulations; and positive if the correlation is negative. The other balancing methods retain all treated units, so avoid this issue.

The other balancing methods are known to be consistent. CEM is not consistent; our simulation provides a counter-example.

VII. CEM Pathologies

We explore in this part a number of features of CEM, some surprising, that can together help to explain why CEM produces results so at variance with other methods and, in the simulations, with known truth.

A. Decile Analyses

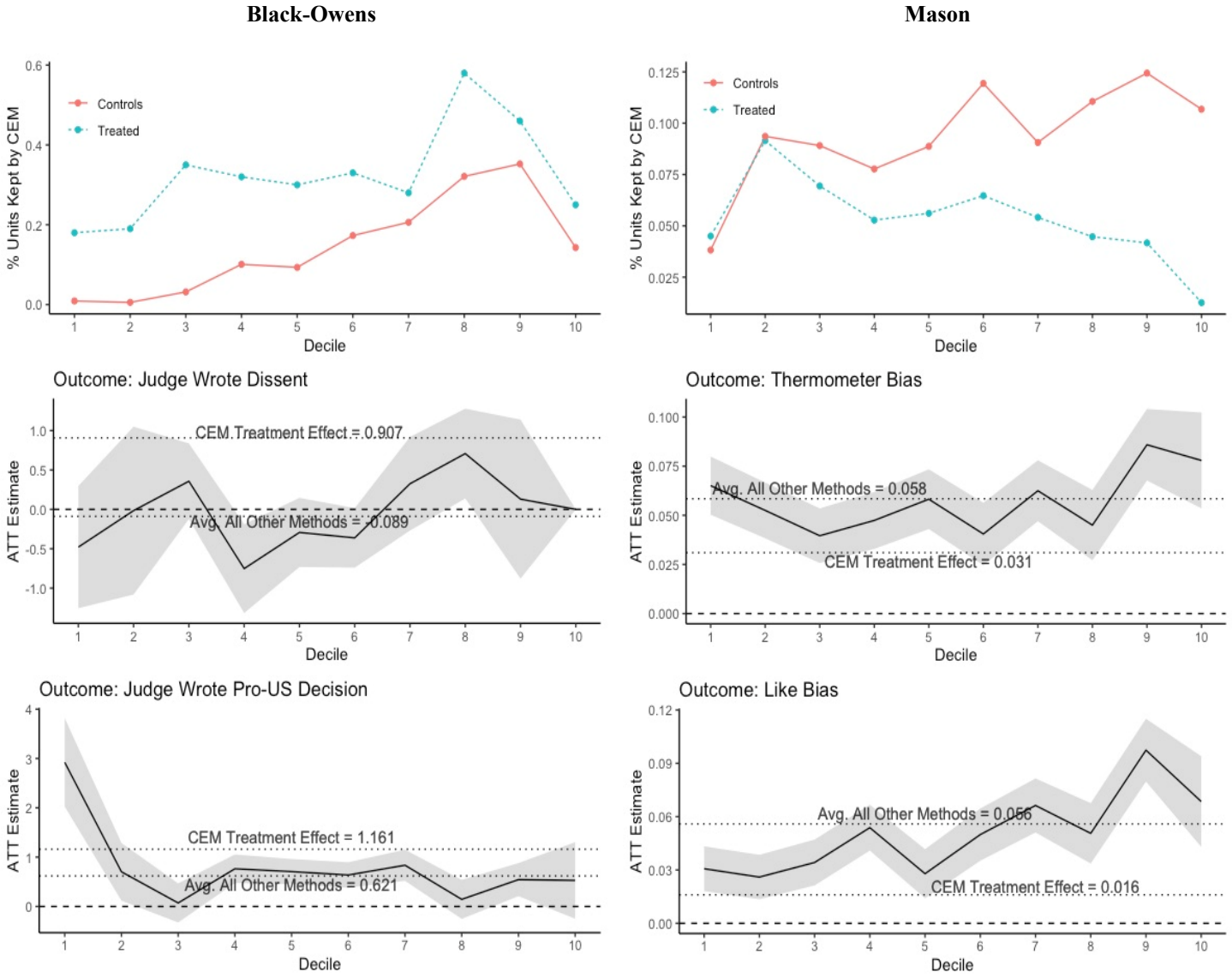
In Figure 8, we divide the Black-Owens and Mason samples into propensity score deciles and report within-decile CEM estimates. The top graphs shows the fraction of treated and control units retained by decile. Variation in this fraction across deciles will produce a biased ATT estimate if there is treatment heterogeneity that is correlated with the propensity score.

