Supplementary materials for "Varying impacts of letters of recommendation on college admissions"

A Additional analyses and robustness checks

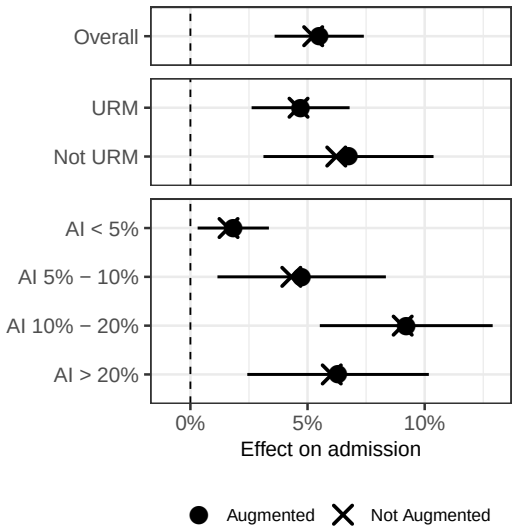
This appendix section addresses how our conclusions change with different estimation approaches, different data and sample definitions, and under violations of the key ignorability assumption.

A.1 Augmented and machine learning estimates

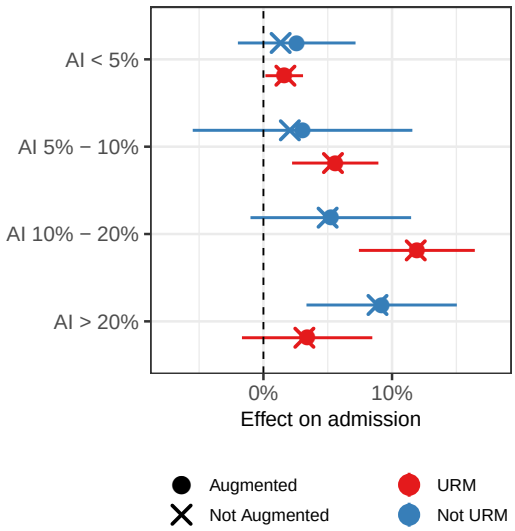
We now consider augmenting the weighting estimator with an estimate of the prognostic score, $\hat{m}(x, g)$. In Appendix Figure D.11 we show estimates after augmenting with ridge regression, fully interacting $\phi(X)$ with the strata indicators; we compute standard errors via Equation (15), replacing $Y_i - \hat{\mu}_{0g}$ with the empirical residual $Y_i - \hat{m}(X_i, g)$. Because the partially pooled balancing weights achieve excellent local balance for $\phi(X)$, augmenting with a model that is also linear in $\phi(X)$ results in minimal adjustment. We therefore augment with random forests, a nonlinear outcome estimator. Tree-based estimators are a natural choice, creating "data-dependent strata" similar in structure to the strata we define for G. For groups where the weights $\hat{\gamma}$ have good balance across the estimates $\hat{m}(x, g)$, there will be little adjustment due to the outcome model. Conversely, if the raw and bias-corrected estimate disagree for a subgroup, then the weights have poor local balance across important substantive data-defined strata. For these subgroups we should be more cautious of our estimates.

Figure D.10 shows the random forest-augmented effect estimates relative to the un-augmented estimates; the difference between the two is the estimated bias. Overall, the random forest estimate of the bias is negligible and, as a result, the un-adjusted and adjusted estimators largely coincide. Augmentation, however, does seem to stabilize the higher-order interaction between AI and URM status, with particularly large adjustments for the highest AI group (AI $\geq 20\%$). This suggests that we should be wary of over-interpreting any change in the relative impacts for URM and non-URM applicants as AI increases.

Finally, we use Bayesian causal forests (BCF; Hahn et al., 2020) to estimate the conditional average treatment effect given the covariates and subgroup, $\hat{\tau}(x, g)$, then aggregate over the treated units in the group to estimate the CATT, $\hat{\tau}_g = \frac{1}{n_{1g}} \sum_{G_i=g} W_i \hat{\tau}(X_i, G_i)$. This approach gives no special consideration to the subgroups of interest: G enters symmetrically with the other covariates X. Appendix Figure D.15 shows the results. The BCF estimate of the overall ATT is nearly



(a) Overall and by URM status and AI.



(b) By URM status interacted with AI.

Figure A.1: Estimated effect of letters of recommendation on admission rates with and without augmentation via a random forest outcome model.

the same as our main estimates and similarly finds no variation between URM and non-URM applicants. However, the BCF estimates find less heterogeneity across admissibility levels and little to no heterogeneity across URM and admissibility subgroups, in part because this approach regularizes higher-order interactions in the CATE function. While this can be beneficial in many cases, with pre-defined subgroups this can lead to over-regularization of the effects of interest, as happens here. Furthermore, the BCF estimates are extremely precise, even in regions with limited overlap (see discussion in Hahn et al., 2020). Overall, we find this approach less credible in our setting with pre-defined subgroups of interest.

A.2 Alternative data and sample definitions

Here we consider alternative data and sample definitions as well as additional data sources. First, we consider effects on an intermediate outcome: whether the second reader — who has access to the LORs — gives a "Yes" score. Because these are *design-based* weights, we use the same set of weights to estimate effects on both second reader scores and admissions decisions. We find a similar pattern of heterogeneity overall.

With this outcome we can also make use of a within-study design to estimate treatment effects, leveraging scores from additional third readers who did not have access to the letters of recommendation. After the admissions process concluded, 10,000 applicants who submitted letters were randomly sampled and the admissions office recruited several readers to conduct additional evaluations of the applicants (Rothstein, 2017). During this supplemental review cycle, the readers were *not* given access to the letters of recommendation, but otherwise the evaluations were designed to be as similar as possible to the second reads that were part of the regular admissions cycle; in particular, readers had access to the first readers' scores.

With these third reads we can estimate the treatment effect by taking the average difference between the second read (with the letters) and the third read (without the letters). One major issue with this design is that readers might have applied different standards during the supplemental review cycle. Regardless, if the third readers applied a different standard consistently across URM and admissibility status, we can distinguish between treatment effects within these subgroups.

Appendix Figures D.13 and D.14 show the results for both approaches. Overall for second reader scores we see a similar structure of heterogeneity as for admission rates, although there does not appear to be an appreciable decline in the treatment effect for the highest admissibility non-URM applicants. The two distinct approaches yield similar patterns of estimates overall, with the largest discrepancy for applicants with AI scores between 5% and 10%, particularly for non-URM applicants. However, this group has a very low effective sample size, and so the weighting estimates are very imprecise.

Finally, recall that an applicant may not have submitted an LOR for one of two reasons: (i) they were not invited to do so, and (ii) they did not submit even though they were invited. We assess the sensitivity of our results to excluding this first group when using the weighting approach. Appendix Figure D.16 shows the estimated effects with the full sample and restricted to applicants who were invited to submit an LOR. The point estimates are similar, although the estimated effect on non-URM low admissibility applicants is higher in the restricted sample. Additionally, although the number of control units is much smaller in the restricted sample — 3,452 vs 29,398 — the standard errors are only slightly larger. This reflects the fact the weighting approach finds that few of the no-invitation control units are adequate comparisons to the treated units.

A.3 Formal sensitivity analysis

We assess sensitivity to the key assumption underlying our estimates, Assumption 1: an applicant's LOR submission is conditionally independent of that applicant's potential admission decision. Since we observe all the information leading to an invitation to submit an LOR, we believe that Assumption 1 is plausible for this step in the process. However, applicants' decisions to submit LORs given that invitation might vary in unobserved ways that are correlated with admission.

To understand the potential impact of such unmeasured confounding, we perform a formal sensitivity analysis. Following the approach in Zhao et al. (2019); Soriano et al. (2020), we allow the true propensity score conditioned on the control potential outcome $e(x, g, y) \equiv P(W = 1 \mid X = x, G = g, Y(0) = y)$ to differ from the probability of treatment given covariates x and group membership g, $e(x, g)$, by a factor of Λ in the odds ratio:

$$\Lambda^{-1} \leq \frac{e(x,g)/1-e(x,g)}{e(x,g,y)/1-e(x,g,y)} \leq \Lambda. \tag{1}$$

This generalizes Assumption 1 to allow for a pre-specified level of unmeasured confounding, where $\Lambda = 1$ corresponds to the case with no unmeasured confounding. The goal is then to find the smallest and largest ATT, $[\tau^{\min}, \tau^{\max}]$, consistent with a given Λ for the marginal sensitivity model in Equation (1). Following Soriano et al. (2020), we use the percentile bootstrap to construct a 95% confidence for this bound, $[L, U]$. In particular, we focus on the largest value of Λ for which the overall ATT remains statistically significant at the 95% level (i.e., $L > 0$), which we compute as $\Lambda = 1.1$.

We then modify this approach to focus on subgroup differences; we focus on differences between URM and non-URM applicants but can apply this to other subgroups as well. First, we use the same

procedure to find bounds on the effect for URM applicants, $[L^{\text{urm}}, U^{\text{urm}}]$, and non-URM applicants, $[L^{\text{non}}, U^{\text{non}}]$. We then construct worst-case bounds on their *difference*, $[L^{\text{urm}} - U^{\text{non}}, U^{\text{urm}} - L^{\text{non}}]$. Although we fail to detect a difference between the two groups, there may be unmeasured variables that confound a true difference in effects. To understand how large the difference could be, we use the sensitivity value that nullifies the overall effect, $\Lambda = 1.1$ and construct a 95% confidence interval for the difference. We find that we cannot rule out a true difference as large as 12 pp, with 95% confidence interval $(-12\%, 8.2\%)$.

