The Trouble with Coarsened Exact Matching

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

Parth Lalkiya
Northwestern University

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 June 2022)

The data and code to generate all results in this project are available at:
[*url to come]

The Trouble with Coarsened Exact Matching

Bernard Black, Parth Lalkiya, and Joshua Y. Lerner*

Abstract: "Balancing" methods, using matching or reweighting to improve the balance between treated and control units, are central methodological tools for causal inference in the social sciences using cross-sectional observational data. We address here 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 substantially worse than other balancing methods and explain why it does so. We replicate five recent papers that use CEM and compare CEM-based results to those from other methods. CEM drops substantially more observations than other methods; is much less precise; can severely misidentify average treatment effects relative to other methods, and to CEM itself (applied by subsetting the sample, applying CEM to each subset, and combining the subset estimates); can produce estimates that are sensitive to adding noninformative covariates; and can over-reject the null when the null is true. Our advice: never use CEM as the sole balancing method, and there is little to be said for using it at all.

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

* 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 a post-baccalaureate research fellow at Northwestern, email: parth.lalkiya@gmail.com Lerner is Research Methodologist and Data Scientist at NORC at the University of Chicago, email: joshlerner1@gmail.com.

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. One method which has become popular, especially in political science, is Coarsened Exact Matching (CEM) (Iacus, King, and Porro 2012). Often, balancing methods are used as a form of "preprocessing" a sample to improve balance prior to regression analysis. CEM is intended to be used in this way, but other balancing methods can also be combined with regression (e.g., Ho et al., 2007; Stuart, 2010).

CEM is a combined 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 (the "coarsening" part), requires an exact match between treated and control units on the binned variables (the matching part), drops unmatched observations (the trimming part), and reweights the remaining observations (the reweighting part). Thus, CEM can usefully be compared both to other matching methods and to other reweighting methods.

We compare CEM, used to measure the average treatment effect on the treated (ATT) to five other well-known, widely used balancing approaches: propensity score matching (PSM), nearest-neighbor matching (nnmatch, Abadie and Imbens, 2011); 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). We chose methods that are well-known in political science; are implemented in Stata, R, or both; and appear to perform well from prior research.[1] As a basis for comparison, we began with a set of all papers using matching and reweighting

[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

published in the *American Journal of Political Science* over 2012-2016, obtained from a colleague. We study all four papers which use CEM: Black and Owens (2016); Mason (2015); Urban and Niebler (2012); Carpenter et al. (2012).[2] We replicate selected results from each paper and compare results using CEM to those from the other methods, in each case followed by regression on the balanced dataset; and also to regression alone without balancing. For a dataset with many covariates, CEM users must decide which to balance on and assess the tradeoff between better balance versus smaller sample size if one balances on more covariates. We used the authors' choices of which variables to balance on.

At the core of causal inference with observational data is the need to impute the unobserved potential outcome for treated units from similar control units. CEM, unlike the comparison methods, drops many treated units, for which matching control units cannot be found. This reduces sample size; the loss of sample size increases rapidly with the number of covariates that one balances on. Thus, CEM suffers strongly from the curse of dimensionality. In contrast, the comparison methods preserve all treated units.

We assess how far the coefficient estimates from each method are from the means from the other methods. We also assess the relative precision of each method. CEM produces higher s.e.'s than the other methods, sometimes much higher. The loss of precision flows from the loss of sample size. Moreover, the CEM loss in sample can come from non-apparent parts of the sample space. If treatment effects are heterogeneous, this can produce treatment effect estimates that are far from true effects.

In the text, we illustrate CEM's odd behavior using principally Black-Owens (2016) and Mason (2015). Using CEM, Black-Owens find support for their conjecture that federal appellate judges who

eBalance and CBPS-weights, perform well relative to IPW. All of these methods are implemented in R in MatchIt and inWeightIt (Greifer and Stuart 2021).

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

are contenders for Supreme Court vacancies write different opinions when a vacancy exists. In contrast, regression alone and all comparison methods produce insignificant, negative coefficients. CEM also produces outlier results for Mason (2015, but we relegate most details to the Appendix). For Mason's Thermometer Bias and Like Bias outcomes, both regression alone 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, CEM coefficients are small and statistically insignificant. While we do not know truth for either paper, that CEM is an outlier compared to all other approaches strongly suggests that the CEM estimate is incorrect.

CEM coefficient estimates are also more sensitive than the other methods to which covariates are used for balancing (followed by regression on all covariates). 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. Moreover, CEM estimates change substantially if we add random covariates (uncorrelated with the outcome or treatment assignment) to the actual covariates, and also balance on the additional random covariates. Yet, by construction, these random covariates should not affect the true ATT. Regression alone and the comparison methods are unaffected by adding random covariates.

Given this array of issues, our advice is to never use CEM as the sole balancing method. We cannot rule out the potential for CEM to outperform the comparison methods for specific datasets or data structures that we did not study. But our analysis suggests that other methods, available in Stata, R, or both, perform better. Although not a focus of this project, we also find evidence, for the real-world

datasets we studied, that the reweighting methods generally outperform the matching methods (PSM and nnmatch).[3]

This project emerged from a broader project using papers drawn principally from the *American Journal of Political Science* (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 planned merely to compare it to other popular approaches. Our decision to write separately about CEM emerged when we observed the stark differences between CEM and other balancing approaches.

This paper proceeds as follows. Part II provides background: a summary of CEM and our other balancing methods, and an overview of the results we replicate across methods. Part III discusses our methods. Part IV provides an overview of the differences between CEM and other methods. Part V discusses the results from each method, as applied to each paper. Part VI discusses the main takeaways from our analysis.

II. Background

We provide in the Appendix an overview of each of the matching methods and how they compare to each other. We provide a more summary treatment here, focusing on CEM.

A. Summary of Coarsened Exact Matching

We offer a non-technical summary of 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

[3] An important technical note: The CEM native code should not be used with binary or categorical variables. The CEM implementation in the MatchIt package corrects this problem.

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 will most users.[4] Each retained treated unit gets weight = 1.[5] 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:[6]

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

These weights are then 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 practice 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.

Something in the CEM code assigns binary variables to multiple bins, not only the lowest and highest. As shown in the Appendix, this produces substantial loss of sample size and important variation

[4] We confirm in the Appendix that departures from the default binning are uncommon. **[*details to come from Parth]**

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

[6] Our notation loosely follows Iacus (2012).

in effect estimates if one uses the CEM default bins. Categorical variables have a similar problem. This coding error is fixed in MatchIt, so users should use the MatchIt implementation.[7]

B. Other Balancing Methods

We compare CEM to regression alone and to five other balancing methods which are commonly used in our experience and have code available in Stata, R, or both, and are amenable to off-the-shelf use.[8] We compare CEM both to methods that provide balance only in expectation and methods that aim at exact covariate balance. For all approaches, we estimate ATT (average treatment effect on the treated). For PSM (propensity score matching), 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 to create a balanced sample, followed by regression; we use it with regression, with 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 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 (Ho et al. 2007).

[7] CEM with manual binning of binary and categorical variables gives the same results as its MatchIt implementation.

[8] We use CEM in R but confirm that we obtain the same results in Stata.

[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] We use eBalance.ado in Stata, but confirmed that we obtain the same results in R. 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.

C. Comparison Papers

We summarize here the four papers which use CEM that we reassess.

1. Black and Owens (2016)

Black and Owens assess whether federal appellate judges who are plausible candidates for the U.S. Supreme Court (candidate judges) change their voting behavior to curry favor with the President, at times when Supreme Court vacancies exist. They use CEM plus regression as their primary 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; to write a pro-US decision, or more likely to support the President's position.

2. Mason (2015)

Mason hypothesizes that "sorting" – the extent to which party identification matches political views -- predicts greater political partisanship. She uses CEM without regression to study whether sorting, controlling for the strength of party identity, predicts four measures of partisanship, which she terms thermometer bias, like bias, activism, and anger (her Figure 5). We study these outcomes, using CEM and other balancing methods with regression.

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, in each case without regression. Genetic matching is similar conceptually to eBalance and CBPS – it provides exact or near-exact balance on covariates in a finite sample. In Black and Lerner (2020), we show that Urban-Niebler misestimated propensity scores, correct this error, and apply a zero-inflated negative binomial model. We find, using

various balancing methods, that spillover ads predict higher contribution amounts, but not higher likelihood of contributing. We examine that outcome here.

4. Carpenter et al. (2012)

Carpenter and coauthors examine whether Federal Drug Administration (FDA) drug approvals, issued close to a Congressional time deadline for FDA action on applications are more likely to lead to approval of drugs that later turn out to have important side effects. They use CEM and optmatch (Hansen, 2004) in robustness checks. We study the three outcomes (black box warnings, safety-based withdrawals, and safety alerts) for which they report statistical significance for their base model (their Table 3).

III. Overview of Performance Measures

For 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 uses ordinary s.e.'s, as do the papers we replicate. We report inference based on randomization inference. We pull the appropriate number of pseudo-treated and pseudo-control firms randomly from the underlying dataset, apply each method to the pseudo-sample, compute the ATT estimate, repeat 1,000 times, and compute the standard deviation (s.d.) of the and standard of the estimates.[11] 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, we report a z-score, as a measure of how far each estimate is from the average "truth" provided by the other methods, defined for each method v relative to the other methods ($-v$) as:

[11] The Appendix contains expanded regression tables that includes standard errors

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

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 matching and reweighting 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 $n_{c,eff}$ (which we round 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, before normalization, 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 compute the means of the absolute values of the ND's across all covariates $\overline{|ND_m|}$.

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

IV. Overview of Differences Between CEM and Other Methods

In this part, we provide an overview of the performance of each method along several dimensions: covariate balance; preserving sample size; precision; and z-scores.

A. Covariate Balance

In Figure 1, we plot the mean of the normalized differences for each method; in the bottom panel, we report the mean of the fractional differences. 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. The average (across papers) of the averages (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. CBPS does next best, with all means of NDs and FDs less than 0.1. CEM often does very well, but less so using the FD measure. Nnmatch, IPW, and PSM have variable performance across papers, with no clear preference between them.

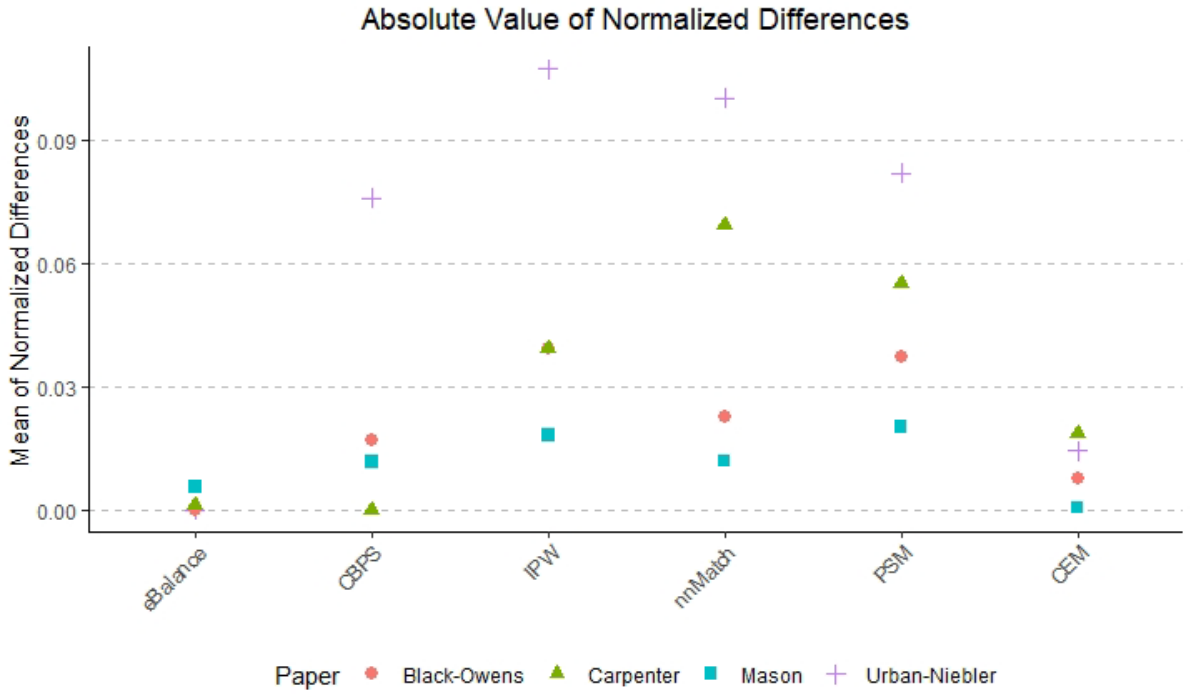
There are two main reasons why CEM achieves imperfect balance, despite discarding observations that it cannot match exactly. First, CEM can produce imbalance for variables that CEM does not try to balance on. However, the comparison papers use CEM to balance on all covariates. Less obviously, CEM generates imperfect balance for continuous variables that are matched on. The culprit is the coarsening. The intuition behind why CEM can produce worse balance than matching on covariate values (as in nnmatch) is straightforward. Consider a vector of continuous variable **x**, which CEM would place in bin b_x, and a distance metric m (say Mahalanobis distance), and initially hold sample size constant. CEM will match treated unit *i* to a control unit *j*, in the same bin with different actual value $\mathbf{x}_j \neq \mathbf{x}_i$ Nnmatch will match unit *i* to unit *j*, if *j* is the closest other available unit, but will match to another unit *k*, perhaps in a different bin, if *k* is closer ($m_{ik} < m_{ij}$). Thus, for each unit, the nnmatch distance will