There are also larger problems with the CEM estimates. The middle and bottom figures show treatment effect estimates and 95% confidence intervals (CIs) by decile for two Black-Owens outcomes (left-hand panels) and two Mason outcomes (right-hand panels. Dotted horizontal lines show the average estimate from other methods (-0.089), the CEM estimate (+0.907), and the average of the CEM decile estimates, weighted by the number of treated units in each decile (0.027). One would expect the weighted average to be close to the full sample estimate. This expectation holds for the other methods (not reported). However, for both Black-Owens outcomes, the CEM weighted average is far below the full-sample estimate. The right-hand figures provide a similar analysis for two Mason results. All within-decile estimates are above the full-sample CEM estimate.

For all four outcomes, the weighted mean of within-decile estimates is much closer to other methods than the CEM full-sample estimate. To be sure, all estimates could be wrong, but given the decile results, it is hard to imagine a data generating process for which the CEM full-sample estimate reflects truth.

Figure 8. CEM: Analysis of Propensity Score Deciles for Black-Owens and Mason

Top graphs. Percent of treated and control units retained by propensity score decile for Black-Owens (left-hand) and Mason (right-hand). **Middle and bottom graphs.** Within-decline treatment effects for indicated outcomes. Horizontal lines show CEM full-sample estimate and average of other methods. Shaded areas show CIs.



B. Varying the Number of Covariates to be Balanced On

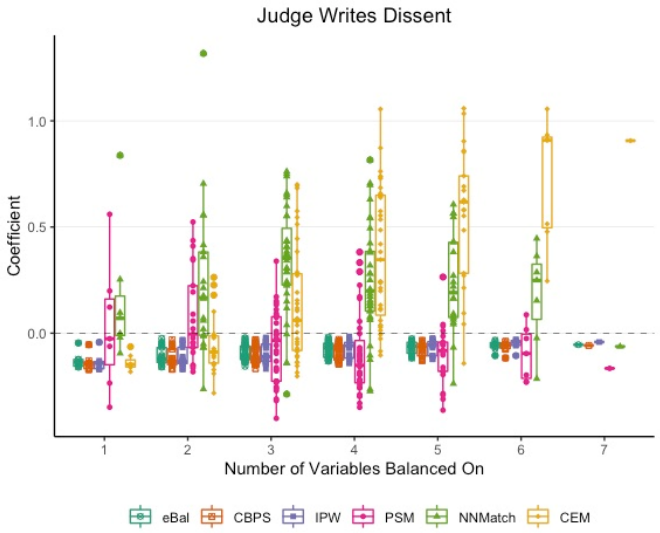
CEM requires researchers to carefully pick the covariates chosen for balancing, and limit their number, to preserve sample size. We therefore explore the sensitivity of estimates from CEM and the other methods to choice of which variables to balance on, for the same Black-Owens and Mason outcomes as in Figure 8. In Figure 9, the second stage regression includes all covariates. We vary the number of covariates balanced on. For a given number of covariates, we randomly choose which to balance on. For example, for 3 covariates for Black-Owens, we randomly select three of the 7 covariates, balance on those, compute ATT estimates, and repeat 1,000 times. The figure shows box-and-whiskers plots, which show the mean, 25th and 75th percentiles (box bounds), and estimates outside the box (whiskers).

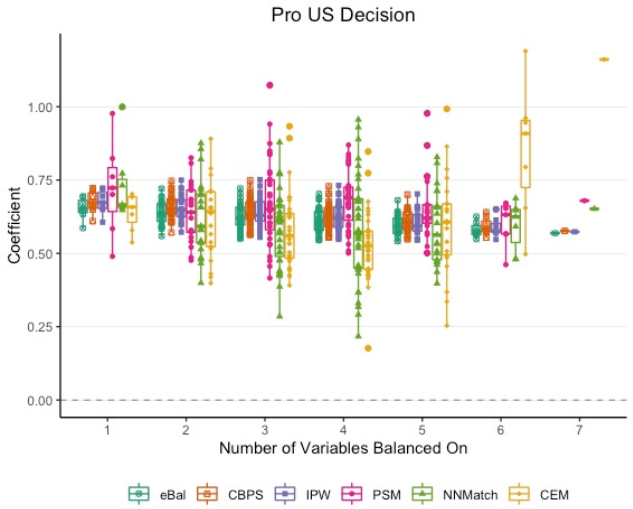
The reweighting estimates are relatively insensitive to the covariates used for balancing. The sensitivity of PSM and nnmatch to this choice can be substantial, but for these methods, one would have no reason to balance on a subset of covariates. Thus, the variation in ATT estimates shown in Figure 6 would not be problematic in the real world. For CEM, for Black-Owens, as the number of covariates increases (and the CEM sample shrinks), the CEM mean departs further and further from the other methods, and the range of ATT estimates becomes large, especially when one leaves out only one covariate. For Mason, the CEM estimates are stable, and similar to other estimates, when balancing on 1-7 covariates, but the CEM estimate drops and diverges from the other estimates as we add additional covariates. This instability is a further troubling feature of the CEM estimates. The CEM estimates are dependent on researcher-driven modeling choices, in a way that the other methods are not, because they don't require researcher decisions on limiting the covariates to be balanced on, imposing coarse bins, or some of both.