To understand this number, Appendix Figure D.17 shows the strength required of an unmeasured confounder in predicting the admission outcome (measured as the magnitude of a regression of the outcome on the unmeasured confounder) to produce enough error to correspond to $\Lambda = 1.1$, for a given level of imbalance in the unmeasured confounder between applicants who did and did not submit LORs.[1] It compares these values to the imbalance in the components of $\phi(X)$ *before weighting* and the regression coefficient for each component where we regress the outcome on $\phi(X)$ for the control units. We find that the unmeasured confounder would have to have a higher level of predictive power or imbalance than any of our transformed covariates, except for the AI. Thus, an unmeasured confounder would have to be relatively strong and imbalanced in order to mask a substantial difference between URM and non-URM applicants.

A.4 Simulation of universal LOR policy

Our main analysis considers the impact of submitting LORs for each applicant in isolation, following the structure of the UC Berkeley pilot study. We now use these estimates to conduct a simple policy simulation for the impact of requiring LORs for *all* UC Berkeley applicants on the composition of the admitted class, relative to the current policy of no LORs. We begin with the same sample as our previous analyses; in this exercise, however, we take explicit account of capacity constraints. The exercise has three basic steps: (1) use our prior estimates to predict the admissions probability for each applicant *in isolation* if no applicants submitted LORs and if all applicants submitted LORs;[2] (2) adjust these probabilities to set the expected number of admitted students under each policy to the observed number;[3] and (3) use the adjusted probabilities to construct an "expected" admitted class under each policy.

This policy simulation is necessarily simplistic, and ignores many complications, such as a possible change in the composition of the applicant pool as a result of requiring LORs. Nonetheless, this exercise captures the important role of an overall cap on undergraduate admissions, which does not affect our individual estimates above: while we estimate that LORs increase the average applicant's admissions probability by around 5 percent in isolation, the feasible policy cannot also increase the total number of all undergraduates by 5 percent. Thus, our focus above on the

[1] Letting δ denote this imbalance, the absolute regression coefficient must be larger than $\max\{|L^{\text{urm}} - U^{\text{non}}|, |U^{\text{urm}} - L^{\text{non}}|\}/\delta = 0.592/\delta$.

[2] We first use the predictions of the baseline probability of admission without LORs using the random forest estimator from Appendix A.1. We then add the estimated treatment effects from the right panel in Figure 5 to the probability of admission for each applicant in the 2016 cohort, according to their subgroup. We assume here that the subgroup ATT equals the subgroup Average Treatment Effect, $\mathbb{E}[Y(1) - Y(0) \mid G = g, W = 1] = \mathbb{E}[Y(1) - Y(0) \mid G = g]$, for each URM $\times$ AI subgroup.

[3] A naive adjustment can lead to estimated probabilities outside $[0, 1]$. To account for this, we iteratively subtract the average difference between the expected and simulated admissions rates and then truncate the probabilities to be between 0 and 1. We repeat these centering and truncation steps until all estimates are in $[0, 1]$ and the expected number of admits equals the observed number of admits for our subsample, 6,874.

differential impacts of LORs is especially relevant for this exercise; see also Section B.1.

Appendix Figure D.18 shows the simulated effect of the LOR policy on the number of admitted applicants in each subgroup. Consistent with our point estimates, we find a negligible overall impact on URM applicants, with roughly 90 fewer URM admitted applicants overall (less than half a percent of all URM applicants in this sample). We see larger (if still small) changes across the Admissibility Index: more medium and high admissibility applicants are admitted, offset by fewer low admissibility applicants, both URM and non-URM. Thus, based on this back-of-the-envelope simulation, we find that requiring LORs for all applicants would have minimal effects on the URM composition of admitted students, while raising the number of applicants with traditional qualifications.

B Additional analytic results

B.1 Balance when estimating the difference in effects

Here we consider targeting the difference in treatment effects for groups g and g'. Because we have access to the treated potential outcomes for treated units, we will focus on the difference in the counterfactual means, $\tilde{\mu}_{0g} - \tilde{\mu}_{0g'}$. Under the linear model in Section 4.1, the estimation error for the difference is

$$\begin{aligned}
\hat{\mu}_{0g} - \hat{\mu}_{0g'} - (\tilde{\mu}_{0g} - \tilde{\mu}_{0g'}) = \bar{\eta}_{gg'} \cdot \Bigg(& \sum_{i=1}^{n} \hat{\gamma}_i (1 - W_i) \left[\frac{1}{n_{1g}} \mathbb{1}\{G_i = g\} - \frac{1}{n_{1g'}} \mathbb{1}\{G_i = g'\} \right] \phi(X_i) \\
- & \sum_{i=1}^{n} W_i \left[\frac{1}{n_{1g}} \mathbb{1}\{G_i = g\} - \frac{1}{n_{1g'}} \mathbb{1}\{G_i = g'\} \right] \phi(X_i) \Bigg) \\
+ & \left(\eta_g - \bar{\eta}_{gg'}\right) \cdot \left(\sum_{G_i = g} \hat{\gamma}_i (1 - W_i) \phi(X_i) - \frac{1}{n_{1g}} \sum_{G_i = g} W_i \phi(X_i) \right) \\
+ & \left(\eta_{g'} - \bar{\eta}_{gg'}\right) \cdot \left(\sum_{G_i = g'} \hat{\gamma}_i (1 - W_i) \phi(X_i) - \frac{1}{n_{1g}} \sum_{G_i = g} W_i \phi(X_i) \right) \\
+ & \frac{1}{n_{1g}} \sum_{G_i = g} (1 - W_i) \hat{\gamma}_i \varepsilon_i \quad + \quad \frac{1}{n_{1g'}} \sum_{G_i = g'} (1 - W_i) \hat{\gamma}_i \varepsilon_i,
\end{aligned} \tag{2}$$

where $\bar{\eta}_{gg'} = \frac{1}{2}(\eta_g + \eta_{g'})$.

From this, we see that the estimation error now includes a combined measure of imbalance (the first term) as well as the level of local balance in subgroups g and g' (the second two terms). However, this is the imbalance in a transformed set of covariates:

$$\tilde{\phi}_{gg'}(X_i) \equiv \left[\frac{1}{n_{1g}} \mathbb{1}\{G_i = g\} - \frac{1}{n_{1g'}} \mathbb{1}\{G_i = g'\} \right] \phi(X_i).$$

Intuitively, the imbalance in this transformation is the imbalance in the *difference* in covariate profiles of the two subgroups, across treatment and control. Practically, we can control this measure by including the transformed covariate vector $\tilde{\phi}_{gg'}(X_i)$ into the set of covariates that we balance, for a given contrast between two groups. In our application, when we include this additional set of covariates for the URM/non-URM contrast the results are nearly identical, indicating that this contrast is already well balanced.

B.2 No interference between applicants

As we discuss in the main text, we consider the availability of LORs as an individual-level treatment, and aim to uncover the average effect of LORs among the students who submit them. We assume independence of admissions across students — in particular, that the availability of LORs for one student does not affect any other student's admission probability. However, this cannot be strictly true in our setting. There are two possible sources of non-independence. First, the number of

admitted students is constrained, so admitting one student reduces the number of admissions offers that can be made to others. Second, many small colleges and elite universities treat admissions as a portfolio problem, aiming to balance a range of characteristics within the pool of admitted students. Under this approach, admitting (say) a tuba player might dramatically reduce the admissions chances of other applicants who play the tuba as well.

Neither of these is a major consideration at UC Berkeley, which admitted over 14,500 first-year students — as well as roughly 4,300 transfer students — for Fall 2017, in a relatively mechanistic way. With the exception of recruited athletes, who we exclude from our sample, each applicant is considered separately, with very limited interaction among the readers reviewing applications and no stage of the process at which balancing considerations can arise. Readers generally do not confer about application scores, and the capacity constraint is enforced only at the very end, where the combined reader scores are converted to admissions decisions; tiebreaking in this process considers very limited information about applicants and does not incorporate the portfolio balancing considerations that are prominent elsewhere (Stevens, 2009). Moreover, the total number of admitted undergraduates fluctuates meaningfully from year to year, with fewer than 12,000 first-year students admitted for Fall 2014 and nearly 15,500 first-year students admitted for Fall 2020, so is not a hard cap. Thus, while each admitted student must on average crowd out about one other student who would otherwise have been admitted, this crowding out is imperfect and is spread across many thousands of students.

We now formalize different types of dependence, following the literature on causal inference with interference; see Miles et al. (2019) for a closely related setup. Formally, let $Y_i(\mathbf{W})$ be the potential outcome associated with the vector of all units' treatment assignments, $\mathbf{W} \in \{0,1\}^N$, which we can equivalently write in terms of unit i's own treatment and the vector of remaining treatments: $Y_i(\mathbf{W}) = Y_i(W_i, \mathbf{W}_{-i})$. The substantive assumption that UC Berkeley does not engage in "portfolio balancing" for the LOR pilot study applicants corresponds to a *stratified interference* assumption: unit i's potential outcomes only depend on the *proportion* of units treated, rather than the identity of those units (Hudgens and Halloran, 2008). Formally, under this assumption can write unit i's potential outcomes as:

$$Y_i(\mathbf{W}) = Y_i(W_i, \mathbf{W}_{-i}) = Y_i\left(W_i, \frac{1}{N-1}\sum_{j\neq i} W_j\right) = Y_i(W_i, \bar{W}_{-i}),$$

where $\bar{W}_{-i} = \frac{1}{N-1}\sum_{j\neq i} W_j$ is the fraction of remaining units treated. Without additional restrictions, analysis is still possible but challenging (see Miles et al., 2019).

In our analysis, we differentiate between two key scenarios based on the fraction of treated units, $\bar{W}$, noting that $\bar{W} \approx \bar{W}_{-i}$ for large N. Let c be a given threshold for proportion treated, such that we consider two scenarios: $\bar{W} \in (0, c)$ and $\bar{W} \in [c, 1]$. Our substantive assumption for the main analysis is that, so long as there are relatively few applicants submitting LORs, the overall limit on undergraduate admissions is not binding and we can effectively ignore interference between units. Formally:

$$\text{for } \bar{W} < c: \quad Y_i(W_i, \bar{W}_{-1}) = Y_i(W_i) \quad \text{for all } i = 1, \ldots, N. \tag{3}$$

In Appendix A.4, we instead consider the opposite scenario in which all applicants are required to submit LORs: $\bar{W} = 1$. Here, we no longer believe that Equation (3) is a reasonable restriction, and instead use a simple policy simulation to assess the overall impact of such a policy.