be equal to or less than the CEM distance: $m_{ik} \leq m_{ij}$. Matters are more complex given that CEM will drop mismatched observations, while nnmatch will not, unless one imposes calipers. The average distance for the treated observations that CEM retains can therefore be either less or more than the average distance in nnmatch for the same observations.

In practice, for the papers we study, the CEM distances are generally smaller than nnmatch (see Figure 1), and also smaller than those for all other methods except eBalance, at the cost of reduced sample size.

Figure 1: Covariate Balance Across Methods

Figure shows mean (across covariates) of absolute values of normalized differences between treated and control units for each method, for each paper, for the covariates used in that paper. Small table below the panels shows the mean of means with no balancing, and for each balancing method.



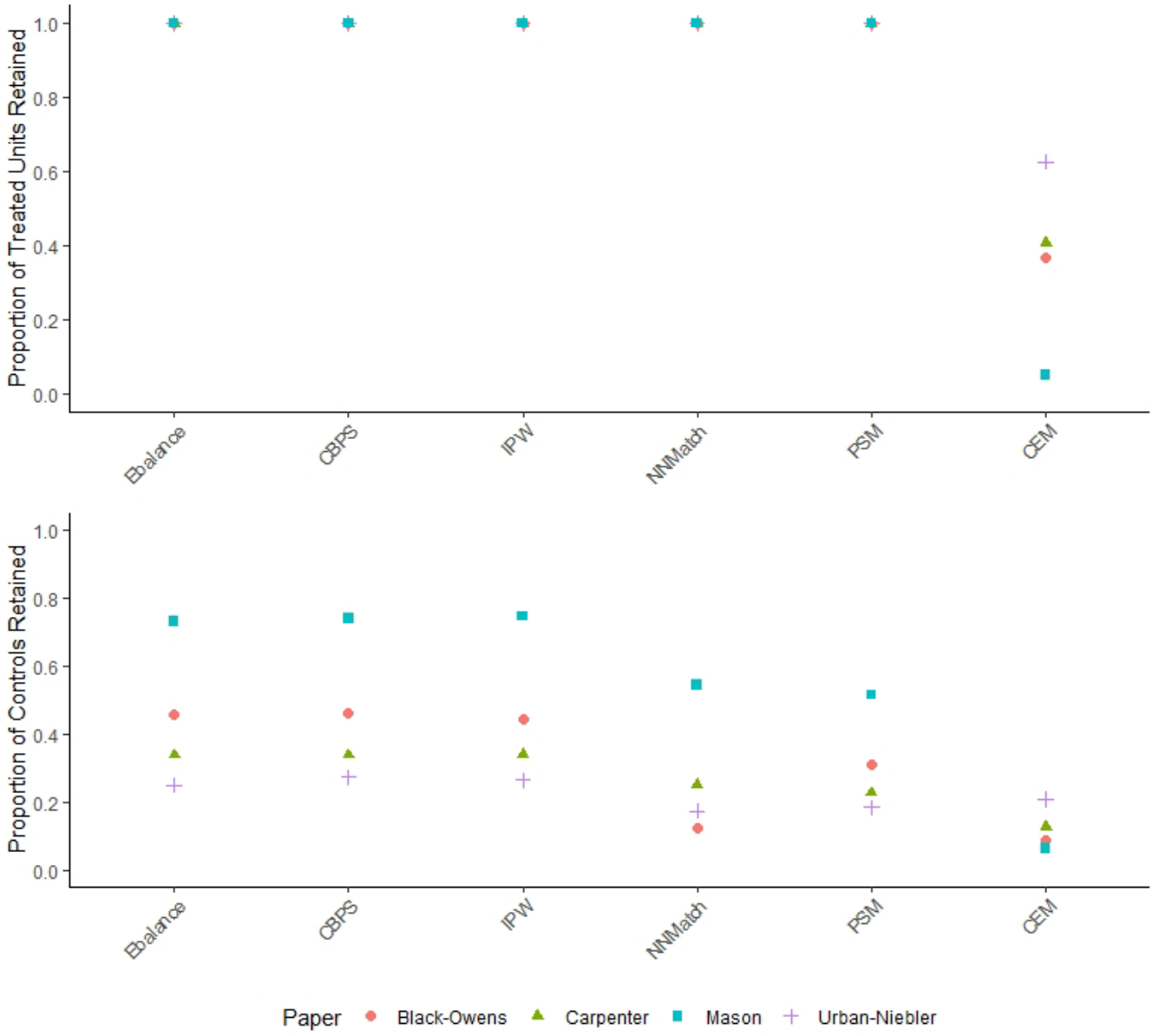
Mean of means	Raw	eBalance	CBPS	IPW	nnmatch	PSM	CEM
NDs	0.265	0.002	0.026	0.051	0.051	0.049	0.011

B. Effect of Different Balancing Methods on Sample Size

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 arrayed along the x-axis in the same order as in Figure 1; 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 varies widely, from 61% for Urban and Niebler to only 5% for Mason. The large loss of sample size for Mason reflects the curse of dimensionality: Mason balances on 10 variables. 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.

Figure 2: Percent of Treated and Control Units Retained

Top panel: Proportion of treated units retained after balancing. All methods other than CEM retain all treated units. **Bottom panel:** Proportion of effective control units retained after balancing. Balancing methods are shown along the x-axis, and proportion of units retained is shown vertically. For Carpenter, we use his larger sample, covering 1993-2007.



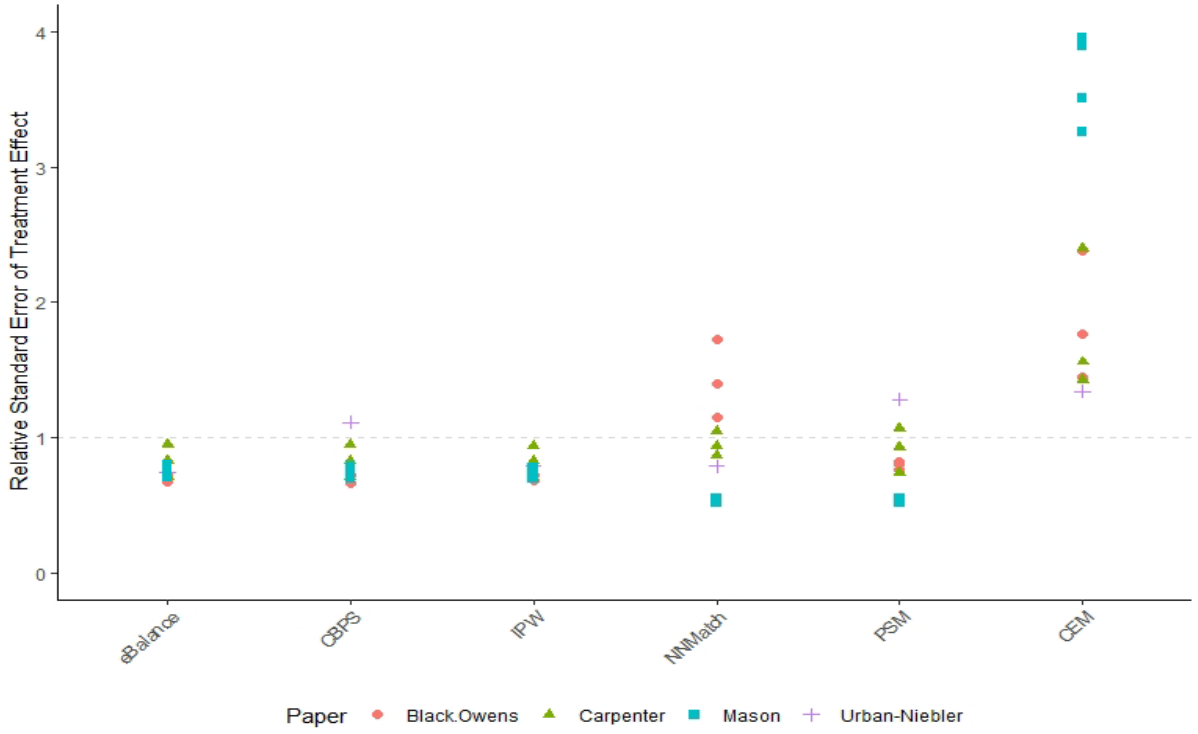
C. Relative Standard Errors

The loss of sample from using CEM has substantial implications for precision. We illustrate this in Figure 3. The figure shows the ratio of the s.e. using a particular method (larger of ordinary or robust

s.e.'s) to the average for the other five methods. For papers for which we examine more than one outcome (Black-Owens, Carpenter, Mason), the figure shows one data point for each outcome. 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. For the other methods, s.e.'s are similar for eBalance, CBPS, IPW, and PSM; and somewhat higher for nnmatch the highest.

Figure 3: Relative Standard Errors

Comparison of standard errors (s.e.'s) for treatment effect estimates by method. Balancing methods are shown along the x-axis. Y-axis shows the ratio of the s.e. using the indicated method to the average of the s.e.'s for the other methods, using for each the larger of ordinary or robust s.e.'s. For papers for which we study multiple outcomes (Black-Owens; Carpenter; Mason), graph shows one data point for each outcome.



	eBalance	CBPS	IPW	nnmatch	PSM	CEM
Mean	0.821	0.835	0.836	0.891	0.779	2.22

Figures 2 and 3 illustrate an important tradeoff for CEM that is 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.