Figure 9: CEM Estimates, Varying the Covariates Balanced On

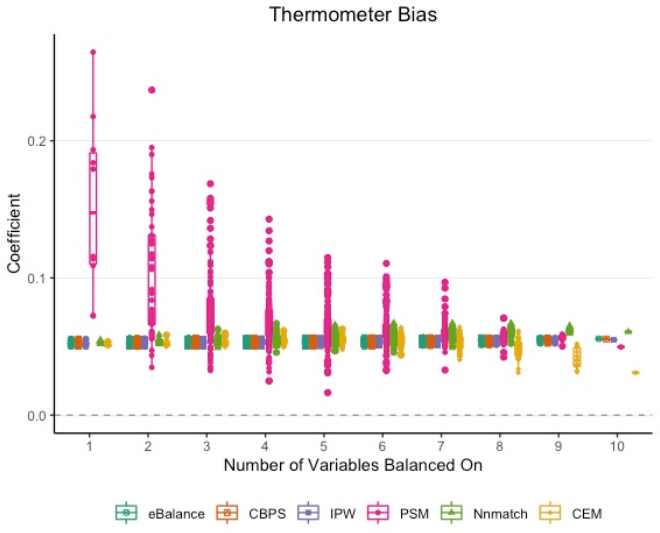
Box-and-whiskers plots of ATT estimates, varying the number of covariates balanced on. Second-stage regression includes all covariates. For each number of covariates, we randomly select this number from the full set of covariates, using 1,000 iterations.

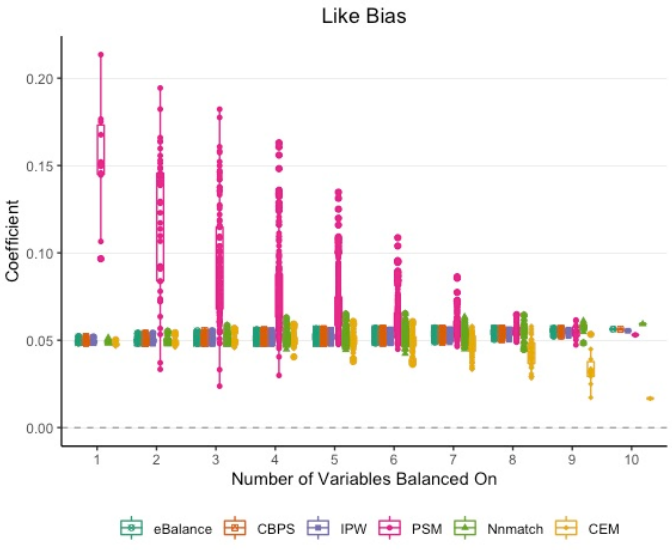
Panel A. Black Owens





Panel B. Mason





C. Sensitivity to Uninformative Covariates

We also explored the sensitivity of the balancing methods to adding additional random covariates. These covariates are, by construction, uncorrelated with both the treatment and the outcome. For regression, adding irrelevant covariates may modestly affect point estimates and precision, but will not produce bias. One should expect the same for a balancing method.

In Figure 10, we add randomly drawn variables to both the balancing and regression stages and iterate 1,000 times for each method. In the left-hand portion of each graph, we show box-and-whiskers plots for results adding a random binary variable (mean = 0.5); in the center portion, we add a continuous, unit-normal variable; in the right-hand portion we add both variables. The plots show the *change* in the ATT estimate after adding the random covariate(s). We winsorize the graphs at ±5 for visual presentation. Panel A shows Black-Owens results, Panel B shows Mason results. In both panels, the reweighting methods are minimally affected by adding random covariates. PSM and nnmatch perform decently, although less well.