C Simulation study

We now present simulations assessing the performance of our proposed approach versus traditional inverse propensity score weights fit via regularized logistic regression as well as outcome modelling with machine learning approaches. To better reflect real-world data, we generate correlated covariates and include binary and skewed covariates. For each simulation run, with $d = 50$ covariates, we begin with a diagonal covariance matrix Σ where $\Sigma_{jj} = \frac{(d-j+1)^5}{d^5}$ and sample a random orthogonal $d \times d$ matrix Q to create a new covariance matrix $\tilde{\Sigma} = Q\Sigma$ with substantial correlation. For $n = 10,000$ units, we draw covariates from a multivariate normal distribution $X_i \overset{iid}{\sim} N(0, \tilde{\Sigma})$. We then transform some of these covariates. For $j = 1, 11, 21, 32, 41$ we dichotomize the variable and define $\tilde{X}_{ij} = \mathbb{1}\{X_{ij} \geq q_{.8}(X_{\cdot j})\}$, where $q_{.8}(X_{\cdot j})$ is the 80th percentile of X_j among the n units. For $j = 2, 7, 12, ..., 47$ we create a skewed covariate $\tilde{X}_{ij} = \exp(X_{ij})$. To match our study, we create discrete subgroup indicators from the continuous variable X_{id}. To do this, we create a grid over X_{id} with grid size $\frac{n}{G}$, and sample $G-1$ points from this grid. We then create subgroup indicators G_i by binning X_{id} according to the $G-1$ points. We consider $G \in \{10, 50\}$ groups.

With these covariates, we generate treatment assignment and outcomes. We use a separate logistic propensity score model for each group following Equation (7),[4]

$$\text{logit } e(X_i, G_i) = \alpha_{G_i} + (\mu_\beta + U_g^\beta \odot B_g^\beta) \cdot X_i, \tag{4}$$

and also use a separate linear outcome model for each group,

$$Y_i(0) = \eta_{0G_i} + (\mu_\eta + U_g^\eta \odot B_g^\eta) \cdot X_i + \varepsilon_i, \tag{5}$$

where $\varepsilon_i \sim N(0,1)$ and $\odot$ denotes element-wise multiplication. We draw the fixed effects and varying slopes for each group according to a hierarchical model with sparsity. We draw the fixed effects as $\alpha_g \overset{\text{iid}}{\sim} N(0,1)$ and $\eta_{0g} \overset{\text{iid}}{\sim} N(0,1)$. For the slopes, we first start with a mean slope vector $\mu_\beta, \mu_\eta \in \{-\frac{3}{\sqrt{d}}, \frac{3}{\sqrt{d}}\}^K$, where each element is chosen independently with uniform probability. Then we draw isotropic multivariate normal random variables $U_g^\beta, U_g^\eta \overset{\text{iid}}{\sim} MVN(0, I_d)$. Finally, we draw a set of d binary variables $B_{gj}^\beta, B_{gj}^\eta$ that are Bernoulli with probability $p = 0.25$. The slope is then constructed as a set of sparse deviations from the mean vector, which is $\mu_\beta + U_g^\beta \odot B_g^\beta$ for the propensity score and $\mu_\eta + U_g^\eta \odot B_g^\eta$ for the outcome model.

To incorporate the possibility that treatment effects vary with additional covariates that are not the focus of our analysis, we generate the treatment effect for unit i as $\tau_i = X_{id} - X_{i3} + 0.3 X_{id} X_{i3}$ and set the treated potential outcome as $Y_i(1) = Y_i(0) + \tau_i W_i$. Note that the effect varies with the underlying continuous variable X_{id} that we use to form groups, as well as the additional variable X_{i3}. The true ATT for group g in simulation j is thus $\tau_{gj} = \frac{1}{n_{1g}} \sum_{G_i=g} W_i(Y_i(1) - Y_i(0))$, and the overall ATT is $\tau_j = \frac{1}{n_1} \sum_{i=1}^n W_i(Y_i(1) - Y_i(0))$.

For $j = 1, \ldots, m$ with $m = 500$ Monte Carlo samples, we estimate the treatment effects for group g, $\hat{\tau}_{gj}$, and the overall ATT, $\hat{\tau}_j$, and compute a variety of metrics. Following the metrics studied by Dong et al. (2020), for subgroup treatment effects we compute (a) the mean absolute bias across the G treatment effects, $\frac{1}{m} \sum_{j=1}^m \left| \frac{1}{g} \sum_{g=1}^G \hat{\tau}_{gj} - \tau_g \right|$, and (b) the mean root mean

[4] The logistic specification differs from the truncated linear odds in Equation 16. If the transformed covariates $\phi(X_i)$ include a flexible basis expansion, the particular form of the link function will be less important.

square error $\sqrt{\frac{1}{mG}\sum_{j=1}^{m}\sum_{g=1}^{G}(\hat{\tau}_{gj}-\tau_g)^2}$. For the overall ATT we measure (a) the absolute bias $\left|\frac{1}{m}\sum_{j=1}^{m}\hat{\tau}_j-\tau_j\right|$ and (b) the root mean square error $\sqrt{\frac{1}{m}\sum_{j=1}^{m}(\hat{\tau}_j-\tau_j)^2}$.

We compute treatment effects for the following eight estimators:

- *Partially pooled balancing weights:* approximate balancing weights that solve Equation (13), using G as the stratifying variable and prioritizing local balance by setting $\lambda_g = \frac{1}{n_{1g}}$.
- *Augmented balancing weights:* augmenting the partially pooled balancing weights as in Equation (21) where $\hat{m}_0(x, g)$ is fit via ridge regression with all interactions.
- *Fully pooled balancing weights:* approximate balancing weights that solve Equation (13), but ignore local balance by setting $\lambda \to \infty$, thus fully pooling towards the global model. This is equivalent to stable balancing weights in Equation (8) with an exact global balance constraint $\delta = 0$.
- *No pooled balancing weights:* approximate balancing weights that solve Equation (13), but without the exact global balance constraint.
- *Full interaction IPW:* traditional IPW with a fully interacted model that estimates a separate propensity score within each stratum as in Equation (7).
- *Fixed effects IPW:* full interaction IPW with stratum-specific coefficients constrained to be equal to a global parameter $\beta_g = \beta$ for all g.
- *Full interaction ridge regression outcome model:* Estimating μ_{0g} via $\hat{\mu}_{0g} = \frac{1}{n_1}\sum_{G_i=g} W_i\hat{m}_0(X_i, g)$, where $\hat{\mu}_0(x, g)$ is the same ridge regression predictor as for augmented balancing weights
- *Bayesian Causal Forests (BCF)*: Estimating τ_g as $\frac{1}{n_{1g}}\sum_{G_i=g} W_i\hat{\tau}_i$, where $\hat{\tau}_i$ are the posterior predictive means from a Bayesian Causal Forest estimator (Hahn et al., 2020).

For the fully interacted specification of the logistic regression in IPW and the ridge regression in the augmented balancing weights, we include a set of global parameters μ_β so that the slope for group g is $\mu_\beta + \Delta_g$, with a squared L^2 penalty for each component. These are correctly specified for the models above. We estimate the models with `glmnet` (Friedman et al., 2010) with the hyperparameter chosen through 5-fold cross validation.

Appendix Figure D.2 shows the results for the overall ATT and for subgroup effects. In most cases, the partially pooled approximate balancing approach has much lower bias and RMSE than the logistic regression-based IPW estimator; however, the RMSEs are comparable with 50 subgroups. Furthermore, prioritizing local balance with partial pooling yields lower bias and RMSE than ignoring local balance entirely with the fully pooled approach. Finally, including the global balance constraint yields much lower bias for the ATT in some settings, with relatively little cost to the subgroup estimates.