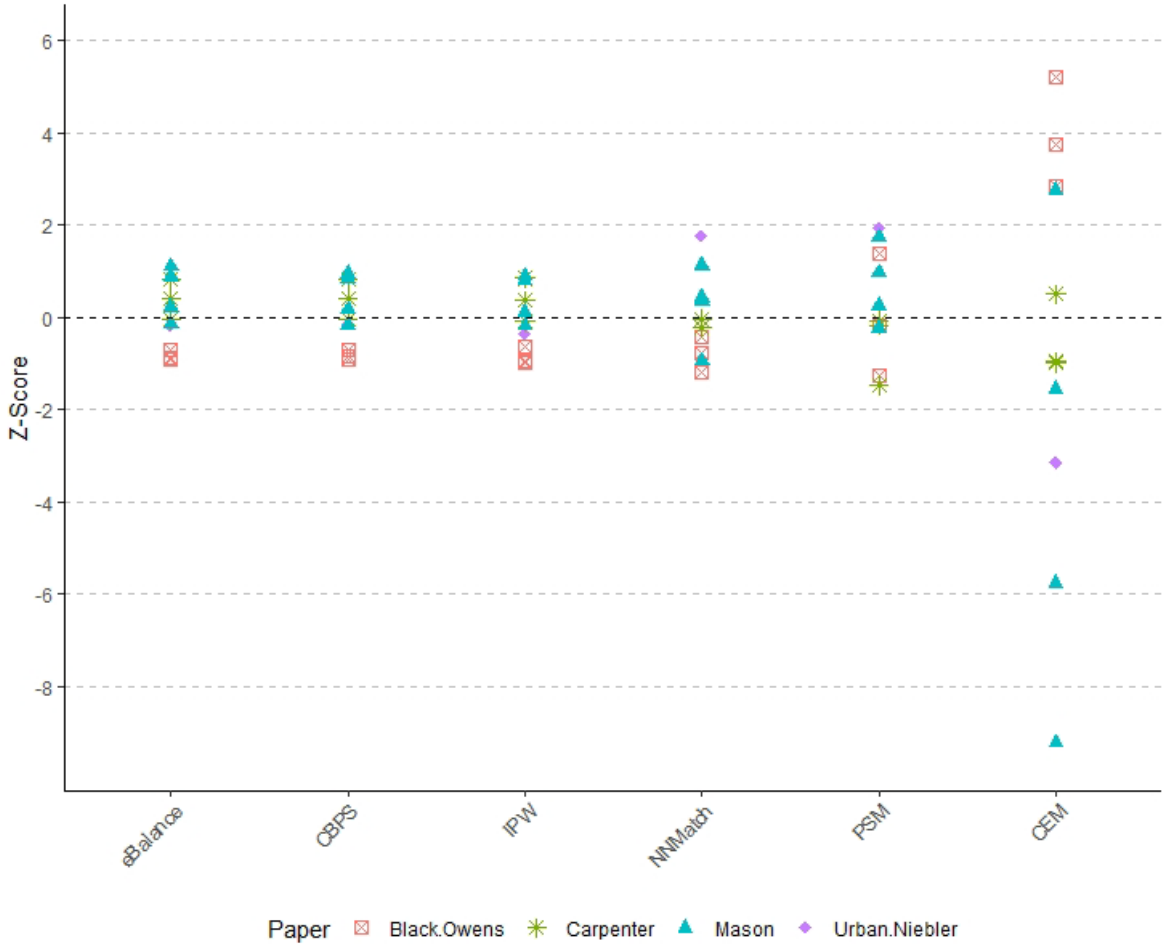
E. Point Estimates: CEM Versus Other Methods

In Figure 4, we plot the z-scores for the treatment effect estimates from each method. We do not know truth, but the average estimate from a number of methods should provide a fair approximation to truth. Thus, a large positive or negative z-score indicates that an estimate is likely to be incorrect. 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 and maximum of the absolute values of the z-scores for each method.

CEM performs dramatically worse than the other methods. A majority of the CEM estimates are outside the 95% confidence bounds from the other methods ($z > \pm 1.96$); the CEM scores range from -9.23 (Mason, like bias) to +5.19 (Black-Owens, judge writes dissent). Among the other methods, the reweighting methods perform well, and the matching methods somewhat less well. nnmatch nearly as well, with z-scores near zero and reasonably symmetric around zero. The stark message from Figure 4: CEM estimates cannot be trusted. Below, we examine more closely two Black-Owens outcomes and two Mason outcomes, and confirm that CEM produces implausible estimates for these outcomes.

Figure 4: Z-Scores for Treatment Effects

Comparison of z-scores for treatment effect estimates by method. Balancing methods are shown along the x-axis. Y-axis shows the z-scores for the indicated method, each measured relative to the mean coefficient and s.e. from the other estimates, using for each method the larger of ordinary or robust s.e.'s. For papers for which we study multiple outcomes (Black-Owens; Carpenter; Mason), graph shows one data point for each outcome.



	eBalance	CBPS	IPW	nnmatch	PSM	CEM
Mean Abs. Value	0.571	0.543	0.563	0.665	0.878	3.34
Max Abs. Value	1.1	0.94	1.05	1.76	1.95	9.24

V. Analysis of Black and Owens (2016)

We discuss Black and Owens (2016) in detail, to illustrate what can go wrong when using CEM. Black and Owens hypothesize that federal appellate judges who are plausible candidates for the U.S. Supreme Court change their voting behavior during periods when the Supreme Court has a vacancy, and make decisions which are closer to the President's political preferences. CEM plus logistic regression is their sole method. We study here the three outcomes for which they find support for their hypothesis using CEM: Judge Writes Dissent, Judge Writes Pro-US Opinion, and Decision Consistent with President's Ideology. Their sample is 11,787 decisions by panels with at least one judge who was a contender for the Supreme Court vacancy during their sample period, from 1946 to 2010.

A. CEM Versus Comparison Methods

We present regression results in Table 1 for logistic regression alone without balancing, and for balancing using CEM and the comparison methods, followed by logistic regression. The last column replicates the Black-Owens results. Consider first the outcome Judge Writes Dissent (Panel A). The CEM coefficient is 0.907 and statistically significant, with z = 5.19. In contrast, all other methods provide small *negative* and statistically insignificant coefficients, with small z-scores. CEM also retains only 1,803 of the 4,901 "treated" decisions (decisions when a vacancy exists) and only 611 effective controls.

Table 1. Black-Owens Results with Different Balancing Methods

Last column provides replication of Black-Owens use of CEM plus logit regression results from their Appendix Table 2, for contender judges, for indicated outcomes and the Black-Owens covariates: Judge's Judicial Common Space (JCS) score; Judge-President ideological distance; Judge-Panel Ideological Distance; circuit median JCS score; Supreme Court median JCS score; trial court is reversed; and case is published. **Panel A** (judge writes dissent); **Panel B** (judge writes pro-US opinion); **Panel C** (opinion consistent with President's ideology). Other columns use indicated balancing methods plus logit regression. Coefficients on covariates and constant term are suppressed. Ordinary standard errors (s.e.'s) in brackets; robust standard errors and randomization inference standard deviations (s.d.'s), based on 1,000 draws, in parentheses. * indicates significance at the 5% level based on randomization inference s.d.'s. Significant results, at 5% level or better, are in **boldface**. Z-score for given method is estimate for that method minus mean estimate from other methods)/(mean s.e. from other methods); z-score, s.e. ratio, and statistical significance are measured using the larger of ordinary or robust s.e.'s.

	(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. [ordinary]	[0.159]	[0.161]	[0.261]	[0.166]	[0.168]	[0.108]	[0.451]
s.e. (robust)	(0.154)	(0.186)	(0.255)	(0.168)	(0.166)	(0.165)	(0.115)
random inf. s.d.	0.142	0.177	0.180	0.142	0.142	0.143	0.196
z-score	--	-1.26	-0.79	-0.63	-0.70	-0.70	5.19
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. [ordinary]	[0.101]	[0.102]	[0.243]	[0.097]	[0.097]	[0.063]	[0.212]
s.e. (robust)	(0.104)	(0.131)	(0.184)	(0.120)	(0.120)	(0.119)	(0.091)
random inf. s.d.							
z-score	--	-0.18	-0.44	-0.94	-0.91	-0.93	3.74
Panel C. Decision Consistent w President's Ideology							
Vacancy Exists	**0.291***	**0.283***	0.126	**0.131***	**0.137***	**0.140***	**0.350***
s.e. [ordinary]	[0.048]	[0.049]	[0.091]	[0.048]	[0.047]	[0.031]	[0.116]
s.e. (robust)	(0.048)	(0.063)	(0.097)	(0.056)	(0.056)	(0.056)	(0.044)
random inf. s.d.							
z-score	--	1.39	-1.18	-0.98	-0.89	-0.84	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

The CEM results for the other two Black-Owens outcomes are also outliers, although less extreme. For Judge Writes Pro-US Decision, all methods produce positive and significant coefficients. All methods other than CEM provide similar coefficients, between 0.57 and 0.68; the CEM coefficient, at 1.16, is almost twice as high (z = 3.74). For Decision Consistent with President's Ideology, most methods produce positive and significant coefficients, but the CEM coefficient of 0.350 is again higher than with any other method, with a substantial z-score (2.84).

B. Decile Analyses

Could the CEM estimates reflect truth, despite being far away from the other estimates? We investigate that question for the first two Black-Owens outcomes, for which the CEM estimate is far from those using other methods. We divide the sample into deciles based on propensity to be treated, running within-decile regressions, and report the within-decile estimates and the average across deciles, weighting each by the number of treated units in each decile $n_{t,d}$. This subclassification approach is recommended by Imbens and Rubin (2015); see also Imbens (2015). Panel A shows the fraction of the treated and control units retained by CEM by decile. A small table under the figure shows the number of treated and control units by decile. CEM generally retains around 30% of treated units, but a lower percentage for deciles 1 and 2 and a higher percentage for deciles 8 and 9. These differences will produce a biased treatment effect estimate if there is treatment heterogeneity across deciles.

In Panel B, we show treatment effect estimates and 95% CIs for each decile for Judge Writes Dissent, from logit regressions similar to those in Table 1. Dotted horizontal lines show the average estimate for all other methods (-0.089) and for CEM (+0.907).[13] There is treatment heterogeneity, which without more, will produce modest upward bias in the CEM estimate. But the actual CEM estimate has deeper problems. The weighted average of the within-subclass estimates should be close to the full sample ATT estimate. This expectation holds for the other balancing methods (not reported), but not for CEM. Instead, the CEM estimate is higher, often much higher, than *any* of the decile estimates. The weighted mean of the decile estimates is 0.027 – effectively zero, and close to the estimates from other methods.

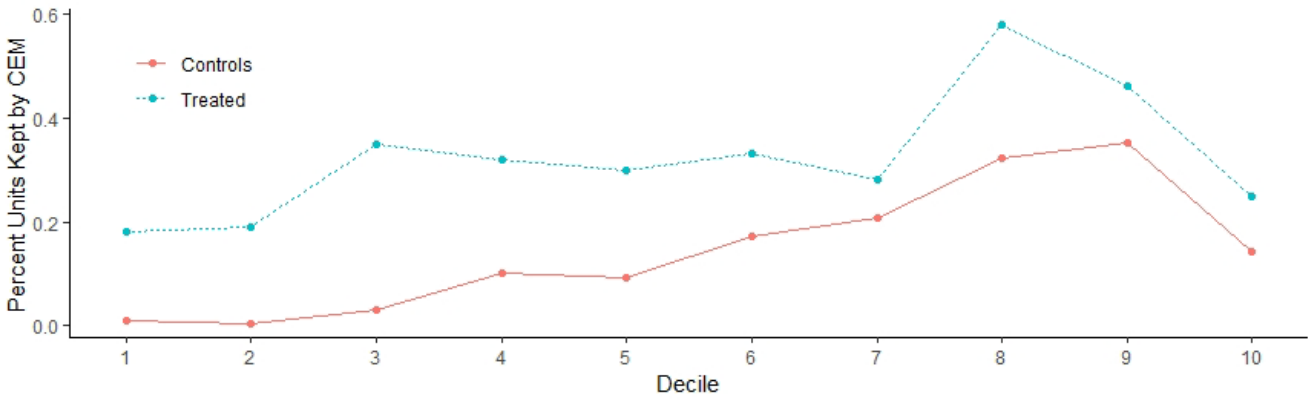
[13] For the top decile, not shown in the graph, the logit regression will not run, but the estimated treatment effect is zero, because the binary outcome is positive for all treated and control units.

