

WORKING PAPER • No. 2021-4

Varying impacts of letters of recommendation on college admissions: Approximate balancing weights for subgroup effects in observational studies

Eli Ben-Michael, Avi Feller, and Jesse Rothstein

August 2021



www.capolicylab.org
[@capolicylab](https://twitter.com/capolicylab)

Varying impacts of letters of recommendation on college admissions: Approximate balancing weights for subgroup effects in observational studies*

Eli Ben-Michael, Avi Feller, and Jesse Rothstein

UC Berkeley and Harvard University

August 2021

Abstract

In a pilot program during the 2016-17 admissions cycle, the University of California, Berkeley invited many applicants for freshman admission to submit letters of recommendation. We use this pilot as the basis for an observational study of the impact of submitting letters of recommendation on subsequent admission, with the goal of estimating how impacts vary across pre-defined subgroups. Understanding this variation is challenging in observational studies, however, because estimated impacts reflect both actual treatment effect variation and differences in covariate balance across groups. To address this, we develop balancing weights that directly optimize for “local balance” within subgroups while maintaining global covariate balance between treated and control units. We then show that this approach has a dual representation as a form of inverse propensity score weighting with a hierarchical propensity score model. In the UC Berkeley pilot study, our proposed approach yields excellent local and global balance, unlike more traditional weighting methods, which fail to balance covariates within subgroups. We find that the impact of letters of recommendation increases with applicant strength but find little average difference for applicants from disadvantaged groups.

*email: ebenmichael@berkeley.edu. We thank Greg Dubrow, Chad Hazlett, Kosuke Imai, Amy Jarich, Luke Miratrix, Jared Murray, Olufeme Ogundole, and James Pustejovsky for helpful conversations and thoughtful comments. We also thank Elsa Augustine, Charles Davis, and Audrey Tiew for excellent research assistance. This work was supported in part by the William T. Grant Foundation and by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D200010. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

1 Introduction and motivation

In a pilot program during the 2016-17 admissions cycle, the University of California, Berkeley invited many applicants for freshman admission to submit letters of recommendation (LORs) as part of their applications. UC Berkeley had (and has) a “holistic review” admissions process, which attempts to examine the whole applicant, taking account of any contextual factors and obstacles overcome without over-reliance on quantitative measures like SAT scores ([Hout, 2005](#)). Unlike other highly selective universities, however, UC Berkeley had not routinely asked applicants to submit letters from teachers and guidance counselors.

The new approach proved controversial within the university. The LORs were intended to help identify students from non-traditional backgrounds who might otherwise be overlooked ([UC Berkeley, 2017](#)). But there was also legitimate concern that applicants from disadvantaged backgrounds might not have access to adults who could write strong letters, and that the use of letters would further disadvantage these students ([Chalfant, 2017](#)).

In this paper, we use the Berkeley pilot as the basis for an observational study of the impact of submitting a letter of recommendation on subsequent admission. Our goal is to assess how impacts vary across pre-defined subgroups, in order to inform the debate over the Berkeley policy and similar debates at other universities.

Assessing treatment effect heterogeneity is challenging in non-randomized settings because variation in estimated impacts reflects both actual treatment effect variation and differences in covariate balance across groups. Inverse Propensity Score Weighting (IPW) is one standard approach for addressing this: first estimate a propensity score model via logistic regression, including treatment-by-subgroup interaction terms; construct weights based on the estimated model; and then compare IPW estimates across subgroups (see [Green and Stuart, 2014](#); [Lee et al., 2019](#)). Estimated weights from traditional IPW methods, however, are only guaranteed to have good covariate balancing properties asymptotically. Balancing weights estimators, by contrast, instead find weights that directly minimize a measure of covariate imbalance, often yielding better finite sample performance ([Zubizarreta, 2015](#); [Athey et al., 2018](#); [Hirshberg and Wager, 2021](#); [Ben-Michael et al., 2020](#)). Both balancing weights

and traditional IPW face a curse of dimensionality when estimating subgroup effects: it is difficult to achieve exact balance on all covariates within each subgroup, or, equivalently, balance all covariate-by-subgroup interactions.

We therefore develop an approximate balancing weights approach tailored to estimating subgroup treatment effects, with a focus on the UC Berkeley LOR pilot study. Specifically, we present a convex optimization problem that finds weights that directly target the level of local imbalance within each subgroup — ensuring *approximate* local covariate balance — while guaranteeing *exact* global covariate balance between the treated and control samples. We show that controlling local imbalance controls the estimation error of subgroup-specific effects, allowing us to better isolate treatment effect variation. We also show that, even when the target estimand is the overall treatment effect, ensuring both exact global balance and approximate local balance reduces the overall estimation error.

Next, we demonstrate that this proposal has a dual representation as inverse propensity weighting with a hierarchical propensity score model, building on recent connections between balancing weights and propensity score estimation (Ben-Michael et al., 2020). In particular, finding weights that minimize both global and local imbalance corresponds to estimating a propensity score model in which the subgroup-specific parameters are partially pooled toward a global propensity score model. Any remaining imbalance after weighting may lead to bias. To adjust for this, we also combine the weighting approach with an outcome model, analogous to bias correction for matching (Rubin, 1973; Athey et al., 2018).

After assessing our proposed approach’s properties, we use it to estimate the impacts of letters of recommendation during the 2016 UC Berkeley undergraduate admissions cycle. Based on the Berkeley policy debate, we focus on variation in the effect on admissions rates based on Berkeley’s preferred markers of student disadvantage (such as low income, or from a low-scoring high school), and on applicant strength estimated using data from the prior year’s admissions cycle. First, we show that the proposed weights indeed yield excellent local and global balance, while traditional propensity score weighting methods yield poor local balance. We then find evidence that the impact of letters increases with the predicted

applicant strength. Applicants who are very unlikely to be admitted see little benefit from letters of recommendation, while applicants on the cusp of acceptance see a larger, positive impact.

The evidence on the differential effects across student groups is more mixed. Overall, the point estimates for disadvantaged and non-disadvantaged applicants are close to each other. However, these estimates are noisy and mask important variation by applicant strength. For applicants with the strongest quantifiable credentials, we estimate larger impacts for non-disadvantaged applicants, though these estimates are sensitive to augmentation with an outcome model. For all other applicants, we estimate the reverse: larger impacts for disadvantaged than non-disadvantaged applicants. Since student disadvantage is correlated with applicant strength, this leads to a Simpson’s Paradox-type pattern for subgroup effects, with a slightly larger point estimate for non-disadvantaged applicants pooled across groups (Bickel et al., 1975; VanderWeele and Knol, 2011).

Using these estimates, we then conduct a simple simulation exercise to project the impact of a policy requiring letters of recommendation for all applicants. We find that LORs would have minimal effects on the demographic composition of admitted students, with a slight increase in the share of admitted students with strong traditional qualifications.

The fact that results hinge on higher-order interactions suggests caution but also highlights the advantages of our design-based approach with pre-specified subgroups, in the sense of separating the design and analysis phases of the analysis (Rubin, 2008). As a result, we can carefully assess covariate balance and overlap in the subgroups of interest — and can tailor the weights to target these quantities directly. By contrast, many black-box machine learning methods estimate the entire conditional average treatment effect function, without prioritizing estimates for pre-specified subgroups (Carvalho et al., 2019). While these approaches require minimal input from the researcher, they can over-regularize the estimated effects for subgroups of interest, making it difficult to isolate higher-order interactions. As we discuss in Section 3.4, we argue that researchers should view these weighting and outcome modeling methods as complements rather than substitutes. Even without pre-specified

subgroups, there can be substantial gains from combining weighting and outcome modeling for treatment effect variation (Nie and Wager, 2017; Künzel et al., 2019). Our proposed augmented estimator is especially attractive because we can incorporate pre-defined subgroups in the weighting estimator while still leveraging off-the-shelf machine learning methods for outcome modeling.

Finally, we conduct extensive robustness and sensitivity checks, detailed in Appendix A. In addition to alternative estimators and sample definitions, we conduct a formal sensitivity analysis for violations of the ignorability assumption, adapting the recent proposal from Soriano et al. (2020). We also explore an alternative approach that instead leverages unique features of the UC Berkeley pilot study, which included an additional review without the letters of recommendation for a sample of 10,000 applicants. Overall, our conclusions are similar across a range of approaches. Thus, we believe our analysis is a reasonable first look at this question, albeit best understood alongside other approaches that rest on different assumptions (such as those in Rothstein, 2017).

The paper proceeds as follows. In the next section we introduce the letter of recommendation pilot program at UC Berkeley. Section 2 introduces the problem setup and notation, and discusses related work. Section 3 proposes and analyzes the approximate balancing weights approach. Section 4 presents empirical results on the effect of letters of recommendation. Section 5 concludes with a discussion about possible extensions. The appendix includes additional analyses and theoretical discussion, as well as an extensive simulation study.

1.1 A pilot program for letters of recommendation

As we discuss above, there is considerable debate over the role of letters of recommendation in college admissions. LORs have the potential to offer insight into aspects of the applicant not captured by the available quantitative information or by the essays that applicants submit (Kuncel et al., 2014). At the same time, letters from applicants from disadvantaged backgrounds or under-resourced high schools may be less informative or prejudicial against the applicant, due, e.g., to poor writing or grammar, or to lower status of the letter

writer ([Schmader et al., 2007](#)). A related concern arises in admissions essays: [Alvero et al. \(2021\)](#) find that essay text is strongly predictive of family income, using Natural Language Processing methods on a corpus from University of California applicants.

These issues were central in the UC Berkeley debates over the LOR policy. One Academic Senate committee, following an inquiry into the “intended and unintended consequences of the interim pilot admissions policy, especially for underrepresented minority students,” concluded that “the burden of proof rests on those who want to implement the new letters of recommendation policy, and should include a test of statistical significance demonstrating measurable impact on increasing diversity in undergraduate admissions” ([UC Berkeley, 2017](#)). The UC system-wide faculty senate ultimately limited the use of LORs following the pilot (though before any results were available), citing concerns that “LORs conflict with UC principles of access and fairness, because students attending under-resourced schools or from disadvantaged backgrounds will find it more difficult to obtain high-quality letters, and could be disadvantaged by a LOR requirement” ([Chalfant, 2017](#)).

The UC Berkeley LOR pilot study is a unique opportunity to assess these questions; [Rothstein \(2017\)](#) discusses implementation details as well as an LOR pilot study in the previous admissions cycle. For this analysis, we focus on non-athlete California residents who applied for freshman admission to either the College of Letters and Science or the College of Engineering at UC Berkeley in the 2016 admissions cycle. There were 40,541 such applicants, 11,143 of whom submitted LORs. We examine the impacts for applicants who both were invited to and subsequently did submit LORs; we consider alternative approaches in [Section 4](#) and [Appendix A](#).

Student disadvantage and its use in undergraduate admissions. Our primary interest is in estimating treatment effects of LORs separately for students who are and are not from groups underrepresented among admitted students. We follow the university in defining an underrepresented (or “URM”) applicant as one who is a low-income student, a student from a low-performing high school, a first-generation college student, or a student from an underrepresented racial or ethnic group (Black, Hispanic, or American Indian or

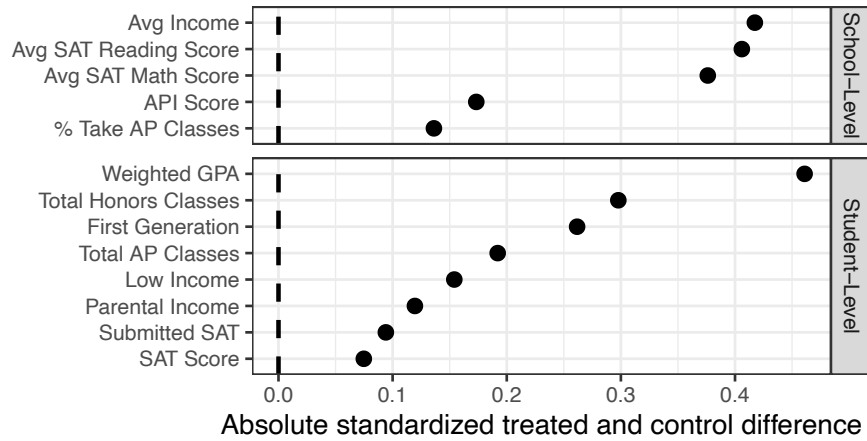


Figure 1: Absolute difference in means, standardized by the pooled standard deviation, between applicants submitting and not submitting LORs for several key covariates. By design, applicants submitting LORs disproportionately have a “Possible” score from the first reader (70% of treated applicants vs. 4% of untreated applicants).

Alaskan Native). Based on this definition, 55% of applicants in our sample are categorized as URM.¹

Indicators for three of these four traits are included in admissions files that are given to application readers and may be considered in evaluating applicants. Under California state law, UC admissions cannot consider race or ethnicity, and this information is not provided to readers (though readers can likely infer it from other application information). Under UC Berkeley’s holistic review, we cannot measure directly what weight readers put on these traits, if any. The debate about the potential disparate effects of LORs hinges on an assumption that readers might evaluate URM applicants differently if given more information.²

Selection into treatment. As in any observational study of causal effects, selection into treatment is central. Selection into submitting letters was a two-step process: A subset of

¹26% are low-income; 37% from low-performing schools; 23% are first generation; and 28% are from underrepresented racial or ethnic groups; there is of course substantial overlap among these categories.