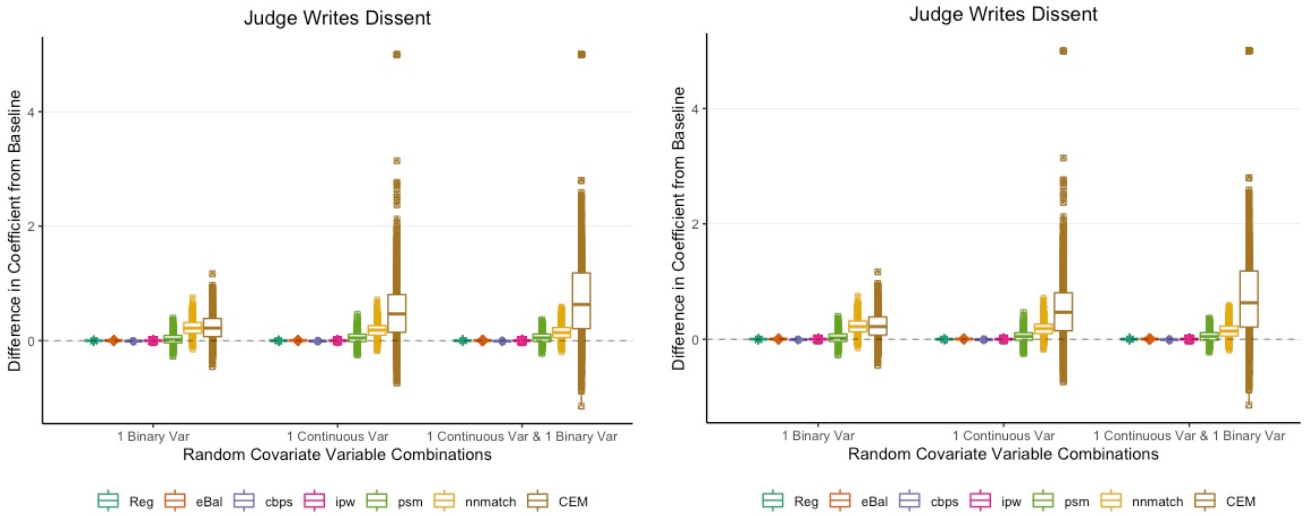
CEM is highly sensitive to adding noise variables, especially a continuous variable, and exhibits substantial skewness. In the Appendix, bias and skewness increase if we add two continuous random variables.

Researchers would not knowingly include a random variable in balancing or in a regression model. But they might include a variable that on theoretical grounds might correlate with the treatment or the outcome, but in the actual data is weakly correlated with both. For CEM, such a variable could greatly affect point estimates. This would not be apparent to the researcher, and the variable would appear highly relevant because it strongly affects point estimates.

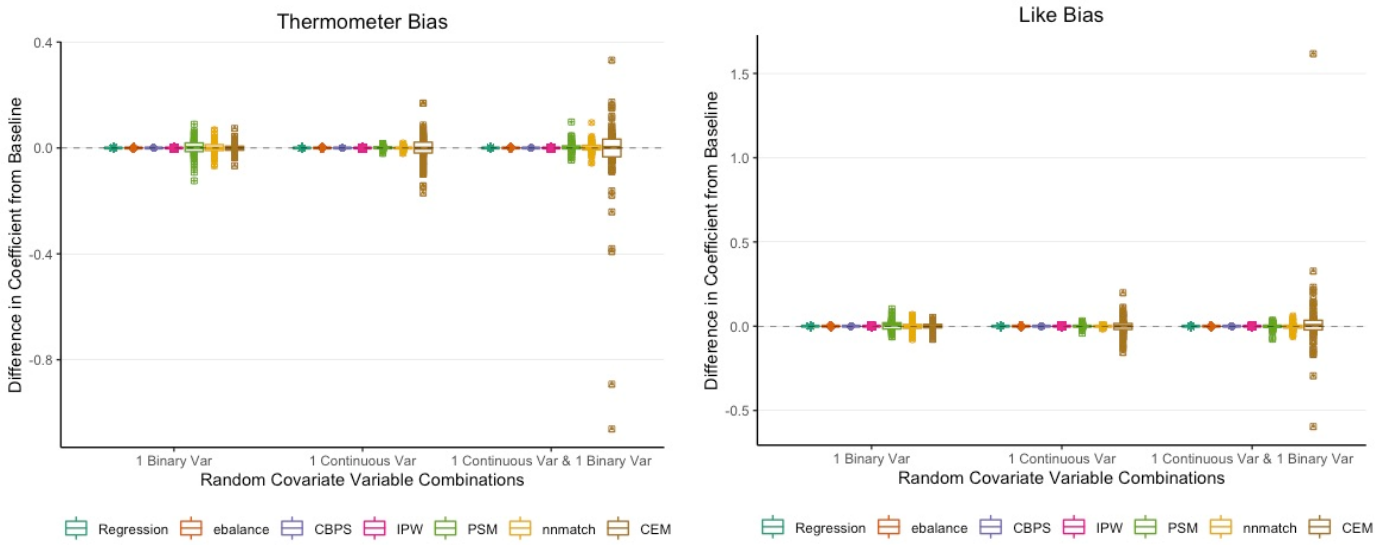
Figure 10. Sensitivity of ATT Estimates to Adding Random Covariates

Box-and-whiskers plots of ATT estimates for indicated outcomes, where we add a random binary variable, a continuous variable, or both. We draw each random covariate from the corresponding distribution and iterate 1,000 times. Maximum difference for estimates capped at (5, -5 for visualization.