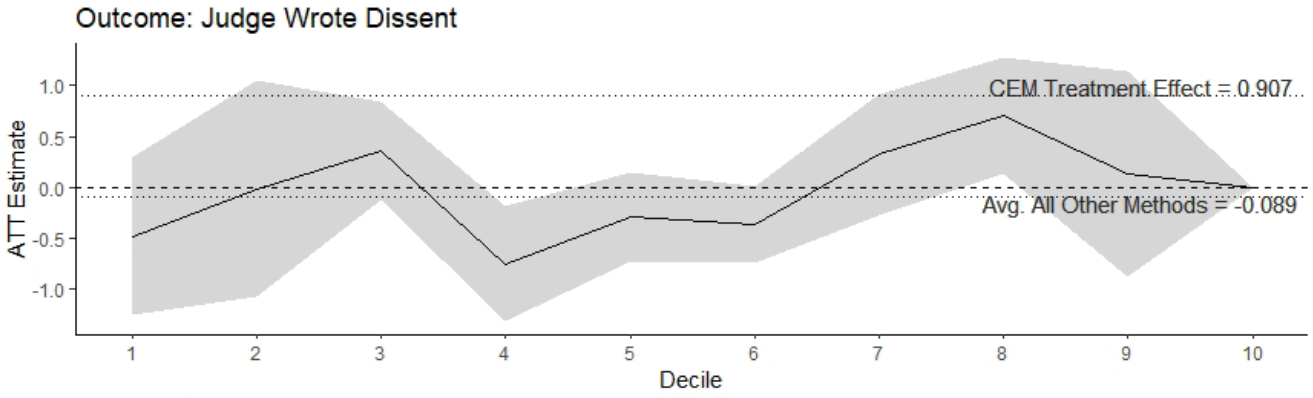
We follow a similar approach in Panel C, for Judge Writes Pro-US Opinion. The CEM estimate is higher than all decile estimates except decile 1, which contains only 1.5% of the treated units. The weighted mean of the decile estimates is 0.236; far below both the CEM overall estimate (1.161) and the average of the estimates from other methods (0.621); and well below the 95% confidence interval (CI) for all methods except the noisiest (nnmatch).

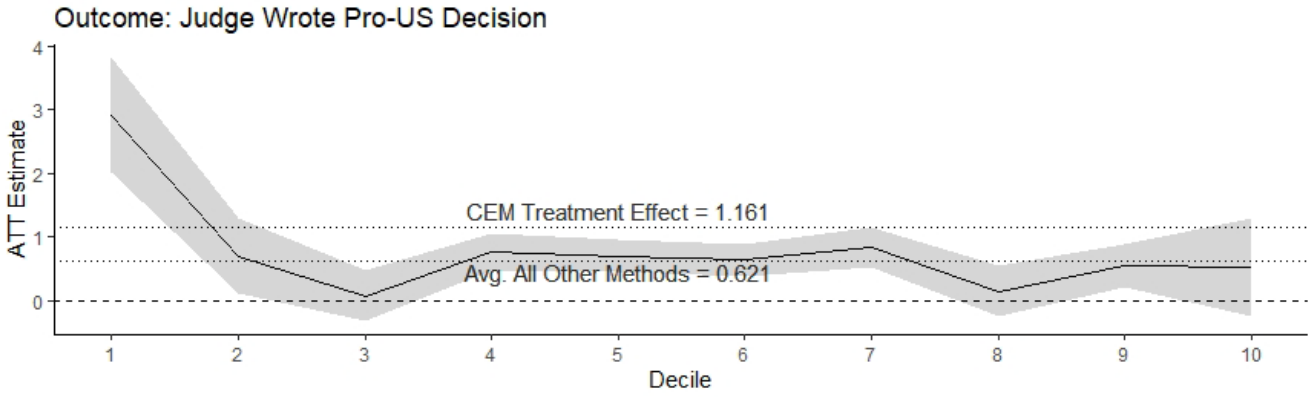
The large departures between the CEM overall estimate and the weighted average of the within-decile estimates provide evidence that the CEM overall estimate is wrong. We find below, for Mason, a similar pattern in which the within-decile estimates are far from the overall CEM estimate.

Figure 5: Black-Owens Analysis of Propensity Score Deciles

Panel A. Percent of treated and control units kept by CEM by propensity score decile. Small table underneath Panel A shows original numbers of treated and control units by decile. **Panel B.** Treatment effects estimated within each decile for Judge Wrote Dissent. **Panel C**. Similar for Judge Wrote Pro-US Opinion. **Panels B and C**. Horizontal lines show estimated treatment effect for CEM and average of all other methods. Shaded areas show 95% confidence intervals.







Treated and Control Units in Each Decile

	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Treated	72	80	199	317	386	640	597	799	863	948
Controls	1107	1099	979	862	793	538	582	380	315	231
Retained by CEM										
Treated	13	15	69	103	117	214	170	466	395	241
Controls	10	6	31	87	74	93	120	122	111	33

C. Varying the Number of Covariates to be Balanced On

We next explore the sensitivity of estimates from the different methods to which variables are balanced on. We again focus on the first two Black-Owens outcomes. To generate Figure 6, we hold constant the second stage regression, which includes all covariates and progressively increase the number of covariates balanced on. For a given number of covariates, we randomly choose which covariates to balance on. For 3 covariates, for example, we randomly select three of the 7 covariates, balance on those, compute ATT estimates for each method, repeat 1,000 times, and show box-and-whiskers plots. These plots show the mean, the box bounds show 25th and 75th percentiles, and whiskers show estimates outside the box.[14]

In Panel A (Judge Writes Dissent), the ATT estimates are not sensitive to the choice of covariates to balance on for the reweighting methods. We see more sensitivity for the matching methods and for CEM. But with all covariates, the ATT estimates with matching eventually settle down at levels close to the reweighting estimates. CEM behaves differently. The range of estimates is unremarkable for 1-3 covariates. But as the number of covariates increases (and the CEM sample therefore progressively shrinks), the CEM estimate departs further and further from the other methods.

In Panel B (Judge Writes Pro-U.S. Opinion), CEM's behavior is unremarkable for 1-5 covariates. But when we add a 6th covariate, the CEM estimate soars to well above the others, and rises further with all 7 covariates. Both this instability, and its tendency to rise as more covariates are balanced on, is a troubling feature of the CEM estimates. The CEM estimates are more model-dependent than at least those using the reweighting methods.[15]

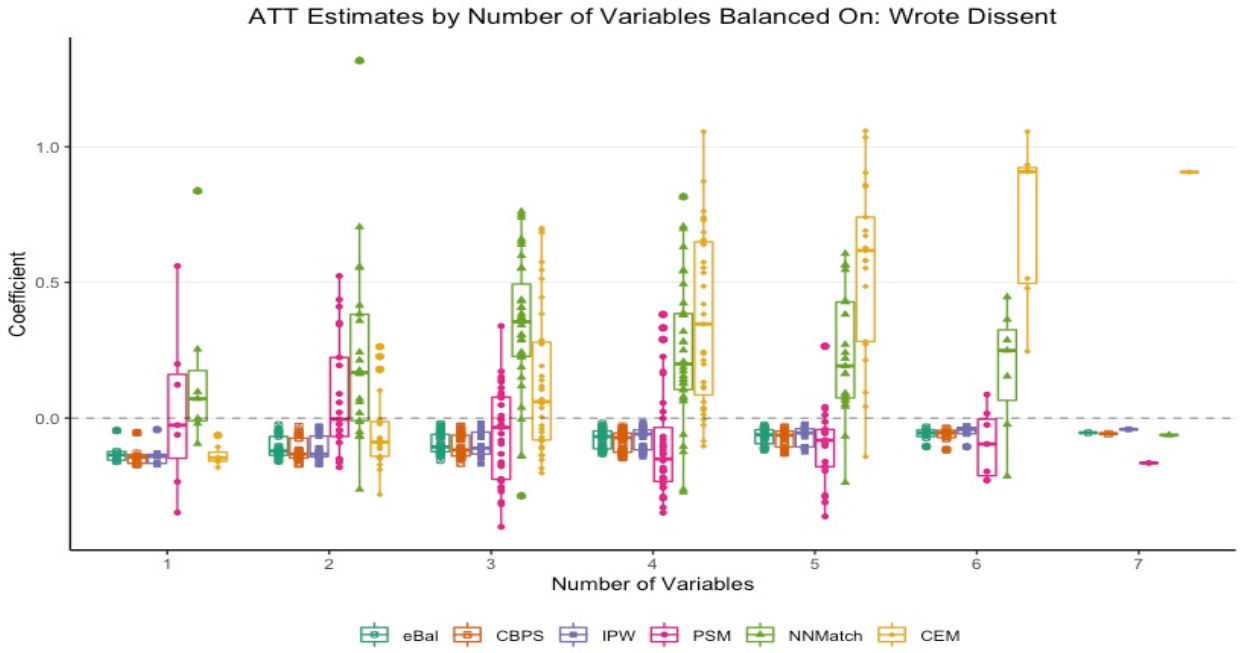
[14] The number of combinations varies with the number of covariates, and is less than 1,000 for 1-4 covariates.

[15] Compare the assertion in King and Nielsen (2019), at xxx **[*add quote; look for something in original CEM paper(s)]**

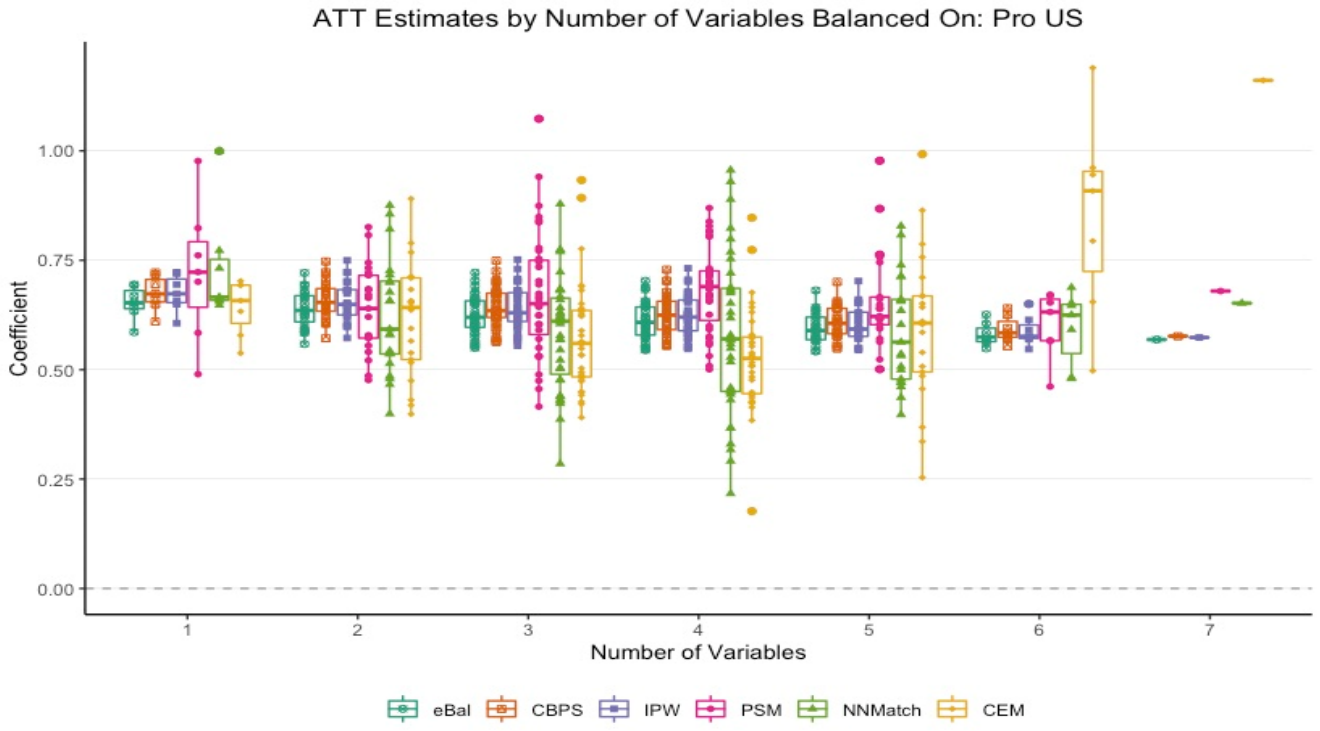
Figure 6: Black- Owens ATT Estimates, Varying the Covariates Balanced on

Figure shows box-and-whiskers plots of ATT estimates from the indicated balancing methods, varying the number of covariates balanced on. Second-stage regression includes all 7 covariates. For each number of covariates, we randomly select this number from the full set of covariates, and iterate 1,000 times. **Panel A.** Dependent variable is Judge Wrote Dissent. **Panel B.** Dependent variable is Judge Wrote Pro-U.S. Opinion.