²A similar issue has arisen in an ongoing lawsuit concerning admissions at Harvard, where a key dispute concerns whether differences in subjective scores given to applicants of different racial groups reflect direct weight placed on race or race-neutral evaluation of essays, LORs, and other materials that may correlate with race.

students were invited to provide letters, and then invited applicants did or did not submit them. The selection of students to be invited was embedded in the application review process, and depended on the initial application review. UC Berkeley uses a two-reader evaluation system. Each reader scores applicants on a three-point scale, as “No,” “Possible,” or “Yes.”³ In the LOR pilot, any applicant who received a “Possible” score from the first reader was invited to submit letters. In addition, due to concerns that the first readers’ scores would not be available in time to be useful, an index of student- and school-level characteristics was generated, and applicants with high levels of the index were invited as well.⁴ When submitted, letters were made available to the second reader for possible consideration, with the sole instruction that applicants’ scores should not be harmed either by the absence of letters or by the content of letters if submitted.

Of our sample of 40,451 applicants, 14,596 were invited to submit a letter. Among those invited, 11,143 (76%) eventually submitted letters, and no applicant submitted a letter who was not invited to. Because the “treatment” of interest is the inclusion of letters in the reader evaluation, we include the 3,453 applicants who were invited but did not submit LORs as part of the possible comparison group; we consider alternative definitions in Appendix A.2.

For this analysis, we assume that submission of LORs is effectively random conditional on the first reader score and on both student- and school-level covariates (Assumption 1 below). In particular, the *interaction* between the covariates and the first reader score plays an important role in the overall selection mechanism, as applicants who received a score of “No” or “Yes” from the first reader could still have been asked to submit an LOR based on their individual and school information. Figure 1 shows covariate imbalance for several key covariates — measured as the absolute difference in means divided by the pooled

³Application decisions are based on the combination of these two scores and the major to which a student has applied. In the most selective majors (e.g., mechanical engineering), an applicant typically must receive two “Yes” scores to be admitted, while in others a single “Yes” is sufficient.

⁴The index was generated from a logistic regression fit to data from the prior year’s admissions cycle, predicting whether an applicant received a “Possible” score (versus either a “No” or a “Yes”). Applicants with predicted probabilities from this model greater than 50% were invited to submit LORs. Because we observe all of the explanatory variables used in the index, this selection depends only on observable covariates. A small share of applicants with low predicted probabilities received first reads after January 12, 2017, the last date that LOR invitations were sent, and were not invited even if they received “Possible” scores.

standard deviation — for applicants who submitted LORs versus those who did not.⁵ We see that there are large imbalances in observable applicant characteristics, most notably average school income, GPA, the number of honors and AP classes taken, and SAT score. There were also large imbalances in first reader scores (not shown in Figure 1): 70% of applicants that submitted LORs had “Possible” scores, compared to only 4% of those who did not. There is a smaller imbalance in URM status, with 61% of those submitting LORs classified as URMs, versus 53% of those who did not submit. As we discuss below, we adjust for these differences in observable characteristics between applicants who do and do not submit LORs. However, differences in *unobservable* characteristics across applicants, e.g. in conscientiousness, may bias our effects, likely upward. To account for this possibility, we assess the sensitivity of our results to unmeasured confounding variables in Appendix A.3.

Independent admissions decisions. 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 (see Section 2.1 below). 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.

⁵The full set of student-level variables we include in our analysis are: weighted and unweighted GPA, GPA percentile within school, parental income and education, SAT composite score and math score, the number of honors courses and percentage out of the total available, number of AP courses, ethnic group, first generation college student status, and fee waiver status. The school level variables we control for are: average SAT reading, writing, and math scores, average ACT score, average parental income, percent of students taking AP classes, and the school Academic Performance Index (API) evaluated through California’s accountability tests. For students that did not submit an SAT score but did submit an ACT score, we imputed the SAT score via the College Board’s SAT to ACT concordance table. For the 992 applicants with neither an SAT nor an ACT score, we impute the SAT score as the average among applicants from the school.

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.

Therefore, we maintain an independence assumption for most of our analysis; see Appendix B.2 for a more formal discussion. We do incorporate capacity constraints in Section 4.3, when we consider the implications of our student-level treatment effects for a potential policy in which all applicants submit LORs.

Heterogeneity across application strength The admissions office provided us with a univariate summary of the large number of applicant- and school-level characteristics, which we refer to as the *Admissibility Index* (AI). This is computed as the prediction from a logistic regression fit to admissions data from the *prior year* (2015), using linear terms for the admissions variables without interactions.⁶

One possible way to estimate the effect of LORs is to use the AI as a direct measure of the admissions probability without LORs, and merely compare the actual admissions outcomes

⁶These variables include those in Figure 1 as well as disaggregated SAT scores, parental education, and an indicator if less than 5% of students from the high school apply to UC Berkeley. Notably, the AI does not include ethnic group or race information.

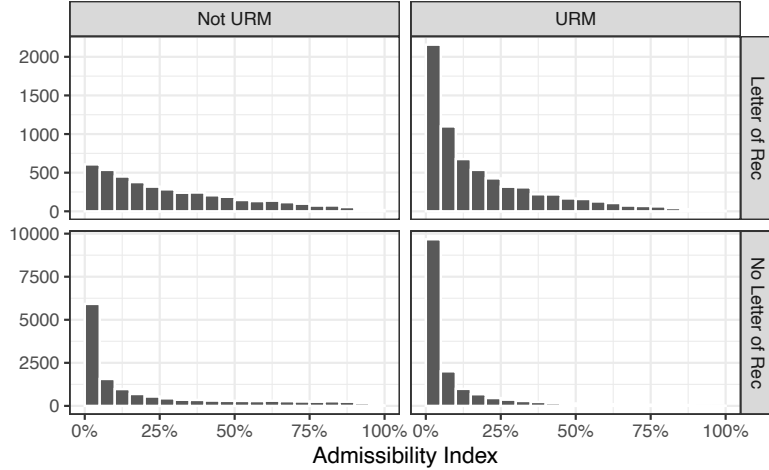


Figure 2: Distribution of the “admissibility index” — an *a priori* estimate of applicant strength — for the 2016 UC Berkeley application cohort, separated into URM and non-URM and those that submitted letters versus those that did not.

of LOR-submitting students to their average AIs. We do not pursue this, for several reasons. First, the AI is estimated from a prior year; we find that it is positively correlated with admissions in our control group, but not perfectly calibrated overall. Second, the AI model is simple and fails to account for important interactions — e.g., students who apply to more competitive colleges or majors have much lower admissions probabilities than students with the same AI who apply to less competitive pools.⁷ Third, about 15% of students in the data used to train the AI model submitted LORs in an earlier iteration of the UC Berkeley pilot; thus, the AI does not predict admissions outcomes without treatment. Finally, and most importantly, the first reader’s score is, as discussed above, an extremely important determinant both of selection into treatment and of outcomes — but is not included in the AI. Thus, we treat the AI as a single *a priori* index of applicant strength, not literally of the probability of admission. We seek to balance the AIs of treated applicants and controls, along with other measures like the first reader’s score.

Figure 2 shows the AI distribution for the 2016 applicant cohort, broken out by URM status and LOR submission. There are several features of this distribution that have im-

⁷ As we show in Appendix Figure D.1, while the AI has decent calibration for the overall sample, it is mis-calibrated for Engineering applicants and high admissibility Letters and Science applicants, both URM and non-URM.

portant implications for our analysis. First, applicants across nearly the full AI support submitted LORs. This is primarily because applicants who received “Possible” scores from the first readers come from a wide range of admissibility levels. This will allow us to estimate heterogeneous effects across the full distribution, with more precision for applicants with lower AIs. Second, because the admissions model disproportionately predicted that URM students had high chances of receiving “Possible” scores, many more URM applicants were invited to submit letters than non-URM applicants, and so our estimates for URM applicants will be more precise than those for non-URM applicants. Third, at higher AI levels large shares of applicants submitted LORs, leaving few comparison observations. This will make it challenging to form balanced comparison groups for high-AI applicants who submit letters.

From Figure 2 we know that the distribution of AI varies between URM and non-URM applicants, and so apparent differences in estimated average effects between the two groups may be due to compositional differences. Therefore, in the subsequent sections we will focus on estimating effects within subgroups defined by both URM status and the AI. To do this, we define subgroups by creating four (non-equally-sized) strata of the AI: $< 5\%$, $5\% - 10\%$, $10\% - 20\%$ and $> 20\%$. Interacting with URM status, this leads to eight non-overlapping subgroups; we will marginalize over these to estimate the other subgroup effects above. Appendix Table D.1 shows the total number of applicants in each of the eight groups, along with the proportion submitting letters of recommendation. As we discuss in Section 4, we will further divide each of these subgroups by first reader score and college, to ensure exact balance on these important covariates.

2 Treatment effect variation in observational studies

2.1 Setup and estimands

We now describe the letter of recommendation study as an observational study where for each applicant $i = 1, \dots, n$, we observe applicant and school level-covariates $X_i \in \mathcal{X}$; a group indicator $G_i \in \{1, \dots, K\}$ denoting a pre-defined subgroup of interest; a binary indicator for

submitting a letter of recommendation $W_i \in \{0, 1\}$; and whether the applicant is admitted, which we denote as $Y_i \in \{0, 1\}$. Let n_{1g} and n_{0g} represent the number of treated and control units in subgroup $G_i = g$, respectively. We assume that for each applicant, (X_i, G_i, W_i, Y_i) are sampled i.i.d. from some distribution $\mathcal{P}(\cdot)$. Following the potential outcomes framework, we assume SUTVA and posit two potential outcomes $Y_i(0)$ and $Y_i(1)$ for each applicant i , corresponding to i 's outcome if that applicant submits a letter of recommendation or not, respectively; the observed outcome is $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$. Importantly, this rules out interference between applicants. As we discuss above, we believe it is reasonable to ignore interference between units when analyzing the LOR pilot study. We discuss this restriction more formally in Appendix B.2, and we relax this restriction in our policy simulation in Section 4.3.

In this study we are interested in estimating two types of effects. First, we wish to estimate the overall Average Treatment Effect on the Treated (ATT), the treatment effect for applicants who submit a letter,

$$\tau = \mathbb{E}[Y(1) - Y(0) \mid W = 1],$$

where we denote $\mu_1 = \mathbb{E}[Y(1) \mid W = 1]$ and $\mu_0 = \mathbb{E}[Y(0) \mid W = 1]$. Second, for each subgroup $G_i = g$, we would like to estimate the Conditional ATT (CATT),

$$\tau_g = \mathbb{E}[Y(1) - Y(0) \mid G = g, W = 1], \tag{1}$$

where similarly we denote $\mu_{1g} = \mathbb{E}[Y(1) \mid G = g, W = 1]$ and $\mu_{0g} = \mathbb{E}[Y(0) \mid G = g, W = 1]$.

Estimating μ_{1g} is relatively straightforward: we can simply use the average outcome for treated units in group g , $\hat{\mu}_{1g} \equiv \frac{1}{n_{1g}} \sum_{G_i=g} W_i Y_i$. However, estimating μ_{0g} is more difficult due to confounding; we focus much of our discussion on imputing this counterfactual mean for the group of applicants who submitted letters of recommendation. To do this, we rely on two key assumptions that together form the usual *strong ignorability* assumption.

Assumption 1 (Ignorability). The potential outcomes are independent of treatment given

the covariates and subgroup:

$$Y(1), Y(0) \perp\!\!\!\perp W \mid X, G. \quad (2)$$

Assumption 2 (One Sided Overlap). The *propensity score* $e(x, g) \equiv P(W = 1 \mid X = x, G = g)$ is less than 1:

$$e(X, G) < 1. \quad (3)$$

In our context, Assumption 1 says that conditioned on the first reader score and applicant- and school-level covariates, submission of LORs is independent of the potential admissions outcomes. Due to the selection mechanism we describe in Section 1.1, we believe that this is a reasonable starting point for estimating these impacts; see Rothstein (2017) and Appendix A.2 for alternatives. In Appendix A.3, we assess the sensitivity of our conclusions to violations of this assumption.

Assumption 2 corresponds to assuming that no applicant would have been guaranteed to submit a letter of recommendation. Although some applicants were guaranteed to be *invited* to submit an LOR, we believe that this is a reasonable assumption for actually submitting a letter. In Section 4.1 we assess overlap empirically.

With this setup, let $m_0(x, g) = \mathbb{E}[Y(0) \mid X = x, G = g]$ be the *prognostic score*, the expected control outcome conditioned on covariates X and group membership G . Under Assumptions 1 and 2, we have the standard identification result:

$$\mu_{0g} = \mathbb{E}[m_0(X, G) \mid W = 1] = \mathbb{E} \left[\frac{e(X, G)}{1 - e(X, G)} Y \mid W = 0 \right]. \quad (4)$$

Therefore we can obtain a plug-in estimate for μ_{0g} with an estimate of the prognostic score, $m_0(\cdot, \cdot)$, an estimate of the propensity score, $e(\cdot, \cdot)$, or an estimate of the treatment odds themselves, $\frac{e(\cdot, \cdot)}{1 - e(\cdot, \cdot)}$.