Panel A. Black-Owens



Panel B. Mason



D. Statistical Power and Over-Rejection of the Null

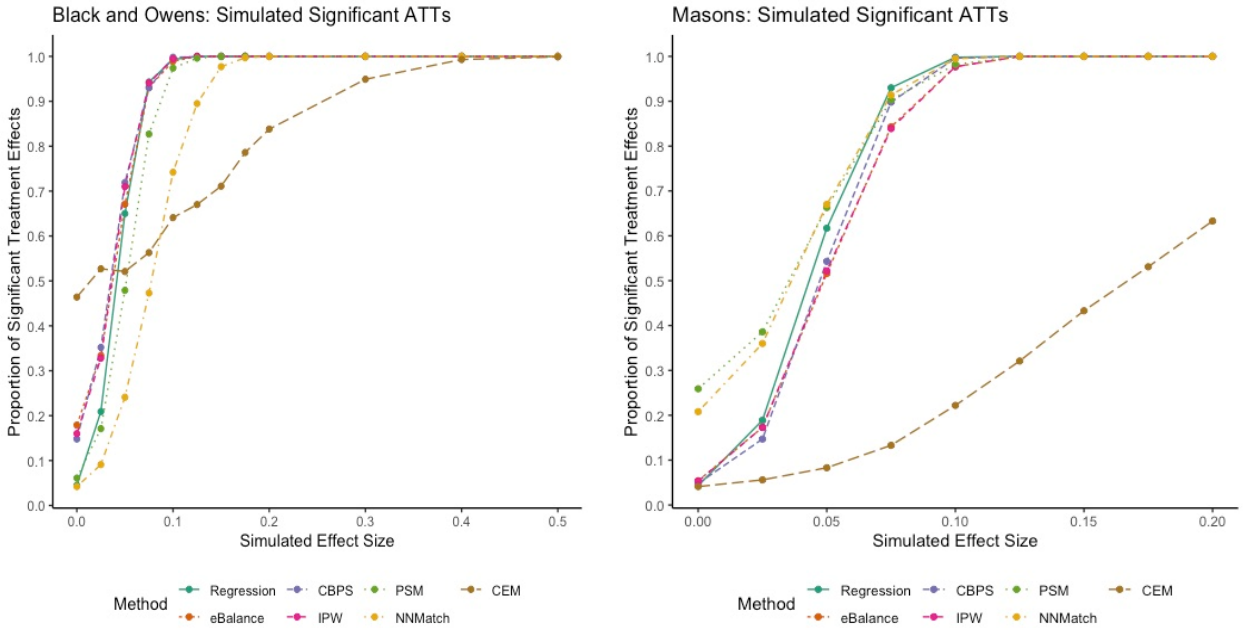
We next investigate the statistical power of each method, and whether it tends to over-or under-reject the null. We show results in Figure 11 for the Black-Owens dataset on the left and the Mason dataset on the right. For each, we begin with the actual sample and treatment assignment. We add a simulated, normally distributed outcome, with an imposed mean and unit standard deviation. We vary the imposed mean from zero (no treatment effect) to 0.20, in 0.025 increments. For each imposed effect, we draw unit-specific estimates from the unit-normal distribution, run each method, obtain the ATT estimate and s.e., and repeat 1,000 times. Figure 8 shows the results.

For Black-Owens, with an imposed null effect, regression has the expected 5% rejection rate; as do the matching methods. The reweighting methods over-reject. This is unexpected, although not ruled out by the double-robustness proofs, which address only consistency (e.g., Bang and Robins, 2005; Kang and Schaefer, 2007; Zhao and Percival, 2017). CEM severely overrejects. As we increase the imposed treatment effect, regression and the reweighting methods reach nearly 100% power for a 0.10 imposed effect of 0.10. CEM has much lower power; the matching methods have greater power than CEM but less than the reweighting methods.

For Mason, regression, the reweighting methods, and CEM have correct size for a zero imposed effect, while the matching methods over-reject the null. However, CEM has far less power than other methods.

Figure 11. Statistical Power

Figure shows for each method, simulated power to reject the null. See text for simulation details.



E. CEM Sensitivity to Number of Bins

Thus far, we have the CEM default choice for number of bins for continuous variables, which is 15 for the sample sizes for Black-Owens, Mason, and 14 for the simulated data (Figure App-1). The papers developing CEM provide little guidance on choosing number of bins, and among the few papers we reviewed that depart from the CEM defaults, almost none explain their choice. But this choice must be important. As the number of bins shrinks toward 1, covariate balance should worsen and the CEM estimate should converge toward the regression estimate, because binning does less work and less sample is lost.