Panel A. Dependent Variable: Wrote Dissent



Panel B. Dependent Variable: Wrote Pro US decision



D. Sensitivity to Uninformative Covariates

Given the sensitivity of CEM to the variables balanced on, we explored sensitivity to adding additional "noise" variables that are, by construction, uncorrelated with both the treatment and the outcome. The precision of regression estimates might be affected by adding irrelevant variables, but adding them they will not produce bias. One should expect the same for a balancing method.

In Figure 7, we add randomly drawn variables to both the balancing and regression stages, and iterate 1,000 times for each method. For the left-hand part of the figure, we add a random binary variable (mean = 0.5). For the center part, we add a continuous, unit-normal variable; for the right-hand part we add both variables.[16] The figure shows box-and-whiskers plots for each method, showing the *change* in the ATT estimate with the random binary covariate added. We winsorize the graph at ±5, but show in the Appendix a version winsorized at ±20. The center part is similar, but shows the change due to adding the continuous random covariate; the right-hand part shows the change after adding both random covariates. Panel A shows results for the Judge Wrote Dissent. The reweighting methods are minimally affected by adding these variables. The matching methods perform less well. For PSM, some estimates are affected but the largest mean for the change in estimate across these three choices is modest at 0.053 (for one binary and one continuous variable), which is a fraction of the s.e. of 0.186 from Table 1 (larger or ordinary or robust). The nnmatch results are more troubling, with a largest mean change of 0.225 (for one binary variable), close to the Table 1 s.e. of 0.261.

The CEM estimates are especially sensitive to adding noise variables, especially a continuous variable, with apparent bias. For one continuous and one binary variable, the mean change is 1.533 (about 0.75 times the Table 1 s.e.). For one continuous variable, the mean change is 0.580 (s.d. = 1.16), with skewness = 11.05. If we add both, the mean change is still larger at 1.533 (s.d. = 3.82),

[16] We draw the binary variable from the binomial distribution with mean = 0.5 and draw the continuous variable from the normal distribution with standard deviation = 1.

with some very large positive changes (95th percentile =16.4). In the Appendix, we find larger bias and skewness if we add two continuous random variables, or two binary and two continuous random variables.

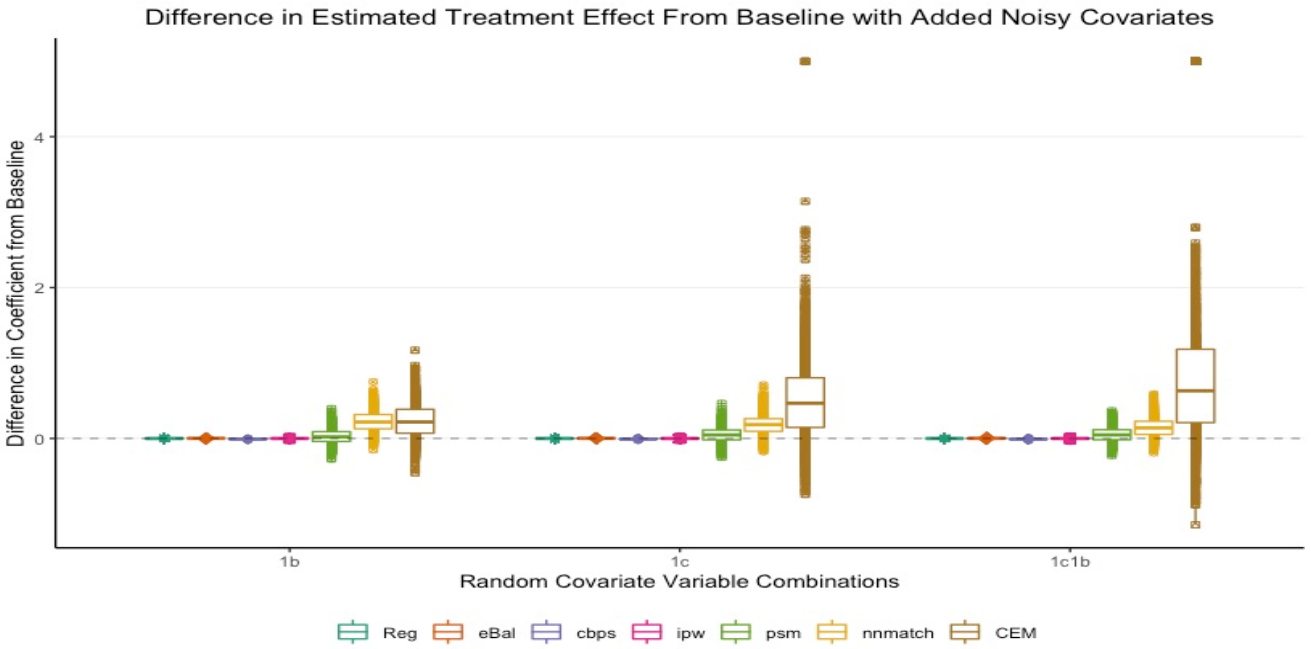
Panel B shows a similar plot for Judge Wrote Pro-U.S. Opinion. The reweighting estimates are again minimally affected. There is noise in the matching but the mean across simulations is similar to the point estimate without the random variable(s). The CEM estimates are strongly affected by adding a random binary variable, a random continuous variable, or one of each.

In the real world, researchers would not knowingly include a truly random variable, but might include a variable that on theoretical grounds might correlate with both treatment and outcome, but in fact is weakly correlated with both. As Figure 7 shows, for CEM, that could produce large bias, which would not be apparent to the researcher. This variable would appear highly relevant because including it strongly affects the treatment effect estimates.

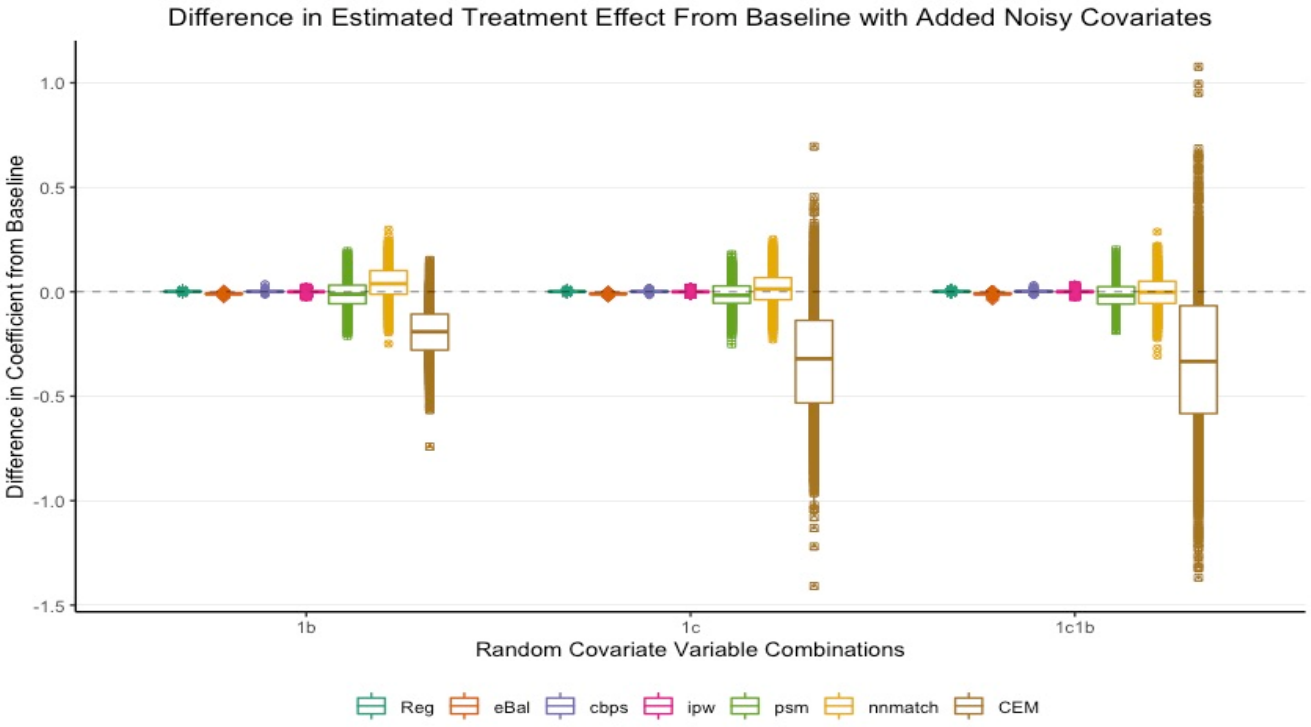
Figure 7: Black- Owens ATT Estimates: Sensitivity to Adding Random Covariates

Figure shows box-and-whiskers plots of ATT estimates from the indicated balancing methods, where we add a random binary variable (mean = 0.5) (left-hand plots), a random, unit normal continuous variable (middle plots), or both together (right-hand plots). Second-stage regression includes both original and added covariates. 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.** Dependent variable is Judge Wrote Dissent. **Panel B.** Dependent variable is Judge Wrote Pro-U.S. Opinion.