2.2 Related work

There is an extensive literature on estimating varying treatment effects in observational studies; see Anoke et al. (2019) and Carvalho et al. (2019) for recent discussions. This is an active area of research, and we narrow our discussion here to methods that assess

heterogeneity across pre-defined, discrete subgroups. In particular, we will focus on linear weighting estimators that take a set of weights $\hat{\gamma} \in \mathbb{R}^n$, and estimate μ_{0g} as a weighted average of the control outcomes in the subgroup:

$$\hat{\mu}_{0g} \equiv \frac{1}{n_{1g}} \sum_{G_i=g} \hat{\gamma}_i (1 - W_i) Y_i. \quad (5)$$

Many estimators take this form; we focus on design-based approaches that do not use outcome information in constructing the estimators (Rubin, 2008). See Hill (2011); Künzel et al. (2019); Carvalho et al. (2019); Nie and Wager (2017); Hahn et al. (2020) for discussions of approaches that instead focus on outcome modeling.

Methods based on estimated propensity scores. A canonical approach in this setting is Inverse Propensity Weighting (IPW) estimators for μ_{0g} (see Green and Stuart, 2014). Traditionally, this proceeds in two steps: first, estimate the propensity score $\hat{e}(x, g)$, e.g., via logistic regression; second, estimate μ_{0g} as in Equation (5), with weights $\hat{\gamma}_i = \frac{\hat{e}(X_i, G_i)}{1 - \hat{e}(X_i, G_i)}$:

$$\hat{\mu}_{0g} = \frac{1}{n_{1g}} \sum_{W_i=0, G_i=g} \frac{\hat{e}(X_i, G_i)}{1 - \hat{e}(X_i, G_i)} Y_i, \quad (6)$$

where these are “odds of treatment” weights to target the ATT. A natural approach to estimating $\hat{e}(X_i, G_i)$, recognizing that G_i is discrete, is to estimate a logistic model for treatment separately for each group or, equivalently, with full interactions between G_i and (possibly transformed) covariates $\phi(X_i) \in \mathbb{R}^p$:

$$\text{logit}(e(x, g)) = \alpha_g + \beta_g \cdot \phi(x). \quad (7)$$

Due to the high-dimensional nature of the problem, it is often infeasible to estimate Equation (7) without any regularization: the treated and control units might be completely separated, particularly when some groups are small. Classical propensity score modeling with random effects is one common solution, but can be numerically unstable in settings similar to this

(Zubizarreta and Keele, 2017). Other possible solutions in high dimensions include L^1 penalization (Lee et al., 2019), hierarchical Bayesian modeling (Li et al., 2013), and generalized boosted models (McCaffrey et al., 2004). In addition, Dong et al. (2020) propose a stochastic search algorithm to estimate a similar model when the number of subgroups is large, and Li et al. (2017) and Yang et al. (2021) propose *overlap weights*, which upweight regions of greater overlap.

Under suitable assumptions and conditions, methods utilizing the estimated propensity score will converge to the true ATT asymptotically. However, in high dimensional settings with a moderate number of subgroups, these methods can often fail to achieve good covariate balance in the sample of interest; as we show in Section 4.1, these methods fail to balance covariates in the UC Berkeley LOR study. The key issue is that traditional IPW methods focus on estimating the propensity score itself (i.e., the conditional probability of treatment) rather than finding weights that achieve good in-sample covariate balance.

Balancing weights. Unlike traditional IPW, balancing weights estimators instead find weights that directly target in-sample balance. One example is the Stable Balancing Weights (SBW) proposal from Zubizarreta (2015), which finds the minimum variance weights that achieve a user-defined level of covariate balance in transformed covariates $\phi(X_i) \in \mathbb{R}^p$:

$$\begin{aligned} \min_{\gamma} \quad & \|\gamma\|_2^2 \\ \text{subject to} \quad & \max_j \left| \frac{1}{n_1} \sum_{W_i=1} \phi_j(X_i) - \frac{1}{n_1} \sum_{W_i=0} \gamma_i \phi_j(X_i) \right| \leq \delta, \end{aligned} \tag{8}$$

for weights γ , typically constrained to the simplex, allowable covariate imbalance δ , and transformations of the covariates $\phi_j(\cdot)$, $j = 1, \dots, p$. These methods have a long history in calibrated survey weighting (see, e.g. Deville et al., 1993), and have recently been extensively studied in the observational study context (e.g. Zubizarreta, 2015; Athey et al., 2018; Hirshberg et al., 2019; Hazlett, 2020). They have also been shown to estimate the propensity score with a loss function designed to achieve good balance (Wang and Zubizarreta, 2020).

While balancing weights achieve better balance than the traditional IPW methods above, we must take special care to use them appropriately when estimating subgroup treatment effects. As we will show in Section 4.1, designing balancing weights estimators without explicitly incorporating the subgroup structure also fails to balance covariates within subgroups in the LOR study. We turn to designing such weights in the next section.

3 Balancing weights for treatment effect variation

Now we describe a specialization of balancing weights that minimizes the bias for subgroup treatment effect estimates. This approach incorporates the subgroup structure into the balance measure and optimizes for the “local balance” within each subgroup. First we show that the error for the subgroup treatment effect estimate is bounded by the level of local imbalance within the subgroup. Furthermore, the error for estimating the overall ATT depends on both the global balance and the local balance within each subgroup. We then describe a convex optimization problem to minimize the level of imbalance within each subgroup while ensuring exact global balance in the full sample. Next, we connect the procedure to IPW with a hierarchical propensity score model, using the procedure’s Lagrangian dual formulation. We conclude by describing how to augment the weighting estimate with an outcome model.

3.1 Local balance, global balance, and estimation error

Subgroup effects. We initially consider the role of local imbalance in estimating subgroup treatment effects. This is the subgroup-specific specialization of standard results in balancing weights; see Ben-Michael et al. (2020) for a recent review. We will compare the estimate $\hat{\mu}_{0g}$ to $\tilde{\mu}_{0g} \equiv \frac{1}{n_{1g}} \sum_{G_i=g} W_i m_0(X_i, g)$, our best approximation to μ_{0g} if we knew the true prognostic score. Defining the residual $\varepsilon_i = Y_i - m_0(X_i, G_i)$, the error is

$$\hat{\mu}_{0g} - \tilde{\mu}_{0g} = \underbrace{\frac{1}{n_{1g}} \sum_{G_i=g} \hat{\gamma}_i (1 - W_i) m_0(X_i, g) - \frac{1}{n_{1g}} \sum_{G_i=g} W_i m_0(X_i, g)}_{\text{bias}_g} + \underbrace{\frac{1}{n_{1g}} \sum_{G_i=g} (1 - W_i) \hat{\gamma}_i \varepsilon_i}_{\text{noise}}. \quad (9)$$

Since the weights $\hat{\gamma}$ are *design-based* in the sense of [Rubin \(2008\)](#), they will be independent of the outcomes, and the noise term will be mean-zero and have variance proportional to the sum of the squared weights $\frac{1}{n_{1g}^2} \sum_{G_i=g} (1 - W_i) \hat{\gamma}_i^2$.⁸ At the same time, the conditional bias term, bias_g , depends on the imbalance in the true prognostic score $m_0(X_i, G_i)$. The idea is to bound this imbalance by the worst-case imbalance in all functions m in a model class \mathcal{M} . While the setup is general,⁹ we describe the approach assuming that the prognostic score within each subgroup is a linear function of transformed covariates $\phi(X_i) \in \mathbb{R}^p$ with L^2 -bounded coefficients; i.e., $\mathcal{M} = \{m_0(x, g) = \eta_g \cdot \phi(x) \mid \|\eta_g\|_2 \leq C\}$. We can then bound the bias by the level of *local imbalance* within the subgroup via the Cauchy-Schwarz inequality:

$$|\text{bias}_g| \leq C \underbrace{\left\| \frac{1}{n_{1g}} \sum_{G_i=g} \hat{\gamma}_i (1 - W_i) \phi(X_i) - \frac{1}{n_{1g}} \sum_{G_i=g} W_i \phi(X_i) \right\|_2}_{\text{local imbalance}}. \quad (10)$$

Based on Equation (10), we could control local bias solely by controlling local imbalance. This approach would be reasonable if we were solely interested in subgroup impacts. In practice, however, we are also interested in overall effects and aggregated subgroup effects.

Overall treatment effect. We can estimate aggregated effects by taking a weighted average of the subgroup-specific estimates, e.g., we estimate μ_0 as $\hat{\mu}_0 = \sum_{g=1}^K \frac{n_{1g}}{n_1} \hat{\mu}_{0g} = \frac{1}{n_1} \sum_{W_i=0} n_{1G_i} \hat{\gamma}_i Y_i$. The imbalance within each subgroup continues to play a key role in estimating this overall treatment effect, alongside global balance. To see this, we again compare to our best estimate if we knew the prognostic score, $\tilde{\mu}_0 = \frac{1}{n_1} \sum_{g=1}^K n_{1g} \tilde{\mu}_{0g}$, and see that the

⁸In the general case with heteroskedastic errors, the variance of the noise term is $\frac{1}{n_{1g}^2} \sum_{G_i=g} \hat{\gamma}_i^2 \text{Var}(\varepsilon_i) \leq \max_i \{\text{Var}(\varepsilon_i)\} \frac{1}{n_{1g}^2} \sum_{G_i=g} \hat{\gamma}_i^2$.

⁹See [Wang and Zubizarreta \(2020\)](#) for the case where the prognostic score can only be approximated by a linear function; see [Hazlett \(2020\)](#) for a kernel representation and [Hirshberg et al. \(2019\)](#) for a general nonparametric treatment.

local imbalance plays a part. The error is

$$\begin{aligned} \hat{\mu}_0 - \tilde{\mu}_0 = & \bar{\eta} \cdot \left(\frac{1}{n_1} \sum_{i=1}^n n_{1G_i} \hat{\gamma}_i (1 - W_i) \phi(X_i) - \frac{1}{n_1} \sum_{i=1}^n W_i \phi(X_i) \right) + \\ & \frac{1}{n_1} \sum_{g=1}^k n_{1g} (\eta_g - \bar{\eta}) \cdot \left(\sum_{G_i=g} \hat{\gamma}_i (1 - W_i) \phi(X_i) - \frac{1}{n_{1g}} \sum_{G_i=g} W_i \phi(X_i) \right) + \\ & \frac{1}{n_1} \sum_{i=1}^n \hat{\gamma}_i (1 - W_i) \varepsilon_i, \end{aligned} \quad (11)$$

where $\bar{\eta} \equiv \frac{1}{K} \sum_{g=1}^K \eta_g$ is the average of the model parameters across all subgroups. Again using Cauchy-Schwarz we see that the overall bias is controlled by the *local imbalance* within each subgroup as well as the *global balance* across subgroups:

$$\begin{aligned} |\text{bias}| \leq & \|\bar{\eta}\|_2 \underbrace{\left\| \frac{1}{n_1} \sum_{i=1}^n n_{1G_i} \hat{\gamma}_i (1 - W_i) \phi(X_i) - \frac{1}{n_1} \sum_{i=1}^n W_i \phi(X_i) \right\|_2}_{\text{global balance}} + \\ & \sum_{g=1}^G \frac{n_{1g}}{n_1} \|\eta_g - \bar{\eta}\|_2 \underbrace{\left\| \sum_{G_i=g} \hat{\gamma}_i (1 - W_i) \phi(X_i) - \frac{1}{n_{1g}} \sum_{G_i=g} W_i \phi(X_i) \right\|_2}_{\text{local balance}}. \end{aligned} \quad (12)$$

In general, we will want to achieve *both* good local balance within each subgroup and good global balance across subgroups. Ignoring local balance can incur bias by ignoring heterogeneity in the outcome model across subgroups, while ignoring global balance leaves potential bias reduction on the table. Equation (12) shows that the relative importance of local and global balance for estimating the overall ATT is controlled by the level of similarity in the outcome process across groups. In the extreme case where the outcome process does not vary across groups — i.e., $\eta_g = \bar{\eta}$ for all g — then controlling the global balance is sufficient to control the bias. In the other extreme where the outcome model varies substantially across subgroups — e.g., $\|\eta_g - \bar{\eta}\|_2$ is large for all g — we will primarily seek to control the local imbalance within each subgroup in order to control the bias for the ATT. Typically, we expect that interaction terms are weaker than “main effects,” i.e., $\|\eta_g - \bar{\eta}\|_2 < \|\bar{\eta}\|_2$ (see [Feller](#)

and Gelman, 2015). As a result, our goal is to find weights that prioritize global balance while still achieving good local balance.