In Table 3, we assess sensitivity to binning choice by varying the number of bins, and confirm the expected convergence to the OLS estimate. For Black-Owens, the CEM estimates with either 2 or 4 bins are close to the regression estimate and those from other methods. But when we increase the number of bins beyond this, sample sizes drop substantially and the CEM estimates diverge from the regression estimate. For Mason, the CEM estimates are close to those from regression and other methods for 2, 4, or 7 bins, but diverge after that. Thus, the CEM estimates are sensitive to binning choice. This sensitivity is not discussed in papers using CEM and is a source of model sensitivity.

Table 3. CEM Estimates: Sensitivity to Number of Bins

CEM estimates using indicated number of bins for continuous variables. CEM default is 15 bins for Black-Ownes and Mason; 14 bins for the simulated data. Regressions are otherwise similar to Tables 1 and 2. Robust s.e.'s in parentheses.

Panel A. Black-Owens Outcomes

	Regression	**CEM Estimates with Indicated Number of Bins**						
		2	4	7	10	15	20	25
Panel A1. Judge Writes Dissent								
Coefficient	-0.155	−0.172	−0.172	0.082	0.443	0.907	1.922	2.837
s.e.	(0.154)	(0.169)	(0.169)	(0.209)	(0.316)	(0.451)	(0.980)	(1.611)
No. of Obs.	11,787	11,787	11,787	6,452	4,330	2,490	1,322	1,296
Panel A2. Pro-U.S. Decision								
Coefficient	0.668	0.657	0.657	0.310	0.349	1.161	−0.063	−0.065
s.e.	(0.104)	(0.104)	(0.104)	(0.130)	(0.171)	(0.212)	(0.265)	(0.305)
No. of Obs.	5,805	5,805	5,805	3,295	2,319	1,356	741	762
Panel A3. Decision Consistent with President's Ideology								
Coefficient	0.291	0.221	0.221	0.137	0.249	0.350	0.488	0.084
s.e.	(0.048)	(0.048)	(0.048)	(0.066)	(0.089)	(0.117)	(0.154)	(0.176)

No. of Obs.	10,171	10,171	10,171		3,862	2,216	1,165	1,154

Panel B. Mason Outcomes

	Regression	CEM Estimates with Indicated Number of Bins						
		2	4	7	10	15	20	25
Panel A. Thermometer								
Coefficient	**0.0563**	**0.054**	**0.061**	**0.047**	0.020	0.031	0.019	0.014
s.e.	(0.005)	(0.005)	(0.006)	(0.008)	(0.013)	(0.016)	(0.022)	(0.021)
No. of Obs.	9,858	8,970	6,317	2,788	1,057	572	360	325
Panel B. Like Bias								
Coefficient	**0.0483**	**0.049**	**0.050**	**0.046**	0.017	0.017	0.016	0.029
s.e.	(0.004)	(0.005)	(0.005)	(0.008)	(0.012)	(0.017)	(0.022)	(0.023)
No. of Obs.	9,858	8,970	6,317	2,788	1,057	572	360	325
Panel C. Activism								
Coefficient	**0.0283**	**0.028**	**0.030**	**0.036**	**0.028**	0.025	0.019	0.010
s.e.	(0.004)	(0.005)	(0.005)	(0.007)	(0.011)	(0.014)	(0.018)	(0.019)
No. of Obs.	9,858	8,970	6,317	2,,788	1,057	572	360	325

VII. Analysis of Other CEM Papers

A. Urban and Niebler (2014)

We discuss Urban-Niebler and Carpenter briefly here, and provide details in the Appendix. Urban-Niebler study the effect on campaign contributions of "spillover" Presidential campaign ads, which reach residents in noncontested states who live in a TV-reception area that overlaps a neighboring contested state. They use PSM as their principal balancing method, and CEM as a robustness check. Appendix Table App-2 is adapted from the Black-Lerner (2022) reexamination of Urban-Niebler. It shows the predicted effect on whether a contribution is made for treated zipcodes (which received at least 1,000 spillover ads during the 2008 Presidential campaign); and on the amount contributed, conditional on a contribution being made. Here too, CEM is an outlier relative to the other methods. The CI for the estimated effect of spillover ads on amount contributed (given that a contribution is made) is bounded away from zero for other balancing methods but crosses zero with CEM.