Panel A. Dependent Variable: Judge Wrote Dissent



Panel B. Dependent Variable: Judge Wrote Pro-U.S. Opinion



E. Statistical Power and Over-Rejection of the Null

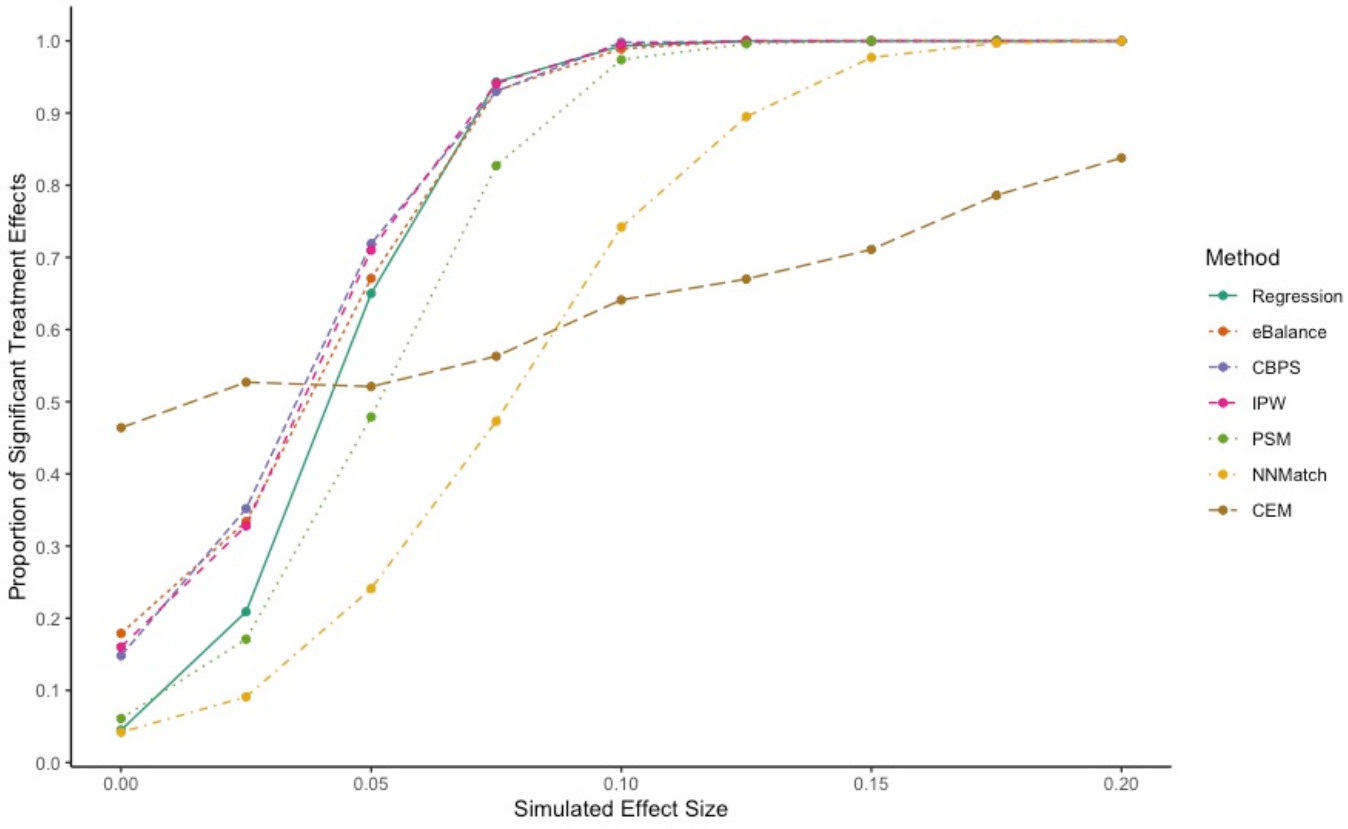
We next investigate the statistical power of each method, and whether the null is over- or under-rejected. We begin with the actual Black-Owens sample, with the actual treatment assignment. We add a simulated, normally distributed outcome, with unit standard deviation and an imposed mean. We begin with an imposed mean of zero (no treatment effect). We run each method, obtain an ATT estimate, and repeat 1,000 times, drawing the outcome at random each time from the unit normal distribution with the specified mean. For each draw we save the coefficient and s.e. We then progressively increase the imposed effect size in increments of 0.025. In Figure 8, we plot the proportion of statistically significant estimate, at the 5% significance level. Figure 8 shows results for the Black-Owens sample.

Consider first an imposed null treatment effect. Regression without balancing performs as it should, with a 5% rejection rate. The matching methods also have rejection rates around 5%. The reweighting methods all over-reject, with rejection percentages around 15%. This is unexpected, for methods which are believed to be doubly robust, although not ruled out by the double-robustness proofs, which address whether estimates are consistent (e.g., Bang and Robins, 2005; Kang and Schaefer, 2007). CEM rejects the null around 45% of the time.

As we increase the imposed treatment effect, regression and the reweighting methods reach nearly 100% power at an imposed effect of 0.1; PSM converges somewhat more slowly; and nnmatch still more slowly, reaching nearly full power more at an imposed effect of 0.15. CEM convergence to full power is much slower. CEM's lower power comes both from a smaller retained sample and greater s.e.'s for the same sample size **[*more to come here]**

Figure 8. Statistical Power for Black-Owens Dataset

Figure shows for each method, statistical power to reject the null at the 95% confidence level, based on 1,000 simulations, for different treatment effects (from 0 to 0.2 in increments of 0.025) imposed on the Black-Owens dataset (4901 treated units; 6886 control units). Treatment effects are drawn from a unit normal variable with imposed mean. For each simulation, we impose the treatment effect, run the balancing method followed by regression, and extract the ATT estimate and s.e.



VI. Analysis of Other Papers

A. Mason (2015)

1. Regression Results

Mason studies voter polarization. She hypothesizes that polarization will increase if voters are "sorted" hold party identification consistent with their ideological views. She studies four outcomes: thermometer bias, like bias, activism, and anger. She tests this hypothesis using CEM, and reports results in her Figure 5, but finds only mild support. All coefficients have the predicted signs, but the coefficient on a sorting dummy (=1 if a sorting measure is roughly above the sample median) is significant only for anger.

In Table 2 we present results for each Mason outcome using OLS alone, our comparison methods, and CEM.[17] The CEM retained sample is much smaller than those for other methods, reflecting Mason's use of 10 covariates (5 are binary and four are categorical). All other methods strongly support Mason's hypothesis, with similar coefficients and small z-scores. The CEM results are very different. The CEM estimates are far below well below those from the other methods for thermometer bias and like bias (z = -5.77 and -9.23, respectively, somewhat below the others for activism (z = -1.56), and above the others for anger (z = 2.75) . CEM is also much less precise, with relative standard errors 3-4 times the average of the other methods.

The combination of mostly lower coefficients and larger s.e.'s strongly weakens inference using CEM. Thermometer bias and like bias have *t*-statistics above 10 for all other methods, yet with CEM thermometer bias is only marginally significant and like bias is insignificant. Activism has *t*-statistics above 5 for the other methods, but is insignificant with CEM. The only outcome which is statistically significant with CEM is for anger, but inference is much stronger with the other methods, despite lower point estimates.

[17] We have substantive concerns with Mason's approach. In Figure 5, she does not control for a variable she calls "ideological strength," which is a core component of her sorting variable. We put those concerns aside here and focus on the implications of different balancing methods, assuming her substantive model is appropriate.

Table 2. Mason Results with Different Balancing Methods

Last column provides CEM plus OLS regression results for indicated outcomes, comparable to Mason (2015), Figure 5, for indicated outcomes. Mason's figure uses CEM without regression, Other columns use indicated balancing method followed by OLS regression We use Mason's exact sample (she requires data for all outcomes and covariates). Sorting dummy = 1 is 1 if idcomplexity ≥ 0.15; which is slightly below the sample median of 0.162. Balancing and regression use the following covariates: partisan strength (we divide Mason's categorical variable into dummy variables for the four levels); Issue Position Extremity, Education, Male, White, Age, South, Urban, Church Attendance, and Evangelical. Coefficients on covariates and constant term are suppressed. Ordinary standard errors (s.e.'s) in brackets; robust standard errors and randomization inference standard deviations (s.d.'s), based on 1,000 draws, in parentheses. *, indicates significance at the 5% level, baswed on randomization inference s.d.'s. Significant results, at .05 level or better, in **boldface**. z-score and s.e. ratio are defined in Table 1.

	(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. [ordinary]	[0.00487]	[0.00406]	[0.00407]	[0.00462]	[0.00540]	[0.00439]	[0.0165]
s.e. (robust)	(0.00493)	(0.00407)	(0.00406)	(0.00536)	(0.00540)	(0.00533)	(0.0170)
random inf. s.d.							
z-score	--	0.95	0.42	0.80	0.87	0.82	-5.77
s.e. ratio	--	0.55	0.55	0.75	0.75	0.74	3.51
Panel B. Like Bias							
Sorting dummy	**0.0483***	**0.0610***	**0.0573***	**0.0558***	**0.0570***	**0.0561***	0.0166
s.e. [ordinary]	[0.00457]	[0.00373]	[0.00376]	[0.00428]	[0.00492]	[0.00407]	[0.0168]
s.e. (robust)	(0.00441)	(0.00373)	(0.00375)	(0.00485)	(0.00492)	(0.00486)	(0.0175)
random inf. s.d.							
z-score	--	1.73	1.12	0.89	1.10	0.94	-9.23
s.e. ratio	--	0.52	0.52	0.70	0.71	0.70	3.96
Panel C. Activism							
Sorting dummy	**0.0283***	**0.0323***	**0.0329***	**0.0314***	**0.0320***	**0.0317***	0.0248
s.e. [ordinary]	[0.00453]	[0.00381]	[0.00374]	[0.00433]	[0.00533]	[0.00413]	[0.0144]
s.e. (robust)	(0.00436)	(0.00381)	(0.00374)	(0.00521)	(0.00533)	(0.00523)	(0.0152)
random inf. s.d.							
z-score	--	0.25	0.35	0.10	0.21	0.15	-1.56
s.e. ratio	--	0.55	0.54	0.78	0.80	0.79	3.26
Panel D. Anger							
Sorting dummy	**0.0812***	**0.0814***	**0.0717***	**0.0822***	**0.0826***	**0.0821***	**0.108***
s.e. [ordinary]	[0.0100]	[0.00851]	[0.00853]	[0.00969]	[0.0114]	[0.00924]	[0.0379]
s.e. (robust)	(0.0103)	(0.00850)	(0.00853)	(0.0112)	(0.0114)	(0.0113)	(0.0397)
random inf. s.d.							
z-score	--	-0.24	-0.95	-0.19	-0.16	-0.19	2.75
s.e. ratio	--	0.52	0.52	0.70	0.72	0.71	3.90
observations	9858	7890	8009	8970	9858	9858	572
treated (control)	5802 (4056)	5802 (2088)	5802 (2207)	5802 (3168)	5802 (4056)	5802 (4056)	294 (278)
Effective controls				3023	2966	3000	254

1. Decile Analysis

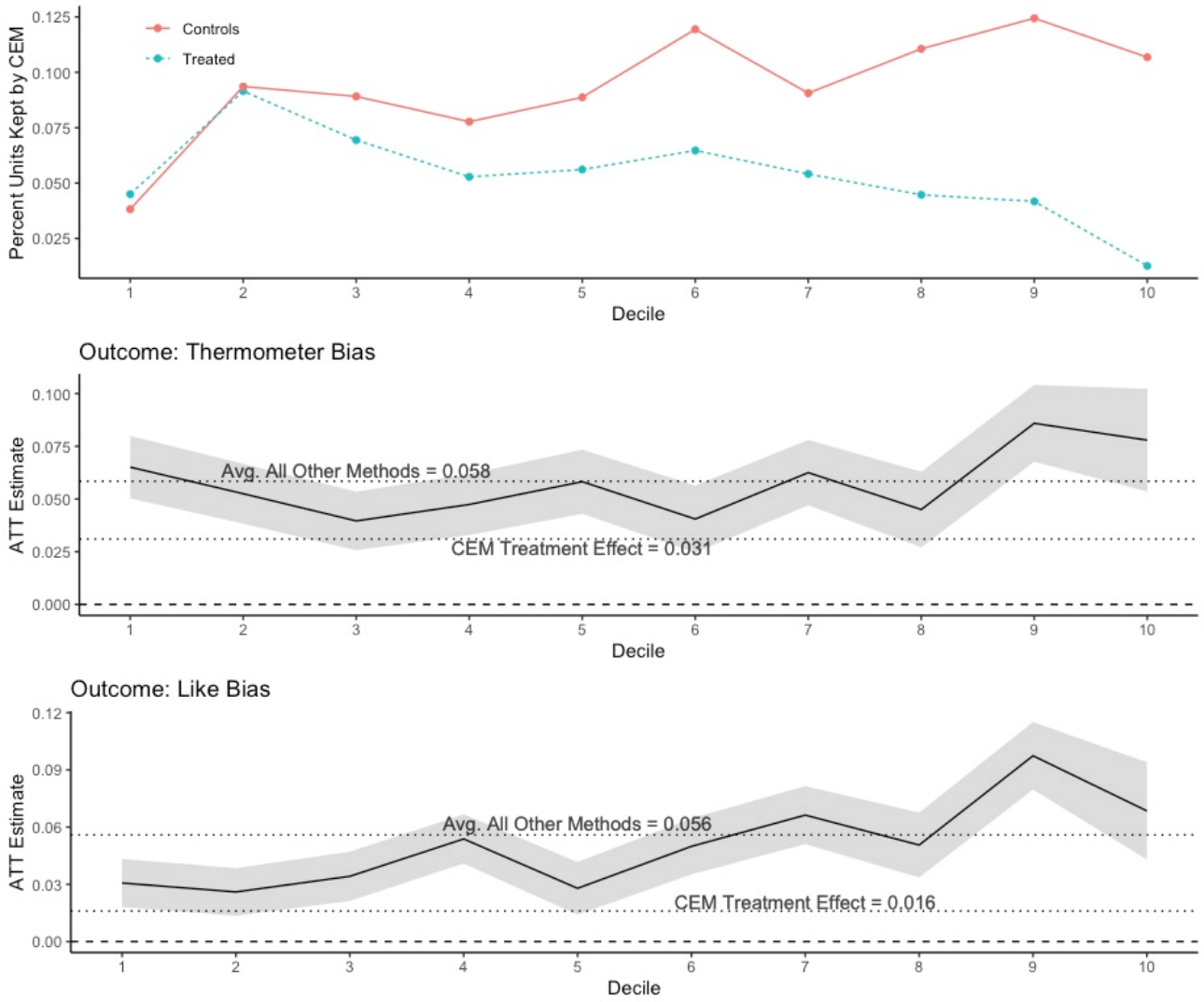
Figure 9 provides a decile analysis of the first two Mason results, similar to that presented above for Black-Owen. Panel A shows the fraction of treated and control units retained by CEM by propensity score decile. The percentage of treated units retained by CEM averages around 5%, but is markedly higher for decile 2 and only 1.3% for decile 10. These differences will produce biased estimates if there is treatment heterogeneity across deciles.