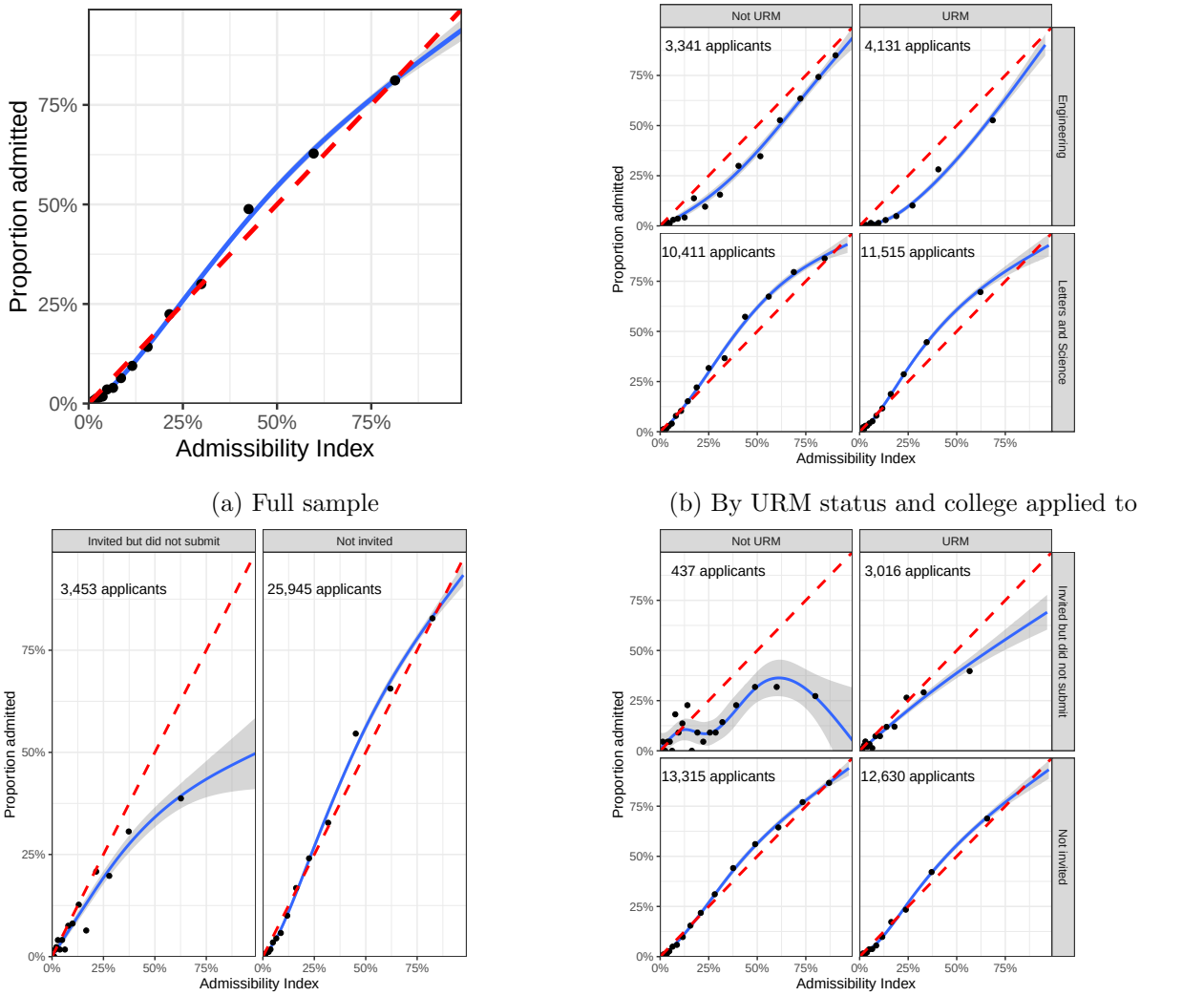
The partially-pooled balancing weights estimator, which is transparent and design-based, also performs nearly as well as the black-box BCF method. Augmenting the partially-pooled weights provides some small improvements to the bias, indicating that the weights alone are able to achieve good balance in these simulations, and a larger improvement in the RMSE in the setting with many groups where the weighting estimator alone have larger variance. Finally, Appendix Figure D.3 shows the coverage of 95% intervals for the different approaches. We see that the weighting

estimator, both with and without augmentation, has reasonable uncertainty quantification, with much better coverage than either of the two model-based approaches.

D Additional figures and tables

AI Range	URM	Number of Applicants	Number Submitting LOR	Proportion Treated
$< 5\%$	URM	11,832	2,157	18%
	Not URM	6,529	607	9%
5% - 10%	URM	3,106	1,099	35%
	Not URM	2,099	536	25%
10% - 20%	URM	2,876	1,212	42%
	Not URM	2,495	828	33%
$> 20\%$	URM	4,645	2,345	50%
	Not URM	6,959	2,359	34%

Table D.1: Number of applicants and proportion treated by subgroup.



(a) Full sample

(b) By URM status and college applied to

(c) By invitation to submit an LOR

(d) By URM status and invitation to submit an LOR

Figure D.1: Calibration of the admissibility index (a) in the full sample, (b) by URM status and college applied to, (c) by invitation to submit an LOR, and (d) by URM status and invitation to submit. The points are binned scatter plots for the quantiles in intervals of 5% and the lines are smoothed estimates of the relationship between the observed admission rate for applicants that did not submit LORs and the AI. The distribution of the quantiles emphasizes that most applicants are very unlikely to be admitted. For most groups, over three quarters have AI scores below 25%.

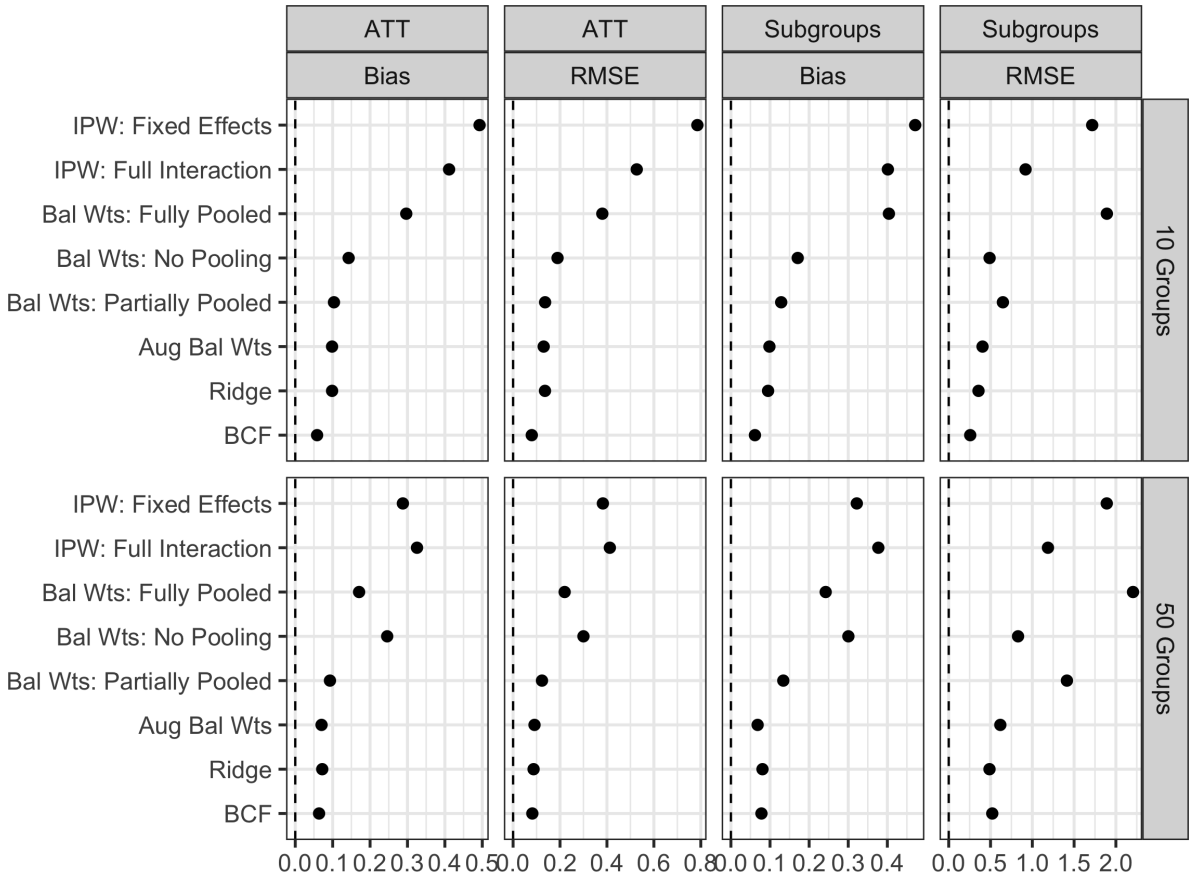


Figure D.2: Performance of approximate balancing weights, traditional IPW with logistic regression, and outcome modelling for estimating subgroup treatment effects.

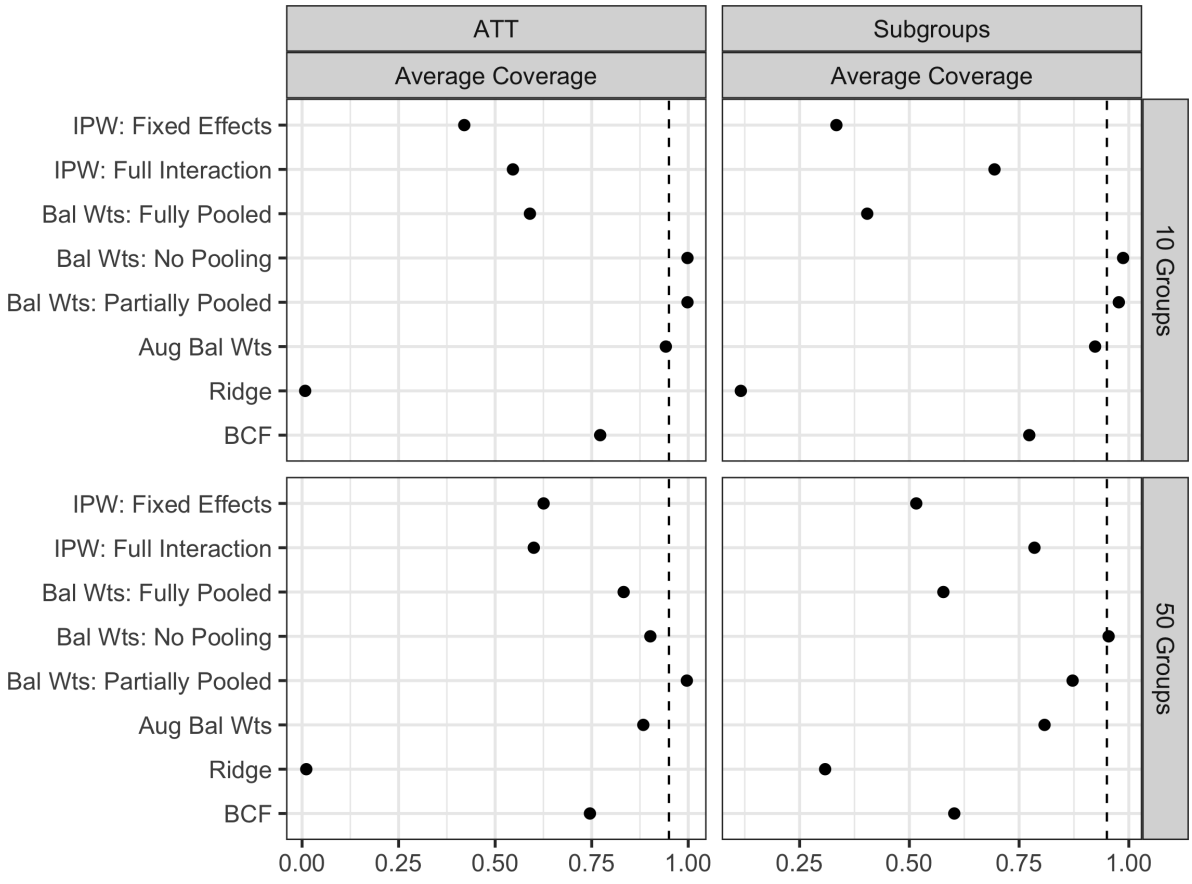


Figure D.3: Coverage for approximate balancing weights, traditional IPW with logistic regression, and outcome modelling for estimating subgroup treatment effects.