C. Carpenter et al. (2012)

Carpenter et al. study how administrative deadlines shape decision timing and quality. They study FDA drug approvals, find that administrative deadlines induce examiners to issue many decisions just before the deadlines, and that just-before-deadline approvals have higher rates of future safety problems (severe, "black box" safety warnings; safety-based withdrawal from the market; and less-severe safety alerts). We present results in Appendix Table App-3.

For all three outcomes, the reweighting estimates are similar with CIs bounded away from zero, and similar to estimates from regression alone. In contrast, the matching methods, including CEM, produce more varied estimates, larger s.d.'s, and CIs that cross zero. CEM keeps only 35 of the 86 treated units, from an already small sample.

VIII. Discussion: Multiple Concerns with CEM

Our major takeaways are as follows. CEM and provides good covariate balance (Figure 1), but at the cost of much smaller retained samples than other methods (Figure 2), and thus lower precision (Figure 3). CEM is strongly biased in simulations with heterogeneous treatment effects (Figure 6), and can be apparently biased for real datasets, for which it produces very different results than the other methods (Figure 4). Full sample CEM results can be inconsistent with subsample results (Figure 8). If one limits the number of variables balanced on to preserve sample size (a tradeoff not needed for other methods), one can obtain widely varying estimates, depending on the variables chosen for balancing (Figure 9). CEM estimates are much more sensitive than those from other methods to including uninformative covariates (Figure 10). CEM can greatly over-reject a true null (Figure 11, Black-Owens), yet has much less power than the other methods to detect a true treatment effect (Figure 11). The CEM estimates are also sensitive to the number of bins used (Table 3).

The other balancing methods are known to be consistent – they converge to truth and thus to each other in asymptopia (e.g., Busso, DiNardo, and McCrary, 2014). For CEM, no consistency proof exists, and our simulation provides a counterexample: bias is driven by the correlation between the treatment effect and the propensity score, and will remain regardless of sample size. Any advantage of one method over others thus principally involves precision or bias in smaller samples.

Iacus, King and Porro (2012), at 2 assert that "CEM dominates commonly used . . . matching methods in its ability to reduce imbalance, model dependence [and] estimation error." CEM reduces imbalance on the balancing variables (comparable to eBalance and CBPS, Figure 1). However, it has higher estimation error (Figure 3), as well as model dependence due to researcher need to choose which covariates to balance on and the number of bins chosen (see also Wan, 2025). With the other methods, binning choices are not needed, one can balance on a broad set of available covariates without loss of sample size, and any choice of covariates has much less influence on effect estimates.

We pursue a more careful comparison of the other methods in separate work (Black and Lerner, 2024), but note that in this project, the reweighting estimates performed well, as did IPW (see also Busso, DiNardo and McCrary, 2014).

In Appendix Table App-4, we report Pearson correlation coefficients across methods for the weights on control units. The correlations across reweighting methods are very high. For Black-Owens, control unit weights for CBPS correlate at 0.995 with IPW; and eBalance weights correlate at 0.979 with IPW and 0.978 with CBPS; for Mason the corresponding correlations are 0.955; 0.853; and 0.925. Perhaps the reweighting methods will perform oddly in other datasets,

but the only warning sign in this project was over-rejection of the null when we imposed an artificial treatment effect on the Black-Owens data.

The matching methods (PSM and nnmatch) performed less well than the reweighting methods (consistent with the concerns about PSM in King and Nielsen, 2019). Still, they did not show the extreme departure from other methods seen for CEM, nor CEM's sensitivity to choice of covariates, adding uninformative covariates, or varying the number of bins.

Our overall judgment: CEM should never be used as a primary balancing method. If used, it requires extensive sensitivity checks. In particular, researchers should: (i) explain their choice of how many bins to use and report how results change when they vary this choice; (ii) explain choice of which variables to balance on and report how results change with different choices; and (iii) whether full-sample estimates are consistent with those built up from subsamples.

References

Abadie, Alberto, and Guido W. Imbens (2011), Bias-Corrected Matching Estimators for Average Treatment Effects, 29 *Journal of Business & Economic Statistics* 1-11.

Angrist, Joshua D., and Jorn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion* (2008).

Bang, Heejung, and James M. Robins (2005), Doubly Robust Estimation in Missing Data and Causal Inference Models, *Biometrika* 61(4): 962-973.

Black, Bernard, and Joshua Y. Lerner (2022), Spillover Presidential Ads and Campaign Contributions in a Polarized System, working paper, at http://ssrn.com/abstract=3xxxxxx.

Black, Bernard and Joshua Lerner, Matching Strategies Compared: Assessing How Results in Observational Studies Vary across Balancing Methods (working paper 2024), at http://ssrn.com/abstract=4xxxxxx.

Black, Ryan C., and Ryan J. Owens (2016). "Courting the president: how circuit court judges alter their behavior for promotion to the Supreme Court." *American Journal of Political Science* 60(1): 30-43.