In Panel B, we show treatment effect estimates and 95% CIs by decile for Thermometer Bias. Dotted horizontal lines show the average estimate for all other methods (+0.058) and for CEM (+0.301). There is treatment heterogeneity, with the estimate highest for deciles 9 and 10. But the actual bias is far more than can be explained by treatment effect heterogeneity. The CEM estimate is lower than *any* decile estimate. Similarly, in Panel C, the CEM estimate for Like Bias is lower than each decile estimate.

For both outcomes, the mean of the CEM decile estimates, weighted by the number of retained treated units, is closer to the estimates from other methods than the CEM full-sample estimate. For Thermometer Bias: CEM weighted mean = 0.038 vs. mean from other methods of 0.058 and CEM estimate of 0.031. For Like Bias: CEM weighted mean = 0.034 vs. mean from other methods of 0.056 and CEM mean of 0.015.

Figure 9: Mason Propensity Decile Comparison

Panel A: Percent of treated and control units kept by CEM by propensity score decile. Small table underneath Panel A shows original numbers of treated and control units by decile. **Panels B & C:** Treatment effect estimated within each decile using logistic regression. Horizontal lines show estimated treatment effect for CEM and for average of all other methods.



Treated and Control Units in Each Decile

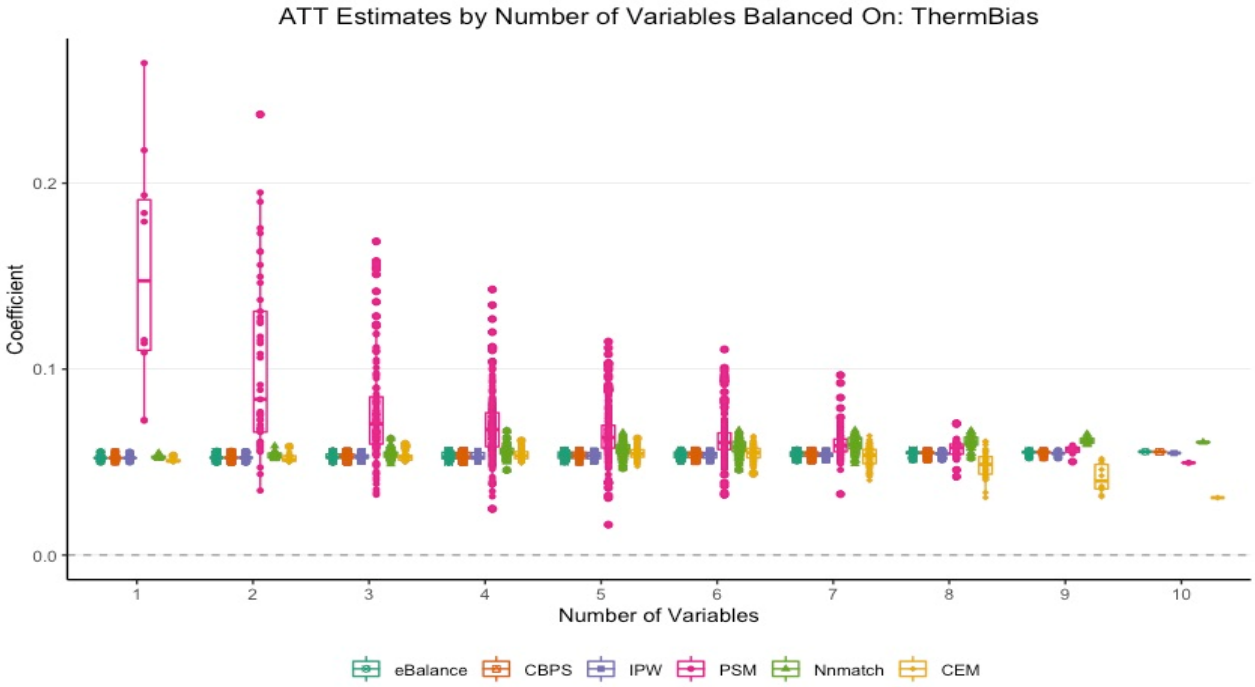
	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10
Treated	400	459	504	511	570	587	610	671	696	794
Controls	497	438	393	386	327	310	287	226	201	103
Retained by CEM										
Treated	18	42	35	27	32	38	33	30	29	10
Controls	19	41	35	30	29	37	26	25	25	11

Figure 10 assesses sensitivity to the covariates balanced on. The structure of this figure is similar to Figure 6 above. The reweighting methods are again insensitive to how many covariates are balanced on, and which ones. This time, the PSM estimates are especially sensitive to the covariates balanced on for small numbers of covariates but converge to the other estimates as the number of covariates balanced on increases. The nnmatch estimates are consistently a bit higher than those from regression and the reweighting methods but not sensitive to number of covariates. CEM appears stable, and similar to other estimates, when balancing on 1-7 covariates, but increasingly departs from the other estimates as we add additional covariates.

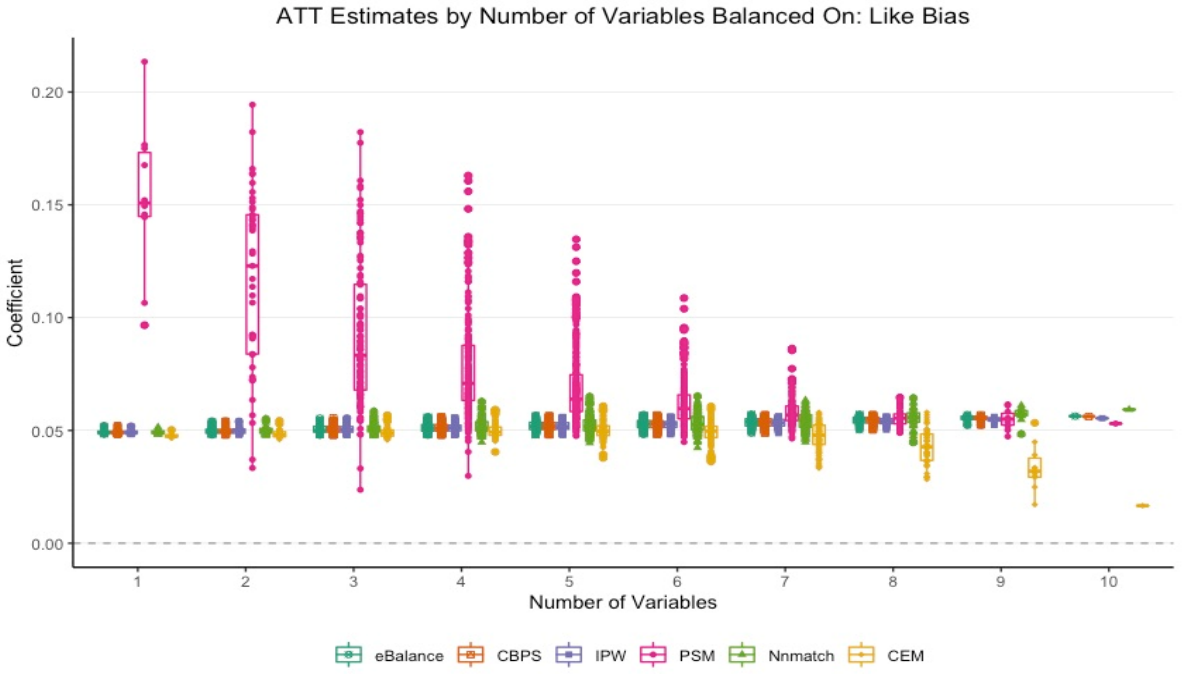
Figure 10: Mason ATT Estimates, Varying the Covariates Balanced on

Figure shows box-and-whiskers plots of ATT estimates from the indicated balancing methods, varying the number of covariates balanced on. Second-stage regression includes all 10 covariates. For each number of covariates, we randomly select this number from the full set of covariates, and iterate 1,000 times. **Panel A.** Dependent variable is Thermometer Bias. **Panel B.** Dependent variable is Like Bias. All 10 covariates are used in the regression stage.

Panel A Dependent Variable: Thermometer Bias



Panel B. Dependent Variable: Like Bias

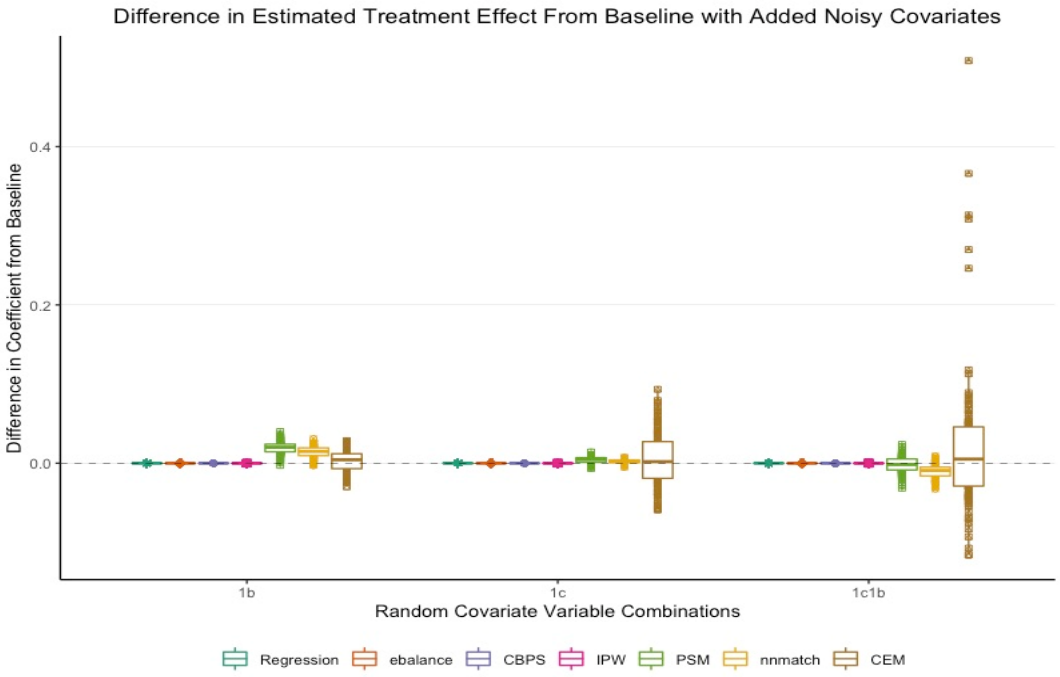


In Figure 11, we assess the sensitivity of each method to adding a random binary covariate, a random continuous covariate, or one of each. Regression and the reweighing methods are minimally affected. For the random binary covariate, the matching methods show both spread in estimates and apparent bias, CEM shows spread but little bias. For the random continuous covariate, CEM's spread is much larger than for the other methods. When we add both a random binary and a random continuous variable, the CEM spread becomes huge, relative to the point estimates. In an actual study, where one has only a single draw, the spread in estimates, even without bias, could produce treatment effect estimates that are far from truth.