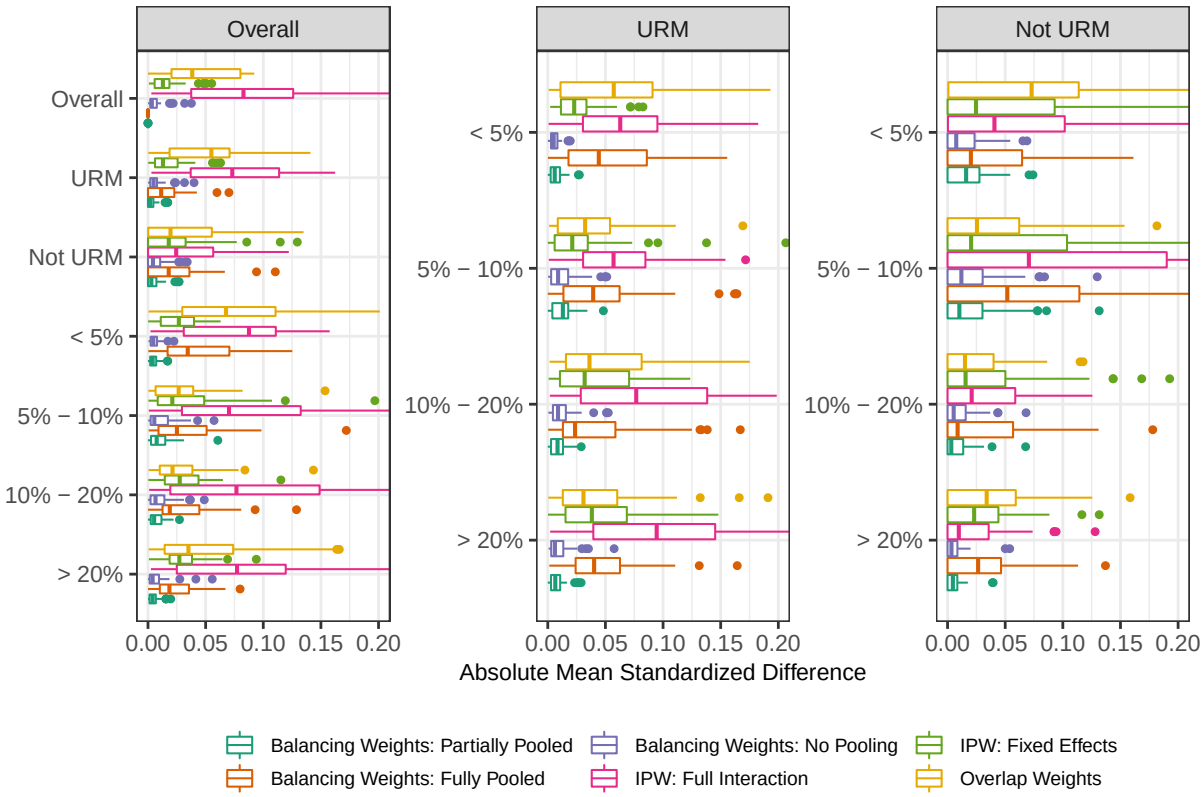
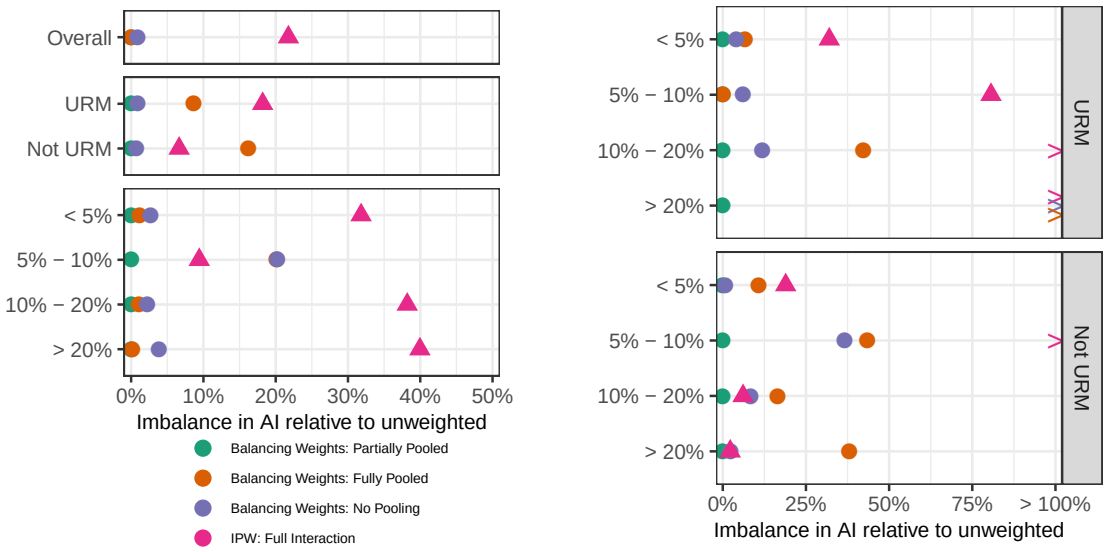


Figure D.4: Distribution of covariate balance measured by the mean standardized difference for different weighting methods.



(a) Overall and by URM status and AI.

(b) By URM status interacted with AI.

Figure D.5: Imbalance in the admissibility index after weighting relative to before weighting, overall and within each subgroup. For several subgroups, the fully pooled balancing weights procedure results in *increased* imbalance in the admissibility index, denoted by an arrow.

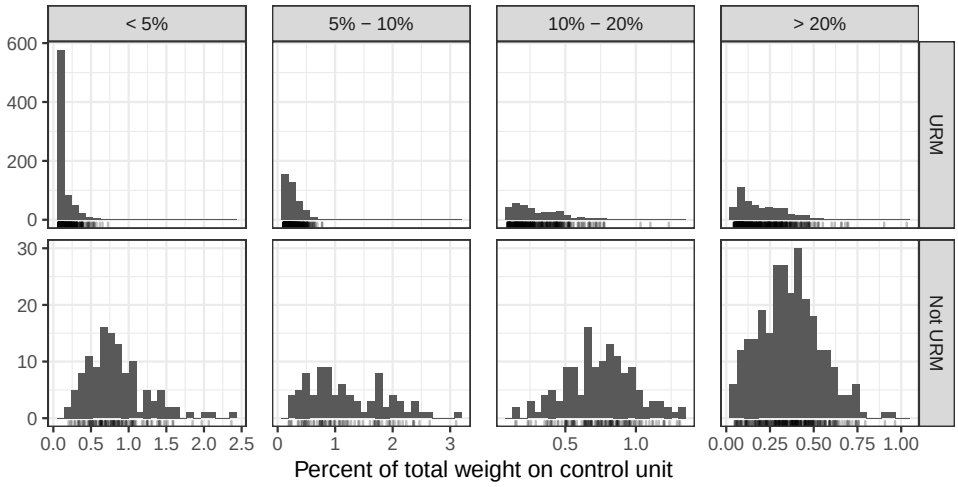


Figure D.6: Weights on control units from solving the approximate balancing weights problem (13). *Not pictured:* the 66% of control units that receive zero weight.

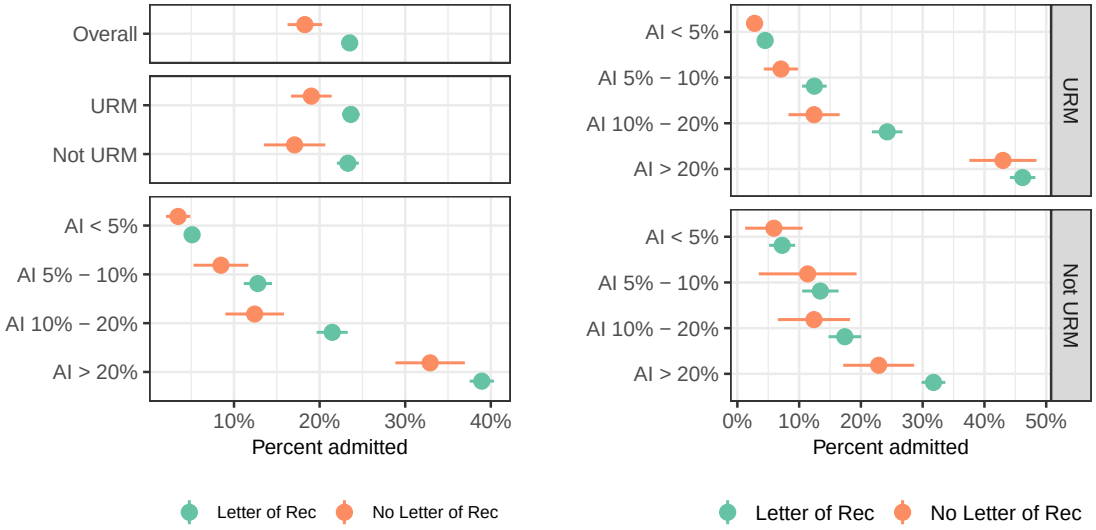
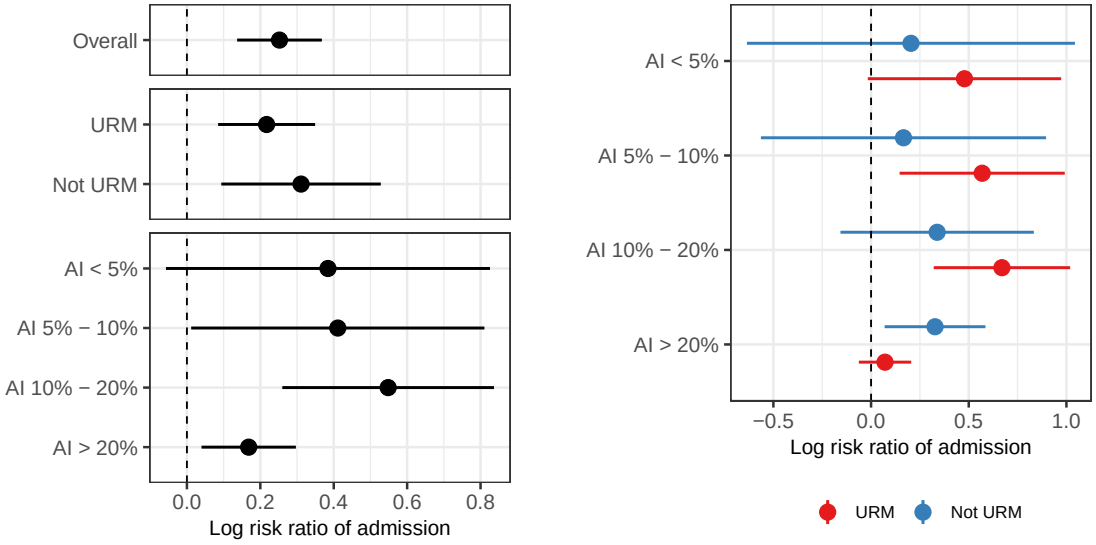