Broockman, David E. (2013), "Black politicians are more intrinsically motivated to advance blacks' interests: A field experiment manipulating political incentives." *American Journal of Political Science* 57(3): 521-536.

Busso, Matias, John DiNardo, and Justin McCrary (2014), New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators, 96 *Review of Economics and Statistics* 885-897.

Carpenter, Daniel, Jacqueline Chattopadhyay, Susan Moffitt, and Clayton Nall (2012), The complications of controlling agency time discretion: FDA review deadlines and postmarket drug safety." *American Journal of Political Science* 56(1): 98-114.

Chattopadhyay, Ambarish, Christopher H. Hase, and Jose R. Zubizarreta (2020), Balancing versus Modeling Approaches to Weighting in Practice, *Statistics in Medicine* 39:3227-3254.

Crump, Ricard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnick (2009), Dealing with Limited Overlap in Estimation of Average Treatment Effects, 96 *Biometrika* 187-199.

Diamond, Alexis, and Jasjeet S. Sekhon. "Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies." *Review of Economics and Statistics* 95, no. 3 (2013): 932-945.

Graham, Bryan S., Cristine Campos de Xavier Pinto, and Daniel Egel (2012), Inverse Probability Tilting for Moment Condition Models with Missing Data," *Review of Economic Studies*, 79(3), 1053–1079.

Greifer, Noah, and Elizabeth A. Stuart. "Matching methods for confounder adjustment: an addition to the epidemiologist's toolbox." *Epidemiologic reviews* 43, no. 1 (2021): 118-129.

Hainmueller, Jens (2012), Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies, 20 *Political Analysis* 25-46.

Hansen, Ben B. (2004), Full Matching in an Observational Study of Coaching for the SAT, 99 *Journal of the American Statistical Association* 609-618.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, (2009) *The Elements of Statistical Learning: data mining, inference, and prediction*. Springer Science & Business Media.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart (2007), Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference, *Political Analysis* 15: 199-236.

Iacus, Stefano M., Gary King, and Giuseppe Porro (2012), Causal inference without balance checking: Coarsened exact matching, *Political Analysis* 20: 1-24.

Imai, Kosuke, and Marc Ratkovic. (2014) "Covariate balancing propensity score." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76, no. 1: 243-263.

Imbens, Guido W. (2015), Matching Methods in Practice: Three Examples, *Journal of Human Resources* 50: 373-419.

Imbens, Guido W, and Donald B. Rubin (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge University Press.

Kang, Joseph D.Y., and Joseph L. Schafer (2007), Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data, *Statistical Science* 22(4), 523-539.

King, Gary, Christopher Lucas, and Richard A. Nielsen (2017), The Balance-Sample Size Frontier in Matching Methods for Causal Inference, 61 *American Journal of Political Science* 473-489.

King, Gary, and Richard Nielsen (2019), Why Propensity Scores Should Not be Used for Matching, 27 *Political Analysis* 435-454.

Kish, Leslie (1965). *Survey Sampling.* New York: John Wiley & Sons, Inc. ISBN 0-471-10949-5.

Mason, Lilliana (2015), "I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science* 59, no. 1 (2015): 128-145.

Ripollone, John E., Krista F. Huybrechts, Kenneth J. Rothman, Ryan E. Ferguson, and Jessica M. Franklin (2020), Evaluating the Utility of Coarsened Exact Matching for Pharmacoepidemiology Using Real and Simulated Claims Data, *American Journal of Epidemiology* 189(6):613–622.

Sekhon, Jasjeet S. (2009), Opiates for the Matches: Matching Methods for Causal Inference, *Annual Review of Political Science* 12: 487-508.

Sloczynski, Tymon, S. Derya Uysal, and Jeffrey M. Wooldridge (2025), Covariate Balancing and the Equivalence of Weighting and Doubly Robust Estimators of Average Treatment Effects, CES working paper 12,152.

Stuart, Elizabeth A. (2010), Matching Methods for Causal Inference: A Review and a Look Forward, *Statistical Science* 25(1): 1-21.

Urban, Carly, and Sarah Niebler (2014), Dollars on the Sidewalk: Should US Presidential Candidates Advertise in Uncontested States?, *American Journal of Political Science* 58(2): 322-336.

Wan, Fei (2025), Examining the Efficacy of Coarsen Exact Matching as an Alternative to Propensity Score Matching, working paper.

Wooldridge, Jeffrey (2025), Introductory Economics: A Modern Approach (8th Ed.).

Zhao, Qingyuan, and Daniel Percival (2017), Entropy Balancing is Doubly Robust, *Journal of Causal Inference* 5: DOI: 10.1515/jci-2016-0010.