3.2 Optimizing for both local and global balance

We now describe a convex optimization procedure to find weights that optimize for local balance while ensuring exact global balance across the sample. The idea is to stratify across subgroups and find approximate balancing weights within each stratum, while still constraining the overall level of balance. To do this, we find weights $\hat{\gamma}$ that solve the following optimization problem:

$$\begin{aligned}
& \min_{\gamma} \quad \sum_{g=1}^K \left\| \sum_{G_i=g, W_i=0} \gamma_i \phi(X_i) - \sum_{G_i=g, W_i=1} \phi(X_i) \right\|_2^2 + \frac{\lambda_g}{2} \sum_{G_i=G, W_i=0} \gamma_i^2 \\
& \text{subject to} \quad \sum_{W_i=0} \gamma_i \phi(X_i) = \sum_{W_i=1} \phi(X_i) \\
& \quad \sum_{G_i=G, W_i=0} \gamma_i = n_{1g} \\
& \quad \gamma_i \geq 0 \quad \forall i = 1, \dots, n
\end{aligned} \tag{13}$$

The optimization problem (13) has several key components. First, following Equation (10) we try to find weights that minimize the local imbalance for each stratum defined by G ; this is a proxy for the stratum-specific bias. We also constrain the weights to *exactly balance* the covariates globally over the entire sample. Equivalently, this finds weights that achieve exact balance marginally on the covariates $\phi(X_i)$ and only approximate balance for the interaction terms $\phi(X_i) \times \mathbb{1}_{G_i}$, placing greater priority on main effects than interaction terms.¹⁰ Taken together, this ensures that we are minimizing the overall bias as well as the

¹⁰ We could extend the optimization problem in Equation (13) to balance intermediate levels between global balance and local balance. Incorporating additional balance constraints for each intermediate level is unwieldy in practice due to the proliferation of hyperparameters. Instead, we can expand $\phi(x)$ to include additional interaction terms between covariates and levels of the hierarchy. We discuss this choice in the letters of recommendation study in Section 4.

bias within each stratum. In principle, weights that exactly balance the covariates within each stratum would also yield exact balance globally. Typically, however, the sample sizes are too small to achieve exact balance within each stratum, and so this combined approach at least guarantees global balance.¹¹

From Equation (12), we can see that if there is a limited amount of heterogeneity in the baseline outcome process across groups, the global exact balance constraint will limit the estimation error when estimating the ATT, even if local balance is relatively poor. In contrast, if there is more heterogeneity, local balance is a higher priority. In principle, incorporating the global balance constraint could lead to worse local balance. However, we show in both the simulations in Appendix C and the analysis of the LOR pilot study in Section 4 that the global constraint leads to negligible changes in the level of local balance and the performance of the subgroup estimators, but can lead to large improvements in the global balance and the performance of the overall estimate. Thus, there seems to be little downside in terms of subgroup estimates from an approach that controls both local and global imbalance — but large potential gains for overall estimates.¹² Note that while we choose to enforce exact global balance, we could also limit to *approximate* global balance, with the relative importance of local and global balance controlled by an additional hyperparameter set by the analyst.

Second, we include an L^2 regularization term that penalizes the sum of the squared weights in the stratum; from Equation (9), we see that this is a proxy for the variance of the weighting estimator. For each stratum, the optimization problem includes a hyper-parameter λ_g that negotiates the bias-variance tradeoff within that stratum. When λ_g is small, the optimization prioritizes minimizing the bias through the local imbalance, and when λ is large it prioritizes minimizing the variance through the sum of the squared weights. As a heuristic, we limit the number of hyperparameters by choosing $\lambda_g = \frac{\lambda}{n_g}$ for a common choice of λ . For larger strata where better balance is possible, this heuristic will prioritize balance

¹¹This constraint induces a dependence across the strata, so that the optimization problem does not decompose into J sub-problems.

¹²In Appendix B.1 we also show that if the estimand is the *difference* in treatment effects between two subgroups there are also possible gains to balancing the difference in covariate values across the two groups.

— and thus bias — over variance; for smaller strata, by contrast, this will prioritize lower variance. We discuss selecting λ in the letters of recommendation study in Section 4.1.

Next, we incorporate two additional constraints on the weights. We include a fine balance constraint (Rosenbaum et al., 2007): within each stratum the weights sum up to the number of treated units in that stratum, n_{1g} . Since each stratum maps to only one subgroup, this guarantees that the weights sum to the number of treated units in each subgroup. We also restrict the weights to be non-negative, which stops the estimates from extrapolating outside of the support of the control units. Together, these induce several stability properties, including that the estimates are sample bounded.

Finally, we compute the variance of our estimator conditioned on the design $(X_1, Z_1, W_1), \dots, (X_n, Z_n, W_n)$ or, equivalently, conditioned on the weights. The conditional variance is

$$\text{Var}(\hat{\mu}_{0g} \mid \hat{\gamma}) = \frac{1}{n_{1g}^2} \sum_{G_i=g} (1 - W_i) \hat{\gamma}_i^2 \text{Var}(Y_i). \quad (14)$$

Using the i^{th} residual to estimate $\text{Var}(Y_i)$ yields the empirical sandwich estimator for the treatment effect

$$\widehat{\text{Var}}(\hat{\mu}_{1g} - \hat{\mu}_{0g} \mid \hat{\gamma}) = \frac{1}{n_{1g}^2} \sum_{G_i=g} W_i (Y_i - \hat{\mu}_{1g})^2 + \frac{1}{n_{1g}^2} \sum_{G_i=g} (1 - W_i) \hat{\gamma}_i^2 (Y_i - \hat{\mu}_{0g})^2, \quad (15)$$

where, as above, $\hat{\mu}_{1g}$ is the average outcome for applicants in subgroup g who submit LORs. This is the fixed-design Huber-White heteroskedastic robust standard error for the weighted average. See Hirshberg et al. (2019) for discussion on asymptotic normality and semi-parametric efficiency for estimators of this form.

3.3 Dual relation to partially pooled propensity score estimation

Thus far, we have motivated the approximate balancing weights approach by appealing to the connection between local bias and local balance. We now draw on recent connections between approximate balancing weights and (calibrated) propensity score estimation through the Lagrangian dual problem. The weights that solve optimization problem (13) correspond

to estimating the inverse propensity weights with a (truncated) linear odds function with the stratum G interacted with the covariates $\phi(X)$,¹³

$$\frac{P(W = 1 \mid X = x, G = g)}{1 - P(W = 1 \mid X = x, G = g)} = [\alpha_g + \beta_g \cdot \phi(x)]_+, \quad (16)$$

where $[x]_+ = \max\{0, x\}$ and the coefficients β_g are *partially pooled* towards a global model.

To show this, we first derive the Lagrangian dual. For each stratum g , the sum-to- n_{1g} constraint induces a dual variable $\alpha_g \in \mathbb{R}$, and the local balance measure induces a dual variable $\beta_g \in \mathbb{R}^p$. These dual variables are part of the *balancing loss function* for stratum z :

$$\mathcal{L}_g(\alpha_g, \beta_g) \equiv \sum_{W_i=0, G_i=g} [\alpha_g + \beta_g \cdot \phi(X_i)]_+^2 - \sum_{W_i=1, G_i=g} (\alpha_g + \beta_g \cdot \phi(X_i)), \quad (17)$$

With this definition we can now state the Lagrangian dual.

Proposition 1. With $\lambda_g > 0$, if a feasible solution to (13) exists, the Lagrangian dual is

$$\min_{\alpha, \beta_1, \dots, \beta_J, \mu_\beta} \underbrace{\sum_{g=1}^K \mathcal{L}_g(\alpha_g, \beta_g)}_{\text{balancing loss}} + \underbrace{\sum_{z=1}^J \frac{\lambda_g}{2} \|\beta_g - \mu_\beta\|_2^2}_{\text{shrinkage to global variable}}. \quad (18)$$

If $\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_J$ are the solutions to the dual problem, then the solution to the primal problem (13) is

$$\hat{\gamma}_i = \left[\hat{\alpha}_{Z_i} + \hat{\beta}_{Z_i} \cdot \phi(X_i) \right]_+. \quad (19)$$

The Lagrangian dual formulation sheds additional light on the approximate balancing weights estimator. First, we apply results on the connection between approximate balancing weights and propensity score estimation (e.g., Wang and Zubizarreta, 2020; Hirshberg and Wager, 2021). We see that this approach estimates propensity scores of the form (16), which corresponds to a fully interacted propensity score model where the coefficients on observed

¹³The truncation arises from constraining weights to be non-negative, and the linear odds form arises from penalizing the L^2 norm of the weights. We can consider other penalties that will lead to different forms; In particular, with an entropy penalty, the weights are linear in the log-odds. See Ben-Michael et al. (2020) for a review of the different choices.

covariates vary across strata. Recall that we find *approximate* balancing weights for each stratum because the number of units per stratum might be relatively small; therefore we should not expect to be able to estimate this fully interacted propensity score well.

The dual problem in Equation (18) also includes a global dual variable μ_β induced by the global balance constraint in the primal problem (13). Because we enforce *exact* global balance, this global model is not regularized. However, by penalizing the deviations between the stratum-specific variables and the global variables via the L^2 norm, $\|\beta_g - \mu_\beta\|_2^2$, the dual problem *partially pools* the stratum-specific parameters towards a global model. Thus, we see that the approximate balancing weights problem in Equation (13) corresponds to a hierarchical propensity score model (see, e.g. Li et al., 2013), as in Section 2.2, fit with a loss function designed to induce covariate balance.

Excluding the global constraint removes the global dual variable μ_β , and the dual problem shrinks the stratum-specific variables β_g towards zero without any pooling. In contrast, ignoring the local balance measure by setting $\lambda_g \rightarrow \infty$ constrains the stratum-specific variables β_g to all be *equal* to the global variable μ_β , resulting in a fully pooled estimator. For intermediate values, λ_g controls the level of partial pooling. When λ_g is large the dual parameters are heavily pooled towards the global model, and when λ_g is small the level of pooling is reduced. By setting $\lambda_g = \frac{\lambda}{n_g}$ as above, larger strata will be pooled less than smaller strata.¹⁴

3.4 Augmentation with an outcome estimator

The balancing weights we obtain via the methods above may not achieve perfect balance, leaving the potential for bias. We can augment the balancing weights estimator with an outcome model, following similar proposals in a variety of settings (see, e.g. Athey et al., 2018; Hirshberg and Wager, 2021; Ben-Michael et al., 2021). Analogous to bias correction for matching (Rubin, 1973) or model-assisted estimation in survey sampling (Särndal et al.,

¹⁴It is also possible to have covariate-specific shrinkage by measuring imbalance in the primal problem (13) with a *weighted* L^2 norm, leading to an additional p hyper-parameters. We leave exploring this extension and hyper-parameter selection methods to future work.

2003), the essential idea is to adjust the weighting estimator using an estimate of the bias. Specifically, we can estimate the prognostic score $m_0(x, g)$ with a working model $\hat{m}_0(x, g)$, e.g., with a flexible regression model. An estimate of the bias in group g is then:

$$\widehat{\text{bias}}_g = \frac{1}{n_{1g}} \sum_{W_i=1, G_i=g} \hat{m}_0(X_i, g) - \frac{1}{n_{1g}} \sum_{W_i=0, G_i=g} \hat{\gamma}_i \hat{m}_0(X_i, g). \quad (20)$$

This is the bias due to imbalance in estimated prognostic score in group g *after* weighting. With this estimate of the bias, we can explicitly bias-correct our weighting estimator, estimating μ_{0g} as

$$\begin{aligned} \hat{\mu}_{0g}^{\text{aug}} &\equiv \hat{\mu}_{0g} + \widehat{\text{bias}}_g \\ &= \frac{1}{n_{1g}} \sum_{W_i=0, G_i=g} \hat{\gamma}_i Y_i + \left[\frac{1}{n_{1g}} \sum_{W_i=1, G_i=g} \hat{m}_0(X_i, g) - \frac{1}{n_{1g}} \sum_{W_i=0, G_i=g} \hat{\gamma}_i \hat{m}_0(X_i, g) \right]. \end{aligned} \quad (21)$$

Thus, if the balancing weights fail to achieve good covariate balance in a given subgroup, the working outcome model, $\hat{m}_0(X_i, g)$, can further adjust for any differences. See [Ben-Michael et al. \(2020\)](#) for further discussion.

4 Differential impacts of letters of recommendation

We now turn to estimating the differential impacts of letters of recommendation on admissions decisions. We focus on the eight subgroups defined in Table [D.1](#), based on the interaction between URM status (2 levels) and admissibility index (4 levels). Due to the selection mechanism described in Section [1.1](#), however, it is useful to create even more fine-grained strata and then aggregate to these eight subgroups. Specifically, we define $G = 41$ fine-grained strata based on URM status, AI grouping, first reader score, and college applied to.¹⁵ While we are not necessarily interested in treatment effect heterogeneity across all 41 strata, this allows us to exactly match on key covariates and then aggregate to obtain the

¹⁵Of the 48 possible strata, we drop 7 strata where no applicants submitted a letter of recommendation. These are non-URM applicants in both colleges in the two lowest AI strata but where the first reader assigned a “Yes” or “No”. This accounts for $\sim 2\%$ of applicants. The remaining 41 strata have a wide range of sizes with a few very large strata. Min: 15, p25: 195, median: 987, p75: 1038, max: 8000.

primary subgroup effects.

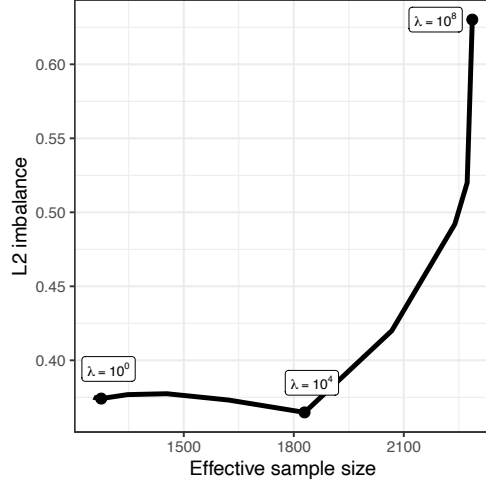
Another key component in the analysis is the choice of transformation of the covariates $\phi(\cdot)$. Because we have divided the applicants into many highly informative strata, we choose $\phi(\cdot)$ to include all of the raw covariates. Additionally, because of the importance of the admissibility index, we also include a natural cubic spline for AI with knots at the sample quantiles.¹⁶ Finally, we include the output of the admissions model and a binary indicator for whether the predicted probability of a “Possible” score is greater than 50%. We further prioritize local balance in the admissibility index by including in $\phi(x)$ the interaction between the AI, URM status, and an indicator for admissibility subgroup; this ensures local balance in the admissibility index at an intermediate level of the hierarchy between global balance and local balance. Finally, we standardize each component of $\phi(X)$ to have mean zero and variance one. If desired, we could also consider other transformations such as a higher-order polynomial transformation, using a series of basis functions for all covariates, or computing inner products via the kernel trick to allow for an infinite dimensional basis (see, e.g. [Hazlett, 2020](#); [Wang and Zubizarreta, 2020](#); [Hirshberg and Wager, 2021](#)).