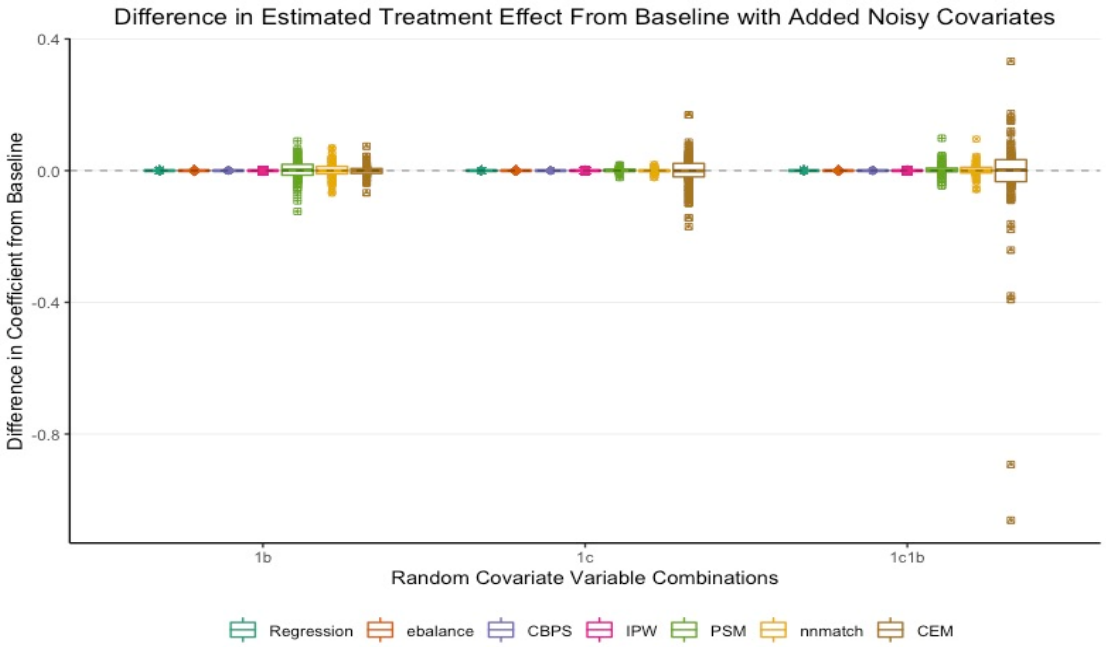
Figure 11: Mason ATT Estimates: Sensitivity to Adding Random Covariates

Figure shows box-and-whiskers plots of ATT estimates from the indicated balancing methods, where we add a random binary variable (mean = 0.5) (left-hand plots), a random, unit normal continuous variable (middle plots), or both together (right-hand plots). Second-stage regression includes both original and added covariates. We draw each random covariate from the corresponding distribution and iterate 1,000 times. **Panel A.** Dependent variable is Thermometer Bias. **Panel B.** Dependent variable is Like Bias.

Panel A. Dependent Variable: Thermometer Bias



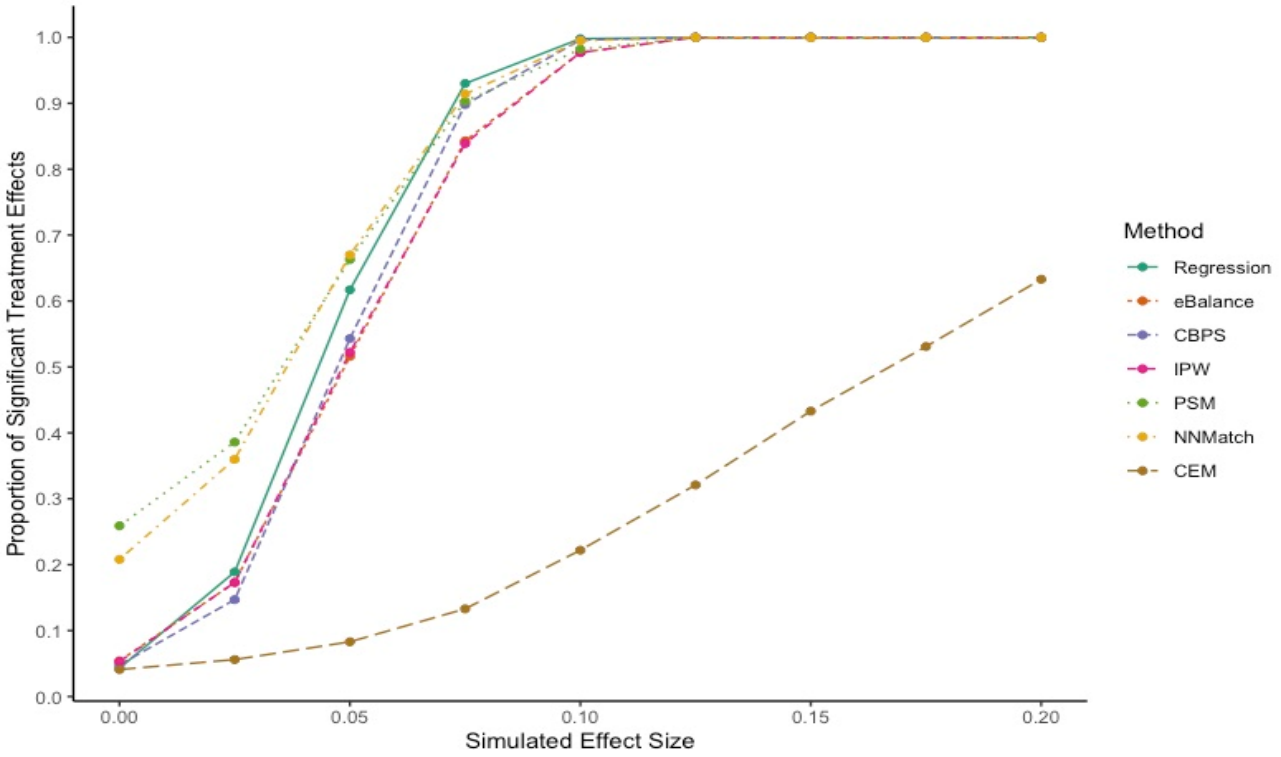
Panel B. Dependent Variable: Like Bias



In Figure 12, we impose artificial treatment effects to the Mason data, using the same approach as in Figure 8. This time, in contrast to the Black-Owens results in Figure 8, 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 much less power than the other methods. All other methods have nearly complete power to detect an 0.1 effect; but at this effect size, CEM remains severely underpowered.

Figure 10. Statistical Power

Figure shows for each method, statistical power to reject the null at the 95% confidence level, based on 1,000 simulations, for different treatment effects (from 0 to 0.2 in increments of 0.025) imposed on the Mason dataset ((5802 treated units; 4056 control units). Treatment effects are drawn from a unit normal variable with imposed mean. For each simulation, we impose the treatment effect, run the balancing method followed by regression, and extract the ATT estimate and s.e.



B. Urban and Niebler (2014)

We discuss Urban-Niebler and Carpenter more 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-1 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 amount contributed conditional on a contribution being made. CEM is again an outlier. The estimated effect of spillover ads on amount contributed (given that a contribution is made) is positive and significant across all five other balancing methods, marginally significant with regression alone, but near zero and insignificant with CEM ($z = -3.16$).

C. Carpenter et al. (2012)

Carpenter et al. are interested in how administrative deadlines shape decision timing and the quality of the decisions made. They study FDA drug approvals, find that administrative deadlines for decisions induce many decisions made by examiners just before the deadlines, and that the "just-before-deadline" approvals are associated with higher rates of future safety problems (severe, "black box" safety warnings; safety-based withdrawal of a drug from the market; and less-severe safety alerts). We present results in Appendix Table App-2 for Black Box Warning, Safety-Based Withdrawal, and Safety Alert.

For all three outcomes, the reweighting estimates are similar, statistically significant, and similar to estimates from regression alone. In contrast, the matching methods, including CEM, produce more varied estimates, larger s.d.'s, and insignificant results. CEM keeps only 35 of the 86 treated units, leading to higher s.d.'s than other methods. Relative to the other matching methods, CEM does not provide outlier estimates with z-scores above 2, but it remains a poor choice because it discards most of the treated units from an already small sample, with resulting lower precision and the potential for biased

estimates. The large differences between reweighting and matching methods suggest that there is value in using both broad approaches.

VII. Discussion: Why Does CEM produce Odd Results?

Our major takeaways are as follows. First, CEM can produce very different results than the other methods, and can produce full sample results that are inconsistent with subsample results. While it provides reasonable covariate balance, this comes at the cost of much smaller retained samples than other methods, and thus lower precision. If one limits the number of variables to be balanced on to preserve sample size, one loses covariate balance for the variables not balanced on and can obtain widely varying estimates, depending on the variables balanced on (Figures 6, 10). If CEM had a consistent direction of bias—if, for example, estimates were always closer to zero, it might still a useful robustness check. However, CEM is consistently inconsistent: CEM estimates can be much farther from zero than all other estimates (for Black-Owens) but can also be near-zero when all other methods produce statistically significant estimates (Mason).

Our blunt judgment: CEM should never be used as a primary balancing method, and it is unclear why researchers should use it at all. There are other, better methods available. We pursue a more careful comparison of the other methods in separate work (Black, Lalkiya, and Lerner, 2022), but note that for the methods and comparison papers we use here, eBalance performed well across the board. The matching methods (PSM and nnmatch) performed worse than the reweighting methods (consistent with the concerns about PSM in King and Nielsen, 2019), but substantially better than CEM.

If very close covariate balance is important, other methods can achieve this without the loss of sample size or CEM's other quirks. We studied one exact balancing method (eBalance) and one near-exact method (CBPS with weights), but other methods also exist, including Jose Zubizarreta's stable balancing weights, Jas Sekhon's genetic matching, and Bryan Graham's inverse propensity tilting. If

easy implementation is important, eBalance is available in both Stata and R. Perhaps eBalance will perform oddly in other datasets, but the only warning signs in this project was over-rejection of the null when we imposed an artificial treatment effect on the Black-Owens data. For R users, CBPS with weights is a reasonable alternative. If one is willing to accept somewhat worse covariate balance, IPW is implemented in both Stata and R and performed reasonably both for us and Busso, DiNardo and McCrary (2014).

References

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

Abadie, Alberto, and Jann Spiess (2021), Robust Post-Matching Inference, *Journal of the American Statistical Association* xxx, yyy-zzz.

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, Parth Lalkiya, and Joshua Lerner, Matching Strategies Compared: Assessing How Results in Observational Studies Vary across Balancing Methods (working paper 2022), at http://ssrn.com/abstract=3xxxxxx.

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, xx *Statistics in Medicine* yyy-zzz.

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.

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 (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* (2012): 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.

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

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.

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

Figure XX: Simulating Sample Loss from CEM by Number of Covariates

Below, we simulate (with 1,000 draws) the proportion of the sample remaining after using CEM for matching, as a function of the number of covariates used for matching, for sample sizes of 250, 500, 1,000, and increments of 500 after that, up to 10,000. The top graph shows the proportion remaining if covariate values are drawn from the normal distribution; the bottom shows the proportion remaining if the covariate values are drawn from the uniform distribution. With the more severe challenge (the uniform distribution), severe sample loss sets in with four covariates, and almost no sample is left with 5 or more covariates, even for very large samples. CEM does better at handling covariates drawn from the normal distribution, but even so, there is major sample loss with five covariates, and essentially no sample is left with seven covariates.