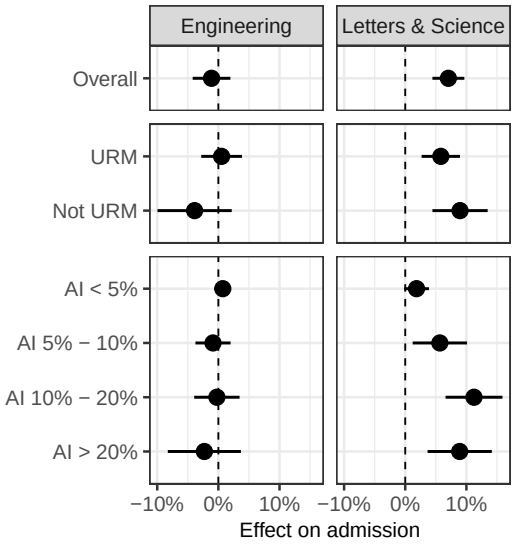
Figure D.7: Estimated treated and control means ± two standard errors: overall, by URM status, by Admissibility Index, and by URM × AI.



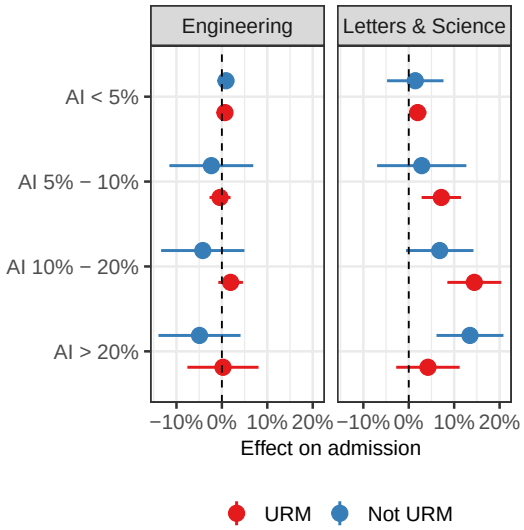
(a) Overall and by URM status and AI.

(b) By URM status interacted with AI.

Figure D.8: Estimated log risk ratio of admission with and without letters of recommendation ± two standard errors computed via the delta method, overall and by URM status and AI.

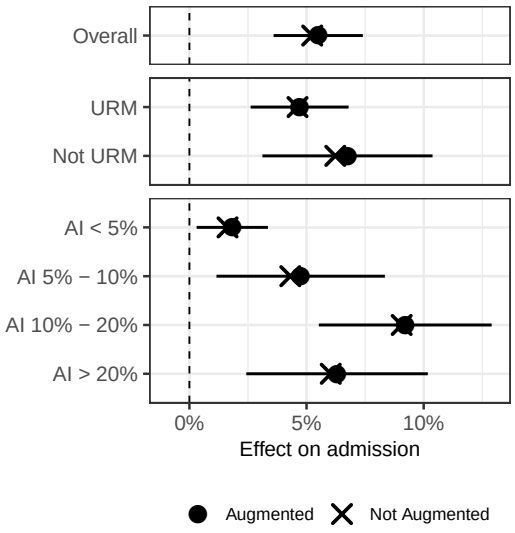


(a) Overall and by URM status and AI.

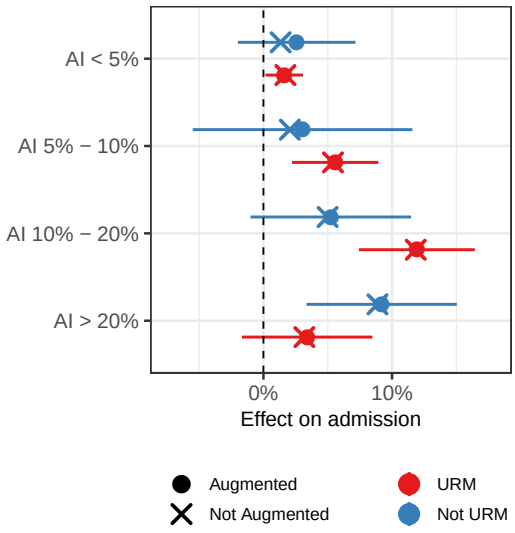


(b) By URM status interacted with AI.

Figure D.9: Estimated effect of letters of recommendation on admission rates for Engineering and Letters and Science applicants.

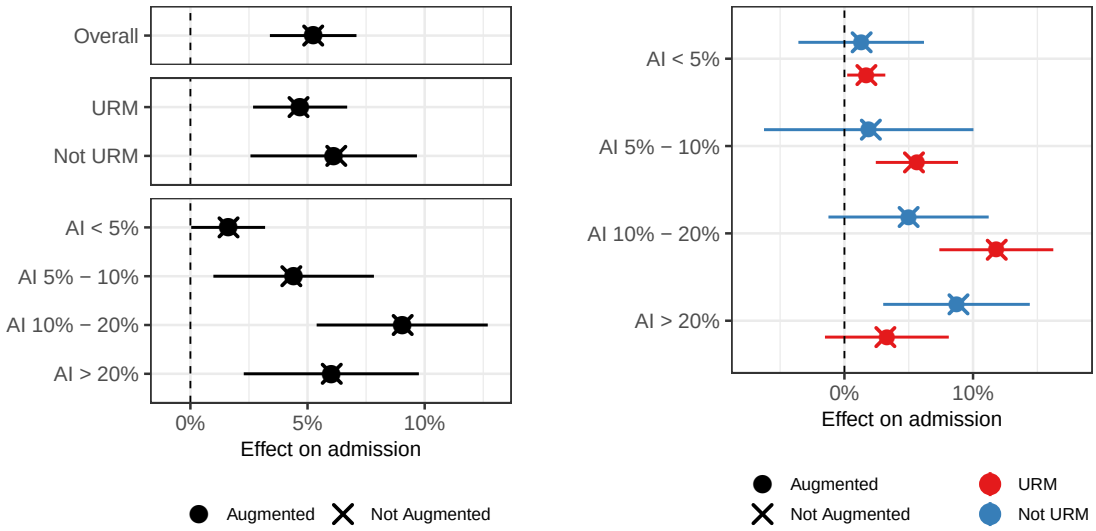


(a) Overall and by URM status and AI.



(b) By URM status interacted with AI.

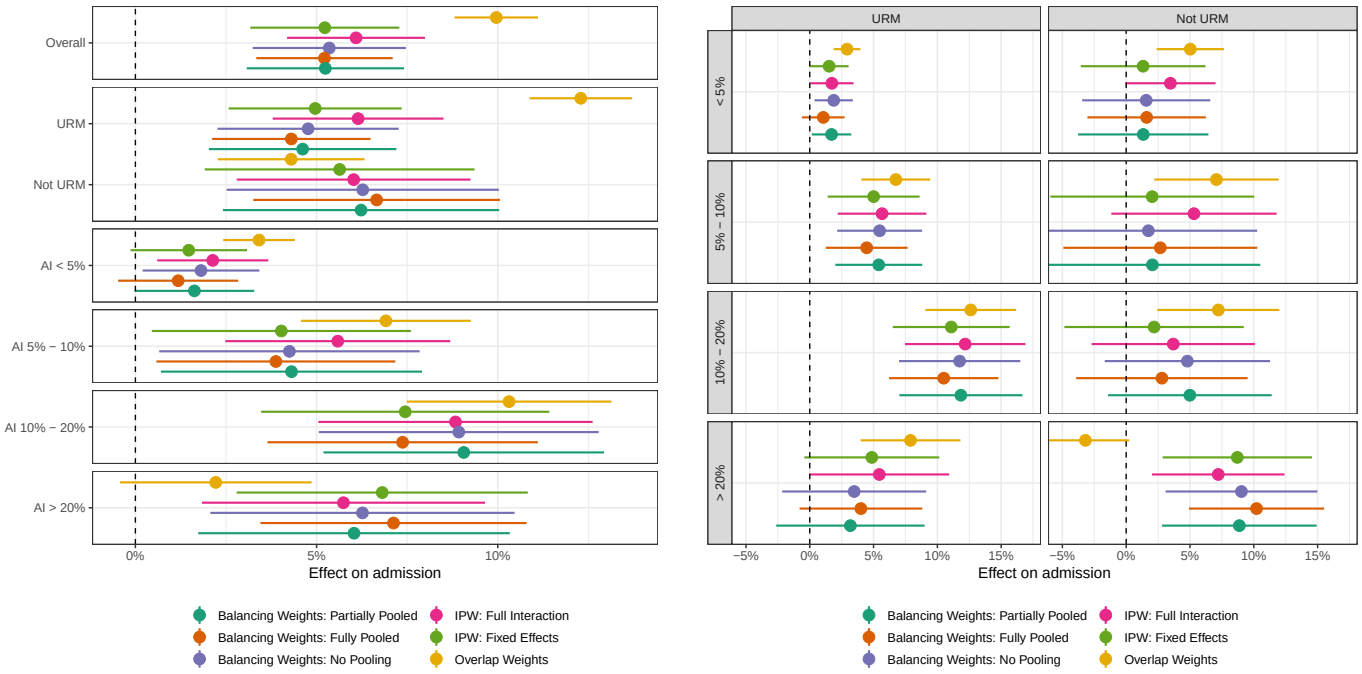
Figure D.10: Estimated effect of letters of recommendation on admission rates with and without augmentation via a random forest outcome model.