4.1 Diagnostics: local balance checks and assessing overlap

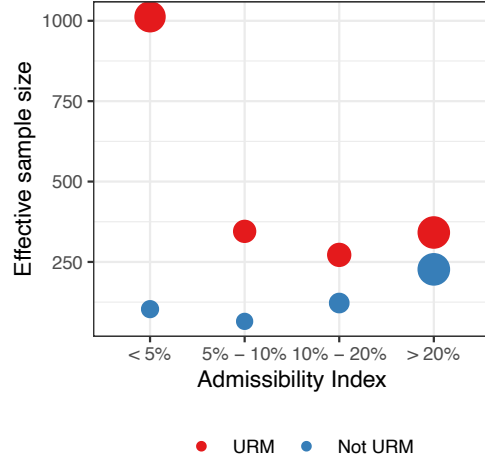
In order to estimate effects, we must first choose values of the common hyperparameter λ in the optimization problem (13), where we set $\lambda_g = \frac{\lambda}{n_g}$. Recall that this hyperparameter negotiates the bias-variance tradeoff: small values of λ will prioritize bias by reducing local imbalance while higher values will prioritize variance by increasing the effective sample size. Figure 3a shows this tradeoff. We plot the square root of the local balance measure in (13) against the *effective sample size* for the re-weighted control group, $n_1 / (\sum_{W_i=0} \hat{\gamma}_i^2)$. Between $\lambda = 10^0$ and 10^4 , we see that the imbalance is relatively flat while the overall effective sample size increases, after which the imbalance increases quickly with λ . We therefore select $\lambda = 10^4$ for the results we present.

Figure 3b shows the effective control group sample size for each of the primary URM and

¹⁶ Since the AI includes many of the raw covariates that are balanced separately, we can view this approach as a useful dimension reduction strategy for balancing an otherwise exponential number of possible combinations and interactions.



(a) Imbalance vs effective sample size.
 $\lambda = 1, 10^4, 10^8$ noted.



(b) Effective sample sizes, area proportional to
number of treated units

Figure 3: (a) Imbalance measured as the square root of the objective in (13) plotted against the effective sample size of the overall control group. (b) Effective sample size of the control group for each subgroup, with weights solving Equation (13) with $\lambda_g = \frac{10^4}{n_g}$.

AI subgroups, scaled by the number of applicants in the group submitting LORs. Across the board, the URM subgroups have larger effective sample sizes than the non-URM subgroups, with particularly stark differences for the lower AI subgroups. For all non-URM subgroups the effective sample size is less than 250. Comparing to the sample sizes in Appendix Table D.1, we see that the weighting approach leads to a large design effect: many applicants who did not submit LORs are not comparable to those who did. However, lower admissibility non-URM applicants also submitted letters at lower rates. This design effect, combined with the smaller percentage of non-URM applicants submitting LORs, means that we should expect to have greater precision in the estimates for URM applicants than non-URM applicants.

We now assess the level of local balance within each subgroup, following the discussion in Section 3.1. We focus on three estimators: fully- and partially-pooled balancing weights, which solve Equation (13) with $\lambda_g \rightarrow \infty$ and $\lambda_g = \frac{10^4}{n_{1g}}$, respectively; and traditional IPW with a fully-interacted propensity score model. See Appendix C for complete descriptions. Figure 4 shows the distribution of the imbalance in each of the 51 (standardized) components of $\phi(X)$. The fully interacted IPW approach has very poor balance overall, due in part to the

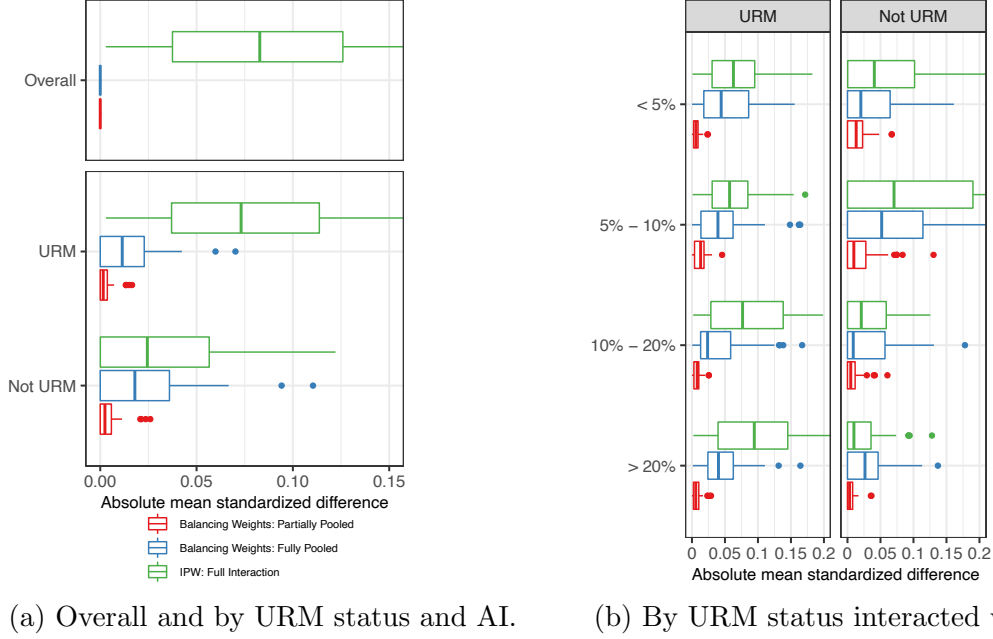


Figure 4: Distribution of imbalance in each component of $\phi(X)$ after weighting, for partially- and fully-pooled balancing weights and fully interacted IPW estimator.

difficulty of estimating the high-dimensional propensity score model. As expected, both the fully- and partially-pooled balancing weights achieve perfect balance overall; however, only the partially pooled balancing weights achieve excellent local balance. Appendix Figure D.4 shows these same metrics for the no-pooled balancing weights and fixed effects IPW estimators we discuss in Appendix C, as well as subgroup overlap weights (Yang et al., 2021). The partially- and no-pooled approaches have similar global and local balance overall, but the partially-pooled approach sacrifices a small amount of local balance for an improvement in global balance. In contrast, both the fixed effects IPW and overlap weights approaches yield poor local balance.

Finally, we assess overlap within each subgroup. A key benefit of weighting approaches is that overlap issues manifest in the distribution of our weights $\hat{\gamma}$. Appendix Figure D.6 plots the distribution of the weights over the comparison applicants by URM status and AI group, normalized by the number of treated applicants in the subgroup. The vast majority of control units receive zero weight and are excluded from the figure. Of the 28,556 applicants

who did not submit LORs, only 9,834 (34%) receive a weight larger than 0.001. This is indicative of a lack of “left-sided” overlap: very many applicants who did not submit a letter of recommendation had nearly zero odds of doing so in the pilot program. This is problematic for estimating the overall average treatment effect, but is less of a concern when we focus on estimating the average treatment effect on the treated.

For each AI subgroup we also see that the distribution of weights is skewed more positively for the non-URM applicants. In particular, for the lower AI, non-URM subgroups we see a non-trivial number of comparison applicants that “count for” over 1% of the re-weighted sample, with a handful of outliers that count for more than 2%. While large weights do not necessarily affect the validity of the estimator, large weights decrease the effective sample size, reducing the precision of our final estimates, as we see in Figure 3b.

4.2 Treatment effect estimates

After assessing local balance and overlap, we can now turn to estimating the differential impacts of letters of recommendation. Figure 5 shows the ATT, $\hat{\mu}_{1g} - \hat{\mu}_{0g}$; Appendix Figure D.7 gives the corresponding means. The standard errors are computed via the sandwich estimator in Equation (15). Overall, we estimate that LORs increased admission rates by 5 percentage points (pp). While we estimate a larger effect for non-URM applicants (6.2 pp) than URM applicants (4.5 pp), there is insufficient evidence to distinguish between the two effects. Second, we see a roughly positive trend between treatment effects and the AI, potentially with a peak for the 10%-20% group. This is driven by the very small estimated effect for applicants with $AI < 5\%$, who are very unlikely to be accepted with or without LORs. LORs thus seem to have a larger effect for applicants closer to the cusp of acceptance.

The right panel of Figure 5 further stratifies the subgroups, showing the effects jointly by URM status and AI. While the point estimate for the overall increase in admission rates is slightly larger for non-URM applicants than for URM applicants, this is mainly a composition effect. For applicants very unlikely to be admitted ($AI < 5\%$) the point estimates are nearly identical for URM and non-URM applicants, although the URM subgroup is estimated much more precisely. For the next two levels of the admissibility index (AI between 5% and 20%),

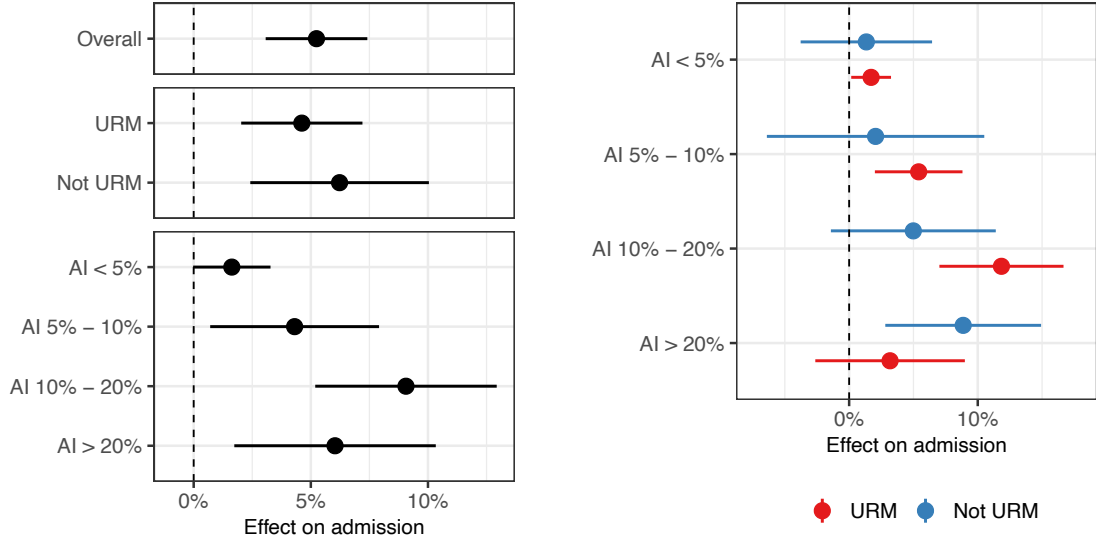


Figure 5: Estimated treatment effects of letters of recommendation on admission \pm two standard errors: overall, by URM status, by Admissibility Index, and by URM \times AI.

URM applicants have a higher estimated impact, with imprecise estimates for non-URM applicants. For the highest admissibility groups (AI > 20%), non-URM applicants have larger positive effects, though again these estimates are noisy. Since URM applicants have lower AI on average, the overall estimate is also lower for URM applicants. Furthermore, the peak in the effect for middle-tier applicants is more pronounced for URM applicants than non-URM applicants. From Appendix Figure D.7 we see that this is primarily because high admissibility URM applicants have very high imputed admission rates.

Appendix A.1 discusses results using alternative estimators on the same data. First, as in Section 3.4, we consider two flexible outcome models, ridge regression and random forests, both alone and in combination with weighting. We can view these two outcome models as creating a new admissibility index directly from the 2016 cohort — with more flexible modeling that accounts for the subgroups of interest — and using the predictions to impute the counterfactual means. Both the augmented estimates and the estimates from outcome modeling alone are broadly consistent with the estimates from weighting alone, although the estimated bias for the highest AI subgroup is large. We also compare our main results

with simple estimates based on linear regression and based on plugging in the AI. Both approaches diverge from our main results for the highest admissibility non-URM applicants, a group with limited overlap. For other subgroups, the estimates are similar for the AI-based estimates, but differ for the linear regression estimates, which has poor control over local imbalance.

Appendix [A.2](#) instead considers estimates based on alternative data and sample definitions. First, we consider effects on an intermediate outcome: whether the second reader — who has access to the LORs — gives a “Yes” score. Because our weights are design-based, 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. We also explore an alternative approach that instead leverages unique features of the UC Berkeley pilot study, which included an additional review without the letters of recommendation from a sample of 10,000 applicants. These results are broadly similar to the estimates from the observational study, again differing mainly in regions with relatively poor overlap. Finally, we repeat our primary analysis but restrict the sample to those applicants who were invited to submit LORs; we find similar point estimates to those from the full sample.¹⁷ Interestingly, even though there are far fewer comparison applicants in the restricted sample (3,452 vs. 29,398) the standard errors are only slightly larger. This reflects the fact that applicants never invited to submit LORs receive less than 25% of the total weight in our main analysis. Thus, applicants who submit LORs are more similar to those invited to submit but did not than they are to those never invited.