(a) Overall and by URM status and AI.

(b) By URM status interacted with AI.

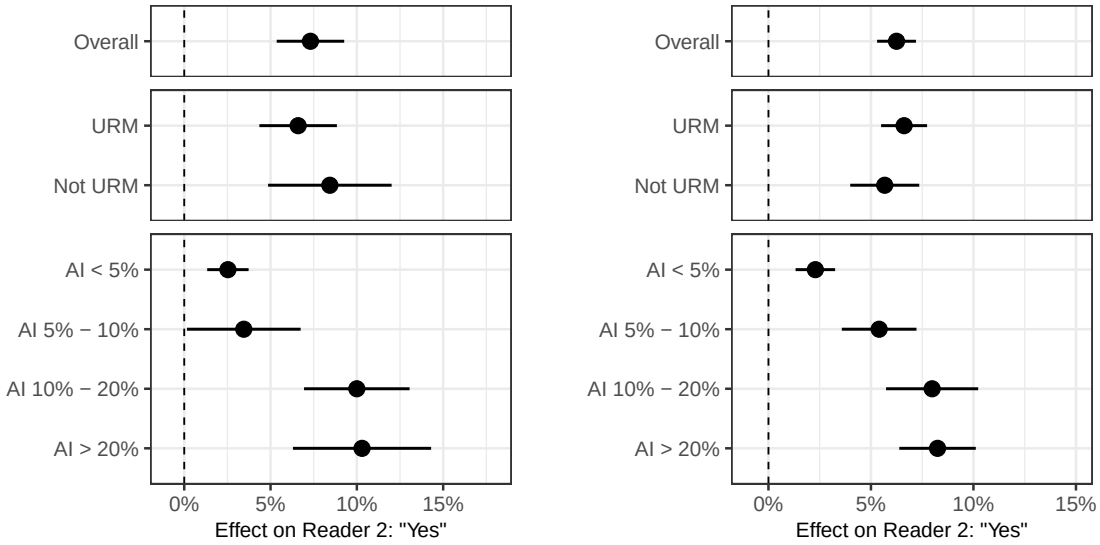
Figure D.11: Estimated effect of letters of recommendation on admission rates with and without augmentation via ridge regression with 5-fold cross validation.



(a) Overall and by URM status and AI.

(b) By URM status interacted with AI.

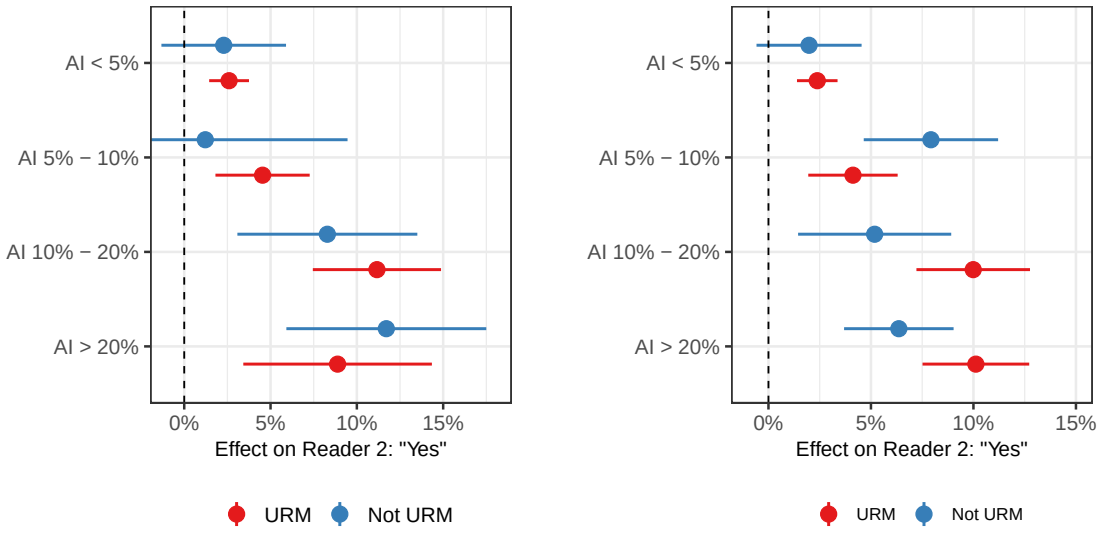
Figure D.12: Estimated effect of letters of recommendation on admission rates for comparable weighting estimators.



(a) Partially pooled balancing weights

(b) Within-subject design

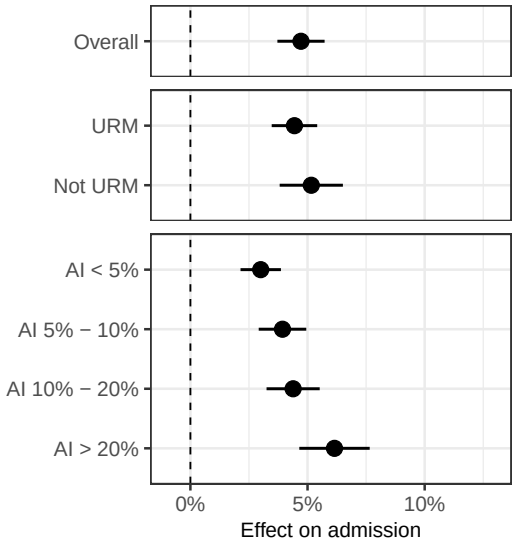
Figure D.13: Effects on second reader scores overall, by URM status, and by AI, estimated via (a) the partially pooled balancing weights estimator and (b) the within-subject design.



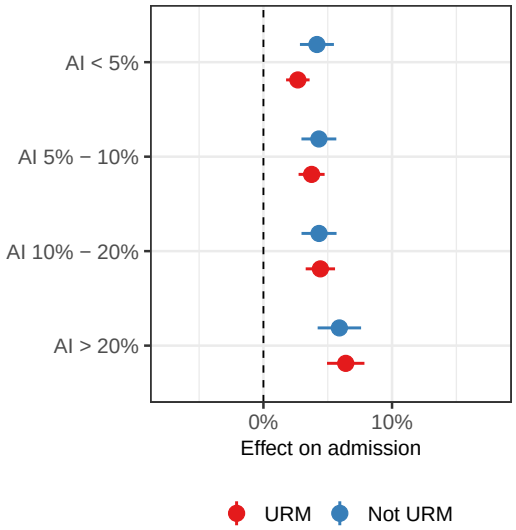
(a) Partially pooled balancing weights

(b) Within-subject design

Figure D.14: Effects on second reader scores by URM status interacted with AI, estimated via (a) the partially pooled balancing weights estimator and (b) the within-subject design.

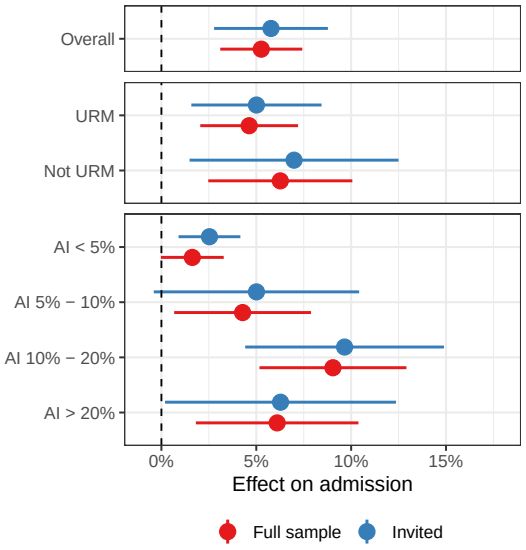


(a) Overall and by URM status and AI.

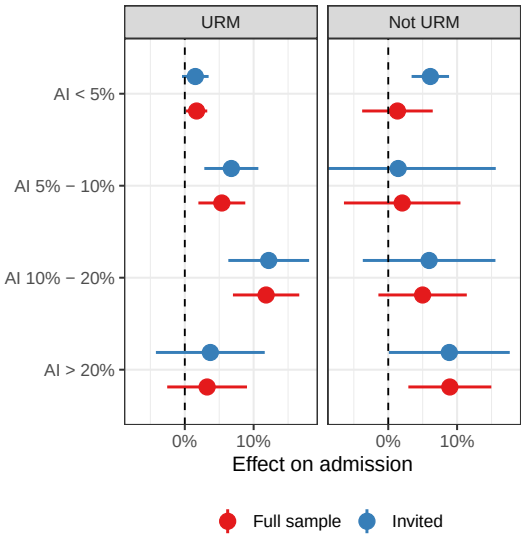


(b) By URM status interacted with AI.

Figure D.15: Estimated effect of letters of recommendation on admission rates via Bayesian Causal Forests (Hahn et al., 2020).