Finally, Appendix [A.3](#) describes a formal sensitivity analysis for violations of the ignorability assumption (Assumption 1), adapting a recent proposal from [Soriano et al. \(2020\)](#). Using this approach we conclude that there would need to be substantial unmeasured confounding, of roughly the same predictive power as the AI, to qualitatively change our conclusions.

¹⁷Recall 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. This analysis excludes this first group.

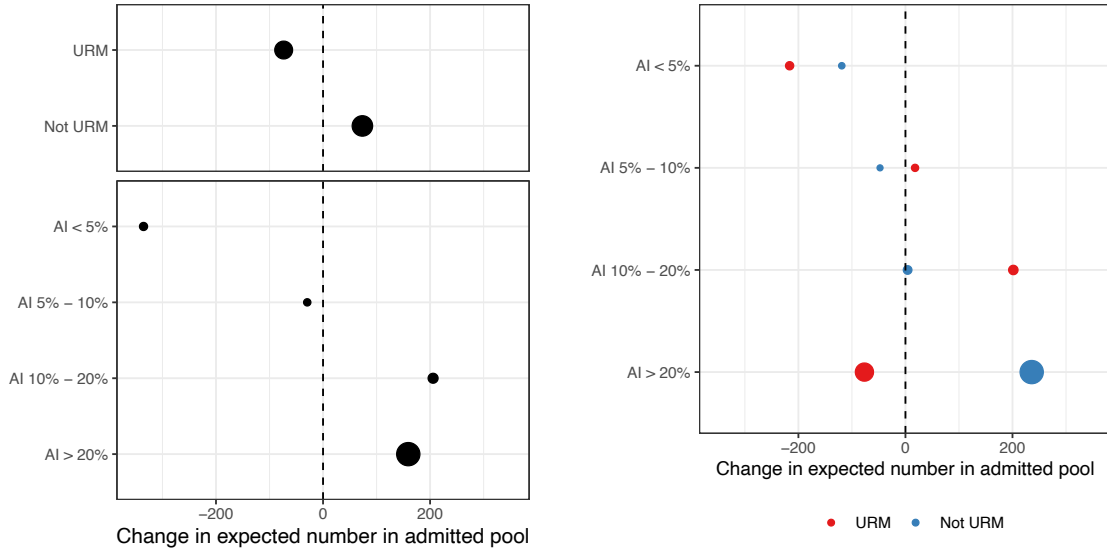


Figure 6: Projected change in number of admitted applicants if all applicants submitted a letter of recommendation: by URM status, by Admissibility Index, and by URM \times AI. Points are scaled by the expected size of the admitted group without submitting LORs; e.g., applicants with high AI are a much larger share of the admitted class than applicants with low AI.

Taken together, our main results and accompanying sensitivity checks paint a relatively clear picture of differential impact of letters of recommendation across applicants’ *a priori* application strength. Treatment effects are low for applicants who are unlikely to be accepted and high for applicants on the margin for whom letters provide useful context, with some evidence of a dip for the highest admissibility applicants. Our estimates of differential impacts between URM and non-URM students are more muddled, due to large sampling errors, and do not support strong conclusions. Point estimates indicate that LORs benefit URM applicants more than they do non-URM applicants at all but the highest academic indexes. Because non-URM applicants are overrepresented in the high-AI category, the point estimate for the average treatment effect is larger for non-URMs; however, there is insufficient precision to distinguish between the two groups.

4.3 Simulation of universal LOR policy

Thus far, we have considered 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 (see [Chalfant, 2017](#); [UC Berkeley, 2017](#)). 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;¹⁸ (2) adjust these probabilities to set the expected number of admitted students under each policy to the observed number;¹⁹ 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 *differential* impacts of LORs is especially relevant for this exercise; see also Section [B.1](#).

Figure [6](#) 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

¹⁸We 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.

¹⁹A 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.

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.

5 Discussion

Estimating heterogeneous treatment effects and assessing treatment effect variation in observational studies is a challenge, even for pre-specified subgroups. Focusing on weighting estimators, we show that the estimation error depends on the level of *local imbalance* between the treated and control groups after weighting. We then present a convex optimization problem that finds approximate balancing weights that directly target the level of local imbalance within each subgroup, while ensuring exact global balance to also estimate the overall effect. Using this method to estimate heterogeneous effects in the UC Berkeley letters of recommendation pilot study, we find evidence that letters of recommendation lead to better admissions outcomes for stronger applicants, with mixed evidence of differences between URM and non-URM applicants.

There are several directions for future methodological work. First, an important limitation of our approach is that subgroups are defined by discrete covariates, requiring us to discretize important continuous measures such as the Admissibility Index. A possible extension is to adapt the recent proposal from [Wang et al. \(2021\)](#) to combine balancing weights and traditional kernel weighting methods in order to estimate a conditional average treatment effect function.

Second, hyperparameter selection for balancing weights estimators is an important question in practice but remains an open problem. We elect to choose the hyperparameter by explicitly tracing out the level of balance and effective sample size as the hyper-parameter changes. However, cross validation approaches such as that proposed by [Wang and Zubizarreta \(2020\)](#) may have better properties. This is an important avenue for future work.

Third, we directly estimate the effect of submitting LORs among those who submit.

However, we could instead frame the question in terms of non-compliance and use the *invitation* to submit LORs as an instrument for submission. Using the approximate balancing weights procedure described above we could adjust for unequal invitation probabilities, and estimate the effect on compliers via weighted two-stage least squares.

Finally, we could consider applying this approach to subgroup effects in randomized trials, adapting recent proposals from, among others, [Zeng et al. \(2021\)](#) and [Yang et al. \(2021\)](#). We expect that the overlap will be substantially improved in this setting, leading to larger effective sample size, even if the randomized trials themselves are smaller. We also anticipate that the design-based approach we advocate here will help researchers navigate Simpsons’ Paradox-type challenges in subgroup effects ([VanderWeele and Knol, 2011](#)).

References

- Alvero, A., S. Giebel, B. Gebre-Medhin, A. Antonio, M. Stevens, and B. Domingue (2021). Essay content is strongly related to household income and sat scores: Evidence from 60,000 undergraduate applications. Technical report. Stanford Center for Education Policy Analysis Working Paper.
- Anoke, S. C., S.-L. Normand, and C. M. Zigler (2019). Approaches to treatment effect heterogeneity in the presence of confounding. *Statistics in medicine* 38(15), 2797–2815.
- Athey, S., G. W. Imbens, and S. Wager (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80(4), 597–623.
- Ben-Michael, E., A. Feller, and J. Rothstein (2021). The augmented synthetic control method. *Journal of the American Statistical Association* (just-accepted), 1–34.
- Ben-Michael, E., D. Hirschberg, A. Feller, and J. Zubizarreta (2020). The balancing act for causal inference.
- Bickel, P. J., E. A. Hammel, and J. W. O’Connell (1975). Sex bias in graduate admissions: Data from Berkeley. *Science* 187(4175), 398–404.
- Carvalho, C., A. Feller, J. Murray, S. Woody, and D. Yeager (2019). Assessing Treatment Effect Variation in Observational Studies: Results from a Data Challenge. *Observational Studies* 5, 21–35.
- Chalfant, J. (2017). Letter from Jim Chalfant, Chair of the Academic Senate, to Janet Napolitano. June 20, 2017.
- Deville, J. C., C. E. Särndal, and O. Sautory (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association* 88(423), 1013–1020.
- Dong, J., J. L. Zhang, S. Zeng, and F. Li (2020). Subgroup balancing propensity score. *Statistical Methods in Medical Research* 29(3), 659–676.

- Feller, A. and A. Gelman (2015). Hierarchical models for causal effects. *Emerging Trends in the Social and Behavioral Sciences: An interdisciplinary, searchable, and linkable resource*, 1–16.
- Green, K. M. and E. A. Stuart (2014). Examining moderation analyses in propensity score methods: Application to depression and substance use. *Journal of consulting and clinical psychology* 82(5), 773.
- 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.
- Hazlett, C. (2020). Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects. *Statistica Sinica*.
- Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20(1), 217–240.
- Hirshberg, D. and S. Wager (2021). Augmented Minimax Linear Estimation. *Annals of Statistics*.
- Hirshberg, D. A., A. Maleki, and J. Zubizarreta (2019). Minimax Linear Estimation of the Retargeted Mean.
- Hout, M. (2005). Berkeley’s comprehensive review method for making freshman admissions decisions: An assessment. Technical report, University of California, Berkeley.
- Kuncel, N. R., R. J. Kochevar, and D. S. Ones (2014). A meta-analysis of letters of recommendation in college and graduate admissions: Reasons for hope. *International Journal of Selection and Assessment* 22(1), 101–107.
- Künzel, S. R., J. S. Sekhon, P. J. Bickel, and B. Yu (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America* 116(10), 4156–4165.
- Lee, Y., T. Q. Nguyen, and E. A. Stuart (2019). Partially Pooled Propensity Score Models for Average Treatment Effect Estimation with Multilevel Data.
- Li, F., K. L. Morgan, and A. M. Zaslavsky (2017). Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association* 113521, 1–11.
- Li, F., A. M. Zaslavsky, and M. B. Landrum (2013). Propensity score weighting with multilevel data. *Statistics in Medicine* 32(19), 3373–3387.
- McCaffrey, D. F., G. Ridgeway, and A. R. Morral (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods* 9(4), 403.
- Nie, X. and S. Wager (2017). Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*.
- Rosenbaum, P. R., R. N. Ross, and J. H. Silber (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association* 102(477), 75–83.
- Rothstein, J. (2017, July). The impact of letters of recommendation on UC Berkeley admissions in the 2016-17 cycle. Technical report, California Policy Lab.
- Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 185–203.

- Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics* 2(3), 808–840.
- Särndal, C.-E., B. Swensson, and J. Wretman (2003). *Model assisted survey sampling*. Springer Science & Business Media.
- Schmader, T., J. Whitehead, and V. H. Wysocki (2007). A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex roles* 57(7-8), 509–514.
- 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.
- UC Berkeley (2017). Notice of Meeting: May 11, 2017. University Committee on Affirmative Action, Diversity, and Equity.
- VanderWeele, T. J. and M. J. Knol (2011). Interpretation of subgroup analyses in randomized trials: heterogeneity versus secondary interventions. *Annals of internal medicine* 154(10), 680–683.
- Wang, J., R. K. Wong, S. Yang, and K. C. G. Chan (2021). Estimation of partially conditional average treatment effect by hybrid kernel-covariate balancing. *arXiv preprint arXiv:2103.03437*.
- Wang, Y. and J. R. Zubizarreta (2020). Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika* 107(1), 93–105.
- Yang, S., F. Li, L. E. Thomas, and F. Li (2021). Covariate adjustment in subgroup analyses of randomized clinical trials: A propensity score approach. *Clinical Trials*, 17407745211028588.
- Yang, S., E. Lorenzi, G. Papadogeorgou, D. M. Wojdyla, F. Li, and L. E. Thomas (2021). Propensity score weighting for causal subgroup analysis. *Statistics in medicine*.
- Zeng, S., F. Li, R. Wang, and F. Li (2021). Propensity score weighting for covariate adjustment in randomized clinical trials. *Statistics in medicine* 40(4), 842–858.
- Zubizarreta, J. R. (2015). Stable Weights that Balance Covariates for Estimation With Incomplete Outcome Data. *Journal of the American Statistical Association* 110(511), 910–922.
- Zubizarreta, J. R. and L. Keele (2017). Optimal multilevel matching in clustered observational studies: A case study of the effectiveness of private schools under a large-scale voucher system. *Journal of the American Statistical Association* 112(518), 547–560.

Supplementary materials for
“Varying impacts of letters of recommendation on college
admissions: Approximate balancing weights for subgroup effects in
observational studies”

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 Alternative estimation approaches

We now consider alternative estimation approaches using the same sample and data definitions as in the primary analysis.

A.1.1 Outcome modeling alone

First, we can see how our main estimates contrast with simple comparisons between those who submitted letters and those who did not. Appendix Figure D.11 shows effects estimated by comparing the admission rate for the treatment group to the admission rate expected by the AI. This comparison shows a large difference between the effects for URM and non-URM applicants, with a *negative* estimated effect on non-URM applicants. This is primarily driven by a large negative estimate for the highest admissibility non-URM applicants, with the other estimates roughly in line with our main estimates. However, as we discuss in Section 1.1, we do not believe that the subgroup effect estimates using the AI in this way are reliable.

Appendix Figure D.11 also shows simple linear regression estimates of the effects, regressing admission on treatment and an additional set of terms from the transformed covariates $\phi(X)$. We estimate effects first with treatment alone, then treatment interacted by URM status, interacted by AI group, and interacted by both. This approach does not directly control for local imbalances and relies on a correctly specified linear additive specification. Because of this, off-the-shelf linear regression disagrees with our main estimates, estimating a smaller overall effect and no effect for non-URM applicants. This is driven by a negative effect for low admissibility non-URM applicants, a region of limited overlap as specified above.

Next, we consider treatment effects estimated via automatic, flexible machine learning approaches that give no special precedence to our pre-defined subgroups. First, Appendix Figure

D.11 shows the result of using ridge regression and random forest estimators to impute the counterfactual means $\hat{\mu}_{0g}$ for each subgroup. We can view these two estimates as creating a new admissibility index directly from the 2016 cohort — with more flexible models that account for the subgroups of interest — and using the predictions to impute the counterfactual means. Because the partially-pooled weights achieve excellent local balance, the ridge and random forest estimates broadly comport with the weighting estimates. However, both models lead to more precise estimates for the subgroups with lower effective sample sizes and fewer comparable control units.