(a) Overall and by URM status and AI.



(b) By URM status interacted with AI.

Figure D.16: Estimated effect of letters of recommendation on admission rates via weighting in the full sample and restricting to applicants who were invited to submit an LOR

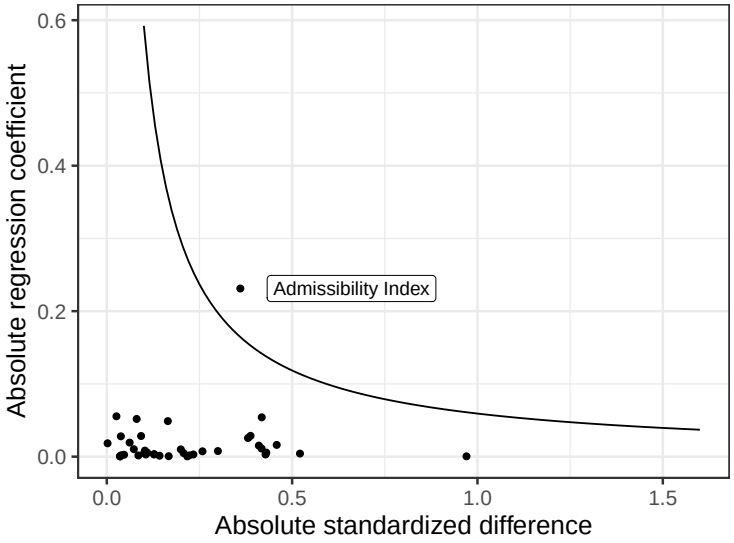


Figure D.17: Amplification of a sensitivity analysis. The line shows the magnitude of the regression coefficient and the magnitude of the imbalance in an unmeasured standardized covariate required to produce enough bias to remove the effect. Points correspond to the regression coefficients and imbalance before weighting for the 51 components of $\phi(X)$.

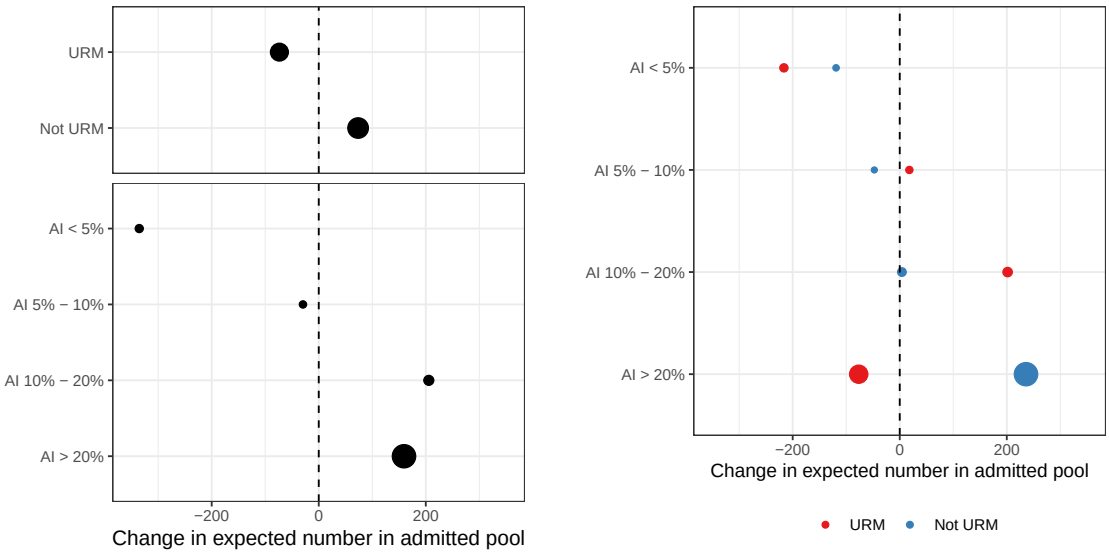


Figure D.18: Projected change in number of admitted applicants if all applicants submitted a letter of recommendation: by URM status, by Admissibility Index, and by URM × AI. Points are scaled by the expected size of the admitted group without submitting LORs; e.g., applicants with high AI are a much larger share of the admitted class than applicants with low AI.

E Proofs

Proof of Proposition 1. First, we will augment the primal optimization problem in Equation (13) with auxiliary covariates $\mathcal{E}_1, \ldots, \mathcal{E}_j$ so that $\mathcal{E}_g = \sum_{G_i=g, W_i=0} \gamma_i \phi(X_i) - \sum_{G_i=g, W_i=1} \phi(X_i)$. Then the optimization problem becomes:

$$\begin{aligned} \min_{\gamma} \quad & \sum_{z=1}^{J} \frac{1}{2\lambda_g} \|\mathcal{E}_j\|_2^2 + \frac{\lambda_g}{2} \sum_{Z_i=z, W_i=0} \gamma_i^2 + \mathcal{I}(\gamma_i \geq 0) \\ \text{subject to} \quad & \sum_{W_i=0} \gamma_i \phi(X_i) = \sum_{W_i=1} \phi(X_i) \\ & \mathcal{E}_j = \sum_{G_i=g, W_i=0} \gamma_i \phi(X_i) - \sum_{G_i=g, W_i=1} \phi(X_i), \quad z = 1, \ldots, J \\ & \sum_{G_i=g, W_i=0} \gamma_i = n_{1g}, \end{aligned} \tag{6}$$

where $\mathcal{I}(x \geq 0) = \begin{cases} 0 & x \geq 0 \\ \infty & x < 0 \end{cases}$ is the indicator function. The first constraint induces a Lagrange multiplier μ_β, the next J constraints induce Lagrange multipliers $\delta_1, \ldots, \delta_J$, and the sum-to-one constraints induce Lagrange multipliers $\alpha_1, \ldots, \alpha_J$. Then the Lagrangian is

$$\begin{aligned} \mathcal{L}(\gamma, \mathcal{E}, \mu_\beta, \delta, \alpha) = & \sum_{z=1}^{J} \left[\frac{1}{2\lambda_g} \|\mathcal{E}_j\|_2^2 - \mathcal{E}_j \cdot \delta_j + \sum_{G_i=g, W_i=0} \frac{1}{2}\gamma_i^2 + \mathcal{I}(\gamma_i \geq 0) - \gamma_i(\alpha + (\mu_\beta + \delta_j) \cdot \phi(X_i)) \right] \\ & + \sum_{z=1}^{J} \sum_{G_i=g, W_i=1} (1 + (\mu_\beta + \delta_j) \cdot \phi(X_i)) \end{aligned} \tag{7}$$

The dual objective is:

$$\begin{aligned} q(\mu_\beta, \delta, \alpha) = & \sum_{z=1}^{J} \left[\min_{\mathcal{E}_j} \left\{ \frac{1}{2\lambda_g} \|\mathcal{E}_j\|_2^2 - \mathcal{E}_j \cdot \delta_j \right\} + \sum_{G_i=g, W_i=0} \min_{\gamma_i \geq 0} \left\{ \frac{1}{2}\gamma_i^2 - \gamma_i(\alpha + (\mu_\beta + \delta_j) \cdot \phi(X_i)) \right\} \right] \\ & + \sum_{z=1}^{J} \sum_{G_i=g, W_i=1} (1 + (\mu_\beta + \delta_j) \cdot \phi(X_i)) \end{aligned} \tag{8}$$

Note that the inner minimization terms are the negative convex conjugates of $\frac{1}{2}\|x\|_2^2$ and $\frac{1}{2}x^2 + \mathcal{I}(X \geq 0)$, respectively. Solving these inner optimization problems yields that

$$\begin{aligned} q(\mu_\beta, \delta, \alpha) = & -\sum_{z=1}^{J} \left[\frac{\lambda_g}{2} \|\delta_j\|_2^2 + \sum_{G_i=g, W_i=0} [\alpha_j + (\mu_\beta + \delta_j) \cdot \phi(X_i)]_+^2 \right] \\ & + \sum_{z=1}^{J} \sum_{G_i=g, W_i=1} (1 + (\mu_\beta + \delta_j) \cdot \phi(X_i)) \end{aligned} \tag{9}$$

Now since there exists a feasible solution to the primal problem (13), from Slater's condition we see that the solution to the primal problem is equivalent to the solution to $\max_{\mu_\beta,\alpha,\delta} q(\mu_\beta, \alpha, \delta)$. Defining $\beta_j \equiv \mu_\beta + \delta_j$ gives the dual problem (18). Finally, note that the solution to the minimization over the weights in Equation (8) is $\gamma_i = [\alpha_j + \beta_j \cdot \phi(X_i)]_+$, which shows how to map from the dual solution to the primal solution. □

References

Dong, J., J. L. Zhang, S. Zeng, and F. Li (2020). Subgroup balancing propensity score. *Statistical Methods in Medical Research 29*(3), 659–676.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software 33*(1).

Hahn, P. R., J. S. Murray, and C. M. Carvalho (2020). Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects. *Bayesian Analysis*, 1–33.

Hudgens, M. G. and M. E. Halloran (2008). Toward causal inference with interference. *Journal of the American Statistical Association 103*(482), 832–842.

Miles, C. H., M. Petersen, and M. J. van der Laan (2019). Causal inference when counterfactuals depend on the proportion of all subjects exposed. *Biometrics 75*(3), 768–777.

Rothstein, J. (2017, July). The impact of letters of recommendation on UC Berkeley admissions in the 2016-17 cycle. Technical report, California Policy Lab.

Soriano, D., E. Ben-Michael, P. Bickel, A. Feller, and S. Pimentel (2020). Interpretable sensitivity analysis for balancing weights. Technical report. working paper.

Stevens, M. L. (2009). *Creating a Class: College Admissions and the Education of Elites*. Harvard University Press.

Zhao, Q., D. S. Small, and B. B. Bhattacharya (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society. Series B: Statistical Methodology 81*(4), 735–761.