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 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.1.2 Augmented estimation

We next consider augmenting the weighting estimator with an estimate of the prognostic score, $\hat{m}(x, g)$. In Appendix Figure D.10 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.

Appendix Figure D.9 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.

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:

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

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.

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

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 3.1, the estimation error for the difference is

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

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 Formalizing the no interference assumption

As we discuss in Section 1.1, we believe that it is reasonable to assume that one applicant's treatment or outcome does not affect another applicant's outcome in the LOR pilot study. Here, we 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:

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

In Section 4.3, 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 $X_i \stackrel{iid}{\sim} N(0, \tilde{\Sigma})$. We then transform some of these covariates. For $j = 1, 11, 21, 32, 41$ we dichotomize the variable and define $\tilde{X}_{ij} = \mathbb{1}\{X_{ij} \geq q_{.8}(X_{.j})\}$, where $q_{.8}(X_{.j})$ is the 80th percentile of X_j among the n units. For $j = 2, 7, 12, \dots, 47$ 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),²

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

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

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

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 $\alpha_g \stackrel{iid}{\sim} N(0, 1)$ and $\eta_{0g} \stackrel{iid}{\sim} N(0, 1)$. 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 $U_g^\beta, U_g^\eta \stackrel{iid}{\sim} MVN(0, I_d)$. 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 $\tau_i = X_{id} - X_{i3} + 0.3X_{id}X_{i3}$ 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 $j = 1, \dots, m$ 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

²The 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 $\lambda \rightarrow \infty$, 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.

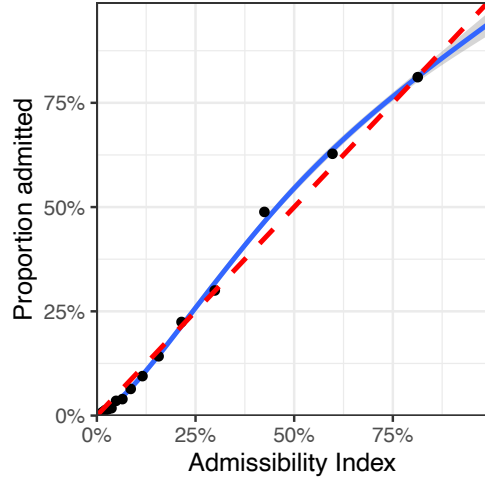
The partially-pooled balancing weights estimator, which is transparent and design-based, also performs nearly as well as the black-box BCF method. Augmenting the partially-pooled weights provides some small improvements to the bias, indicating that the weights alone are able to achieve good balance in 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.

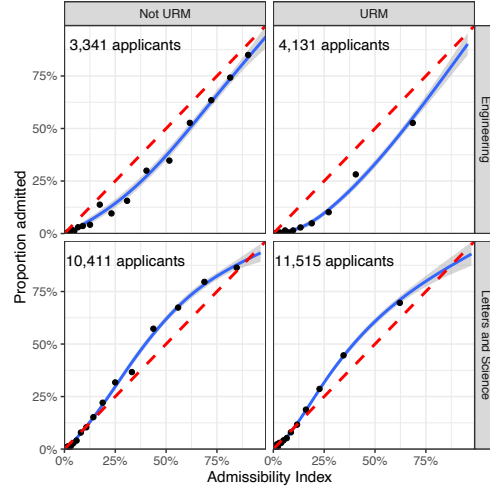
D Additional figures and tables

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%

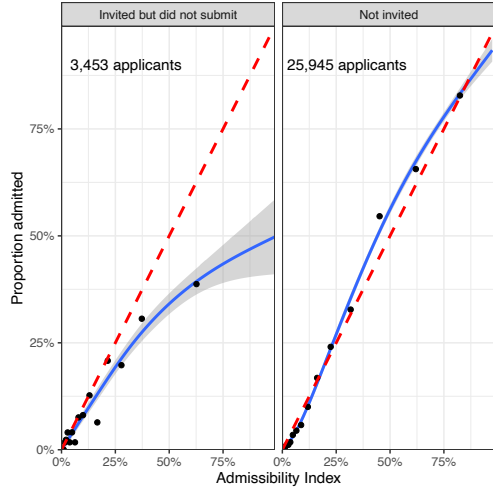
Table D.1: Number of applicants and proportion treated by subgroup.



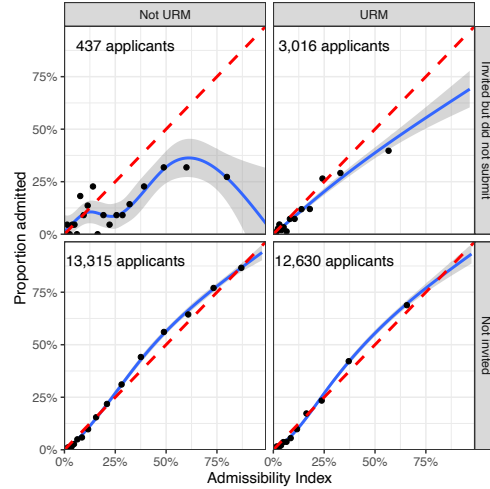
(a) Full sample



(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%.

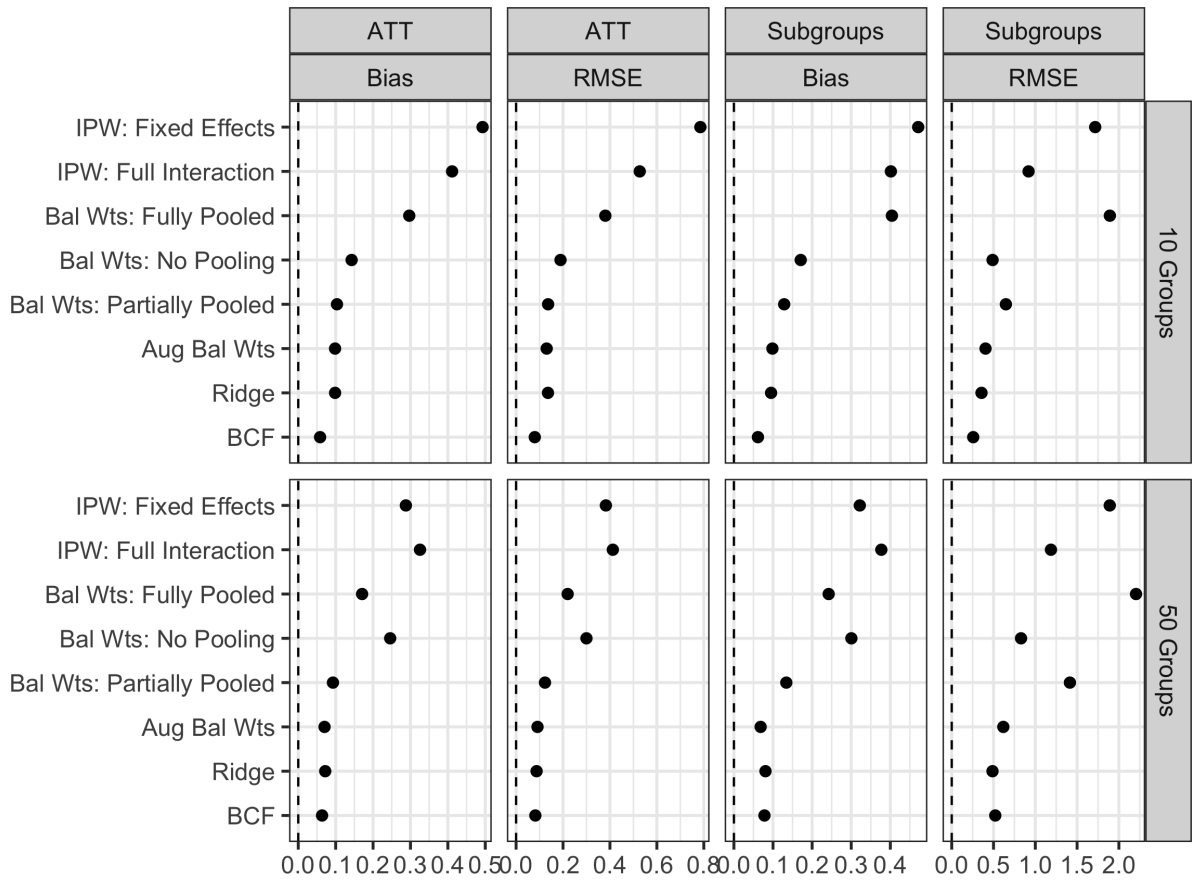


Figure D.2: Performance of approximate balancing weights, traditional IPW with logistic regression, and outcome modelling for estimating subgroup treatment effects.

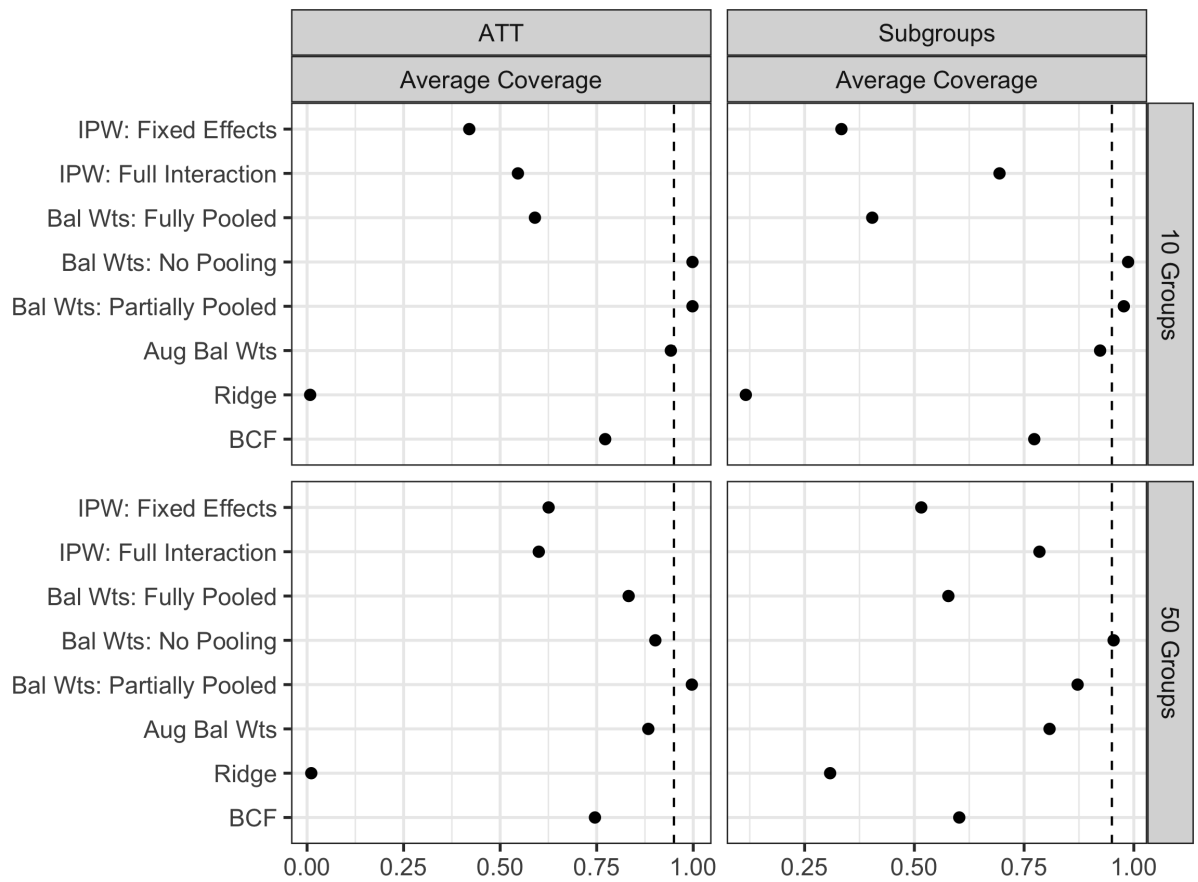


Figure D.3: Coverage for approximate balancing weights, traditional IPW with logistic regression, and outcome modelling for estimating subgroup treatment effects.

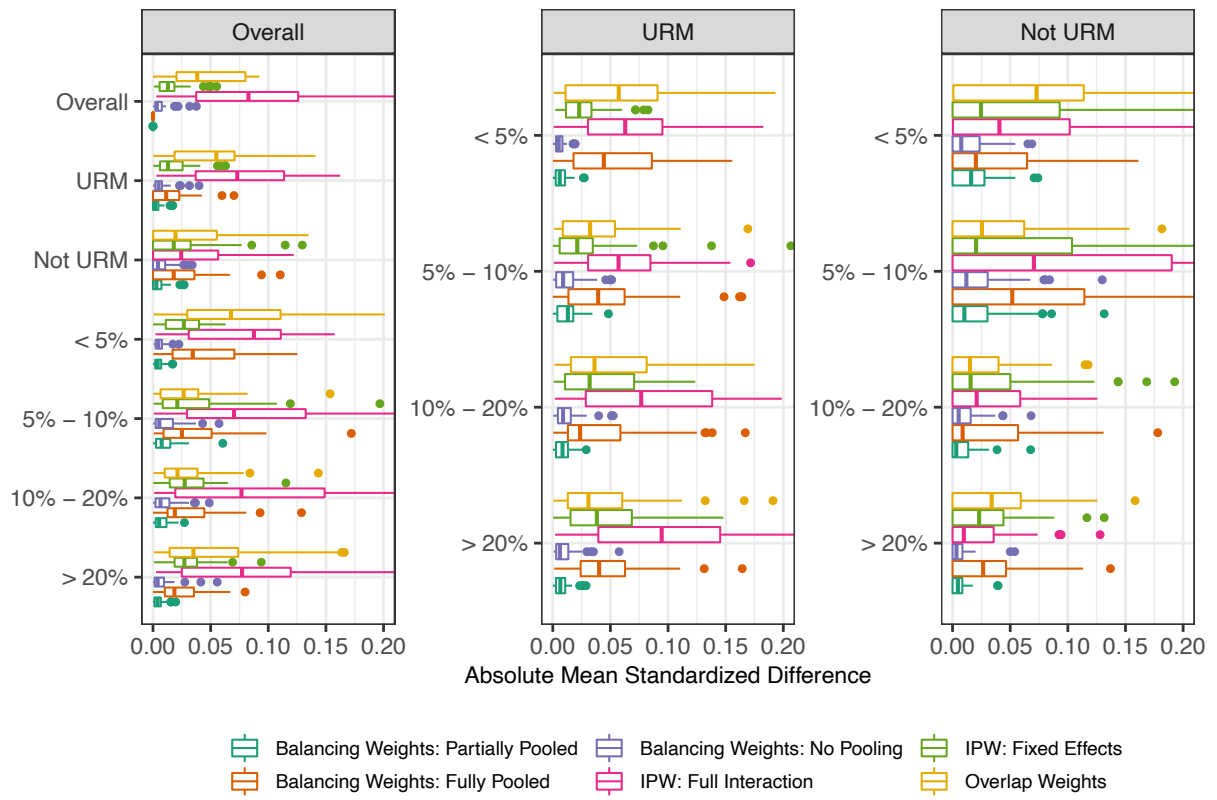
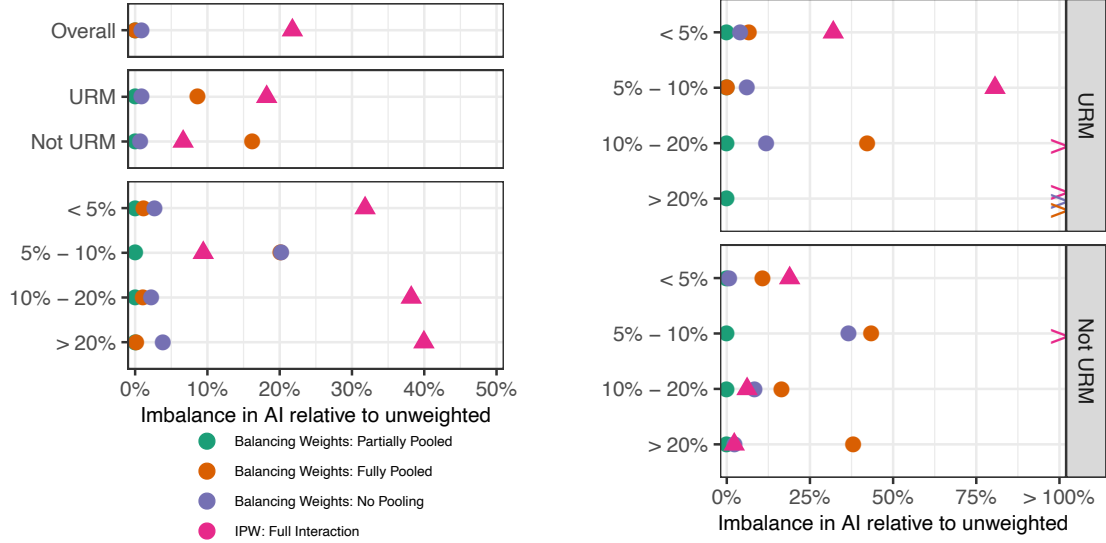


Figure D.4: Distribution of covariate balance measured by the mean standardized difference for different weighting methods.



(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.

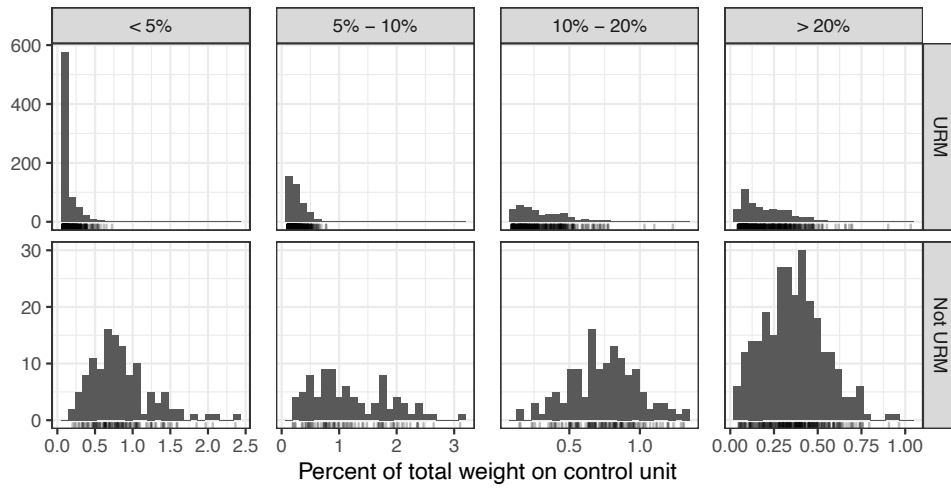


Figure D.6: Weights on control units from solving the approximate balancing weights problem (13). *Not pictured*: the 66% of control units that receive zero weight.

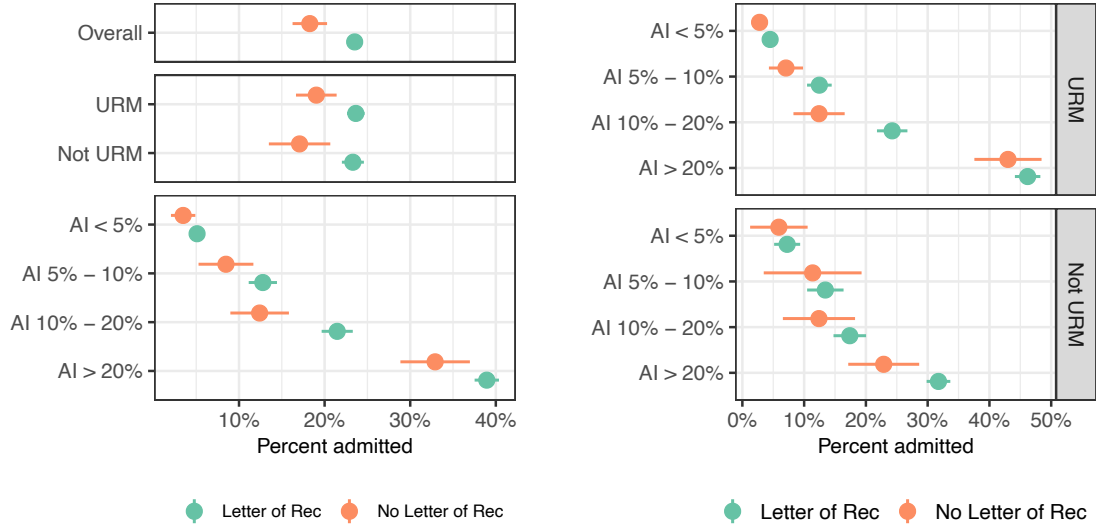
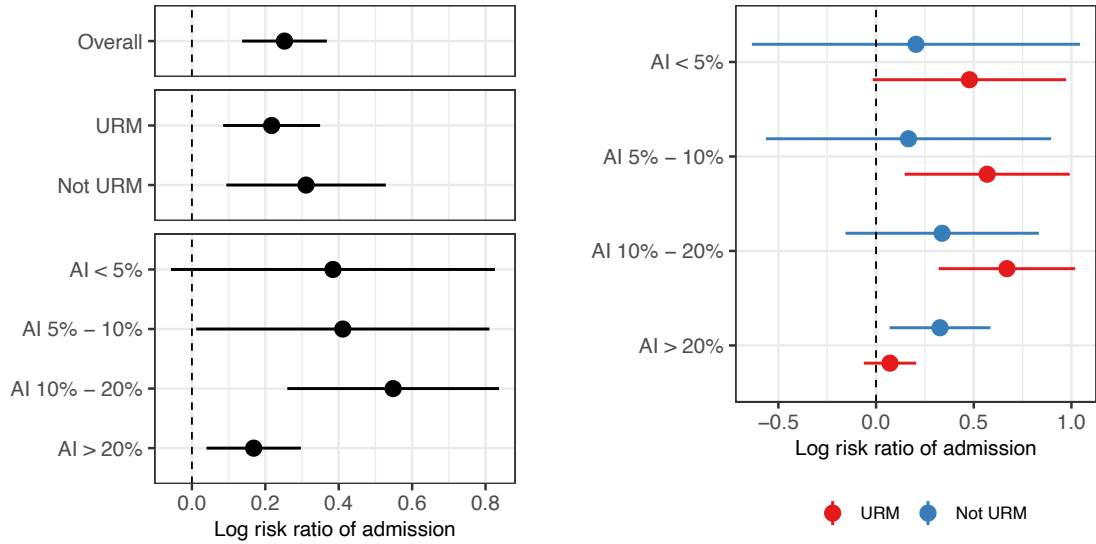


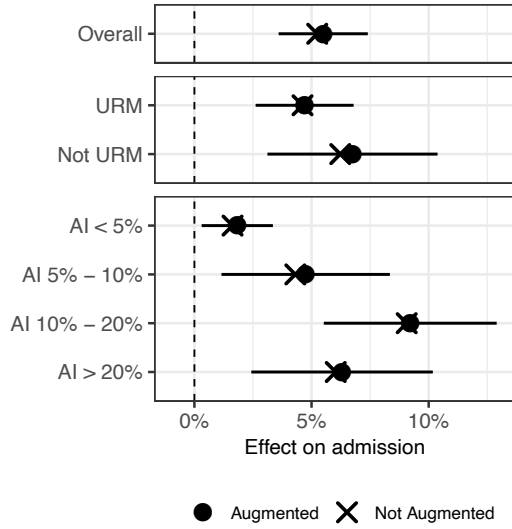
Figure D.7: Estimated treated and control means \pm two standard errors: overall, by URM status, by Admissibility Index, and by URM \times AI.



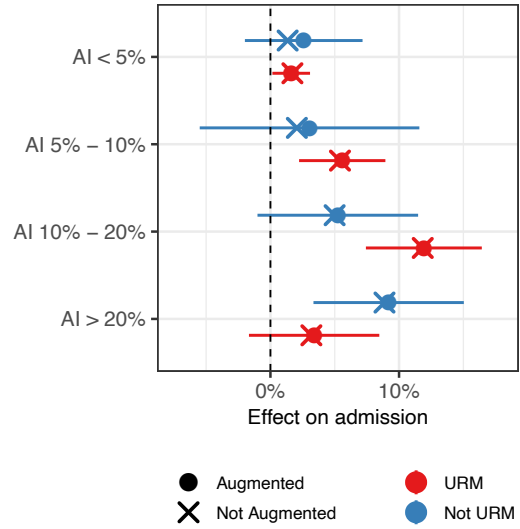
(a) Overall and by URM status and AI.

(b) By URM status interacted with AI.

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

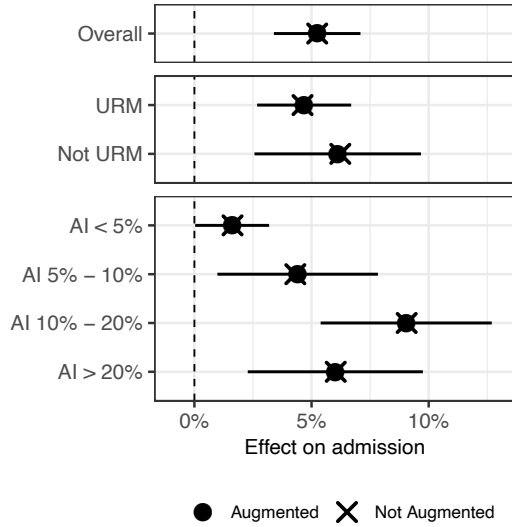


(a) Overall and by URM status and AI.

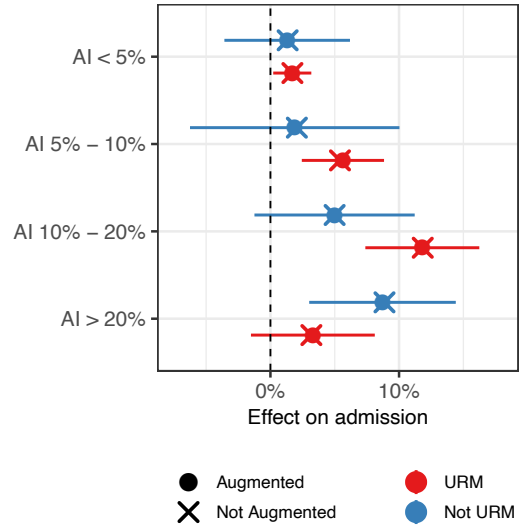


(b) By URM status interacted with AI.

Figure D.9: Estimated effect of letters of recommendation on admission rates with and without augmentation via a random forest outcome model.

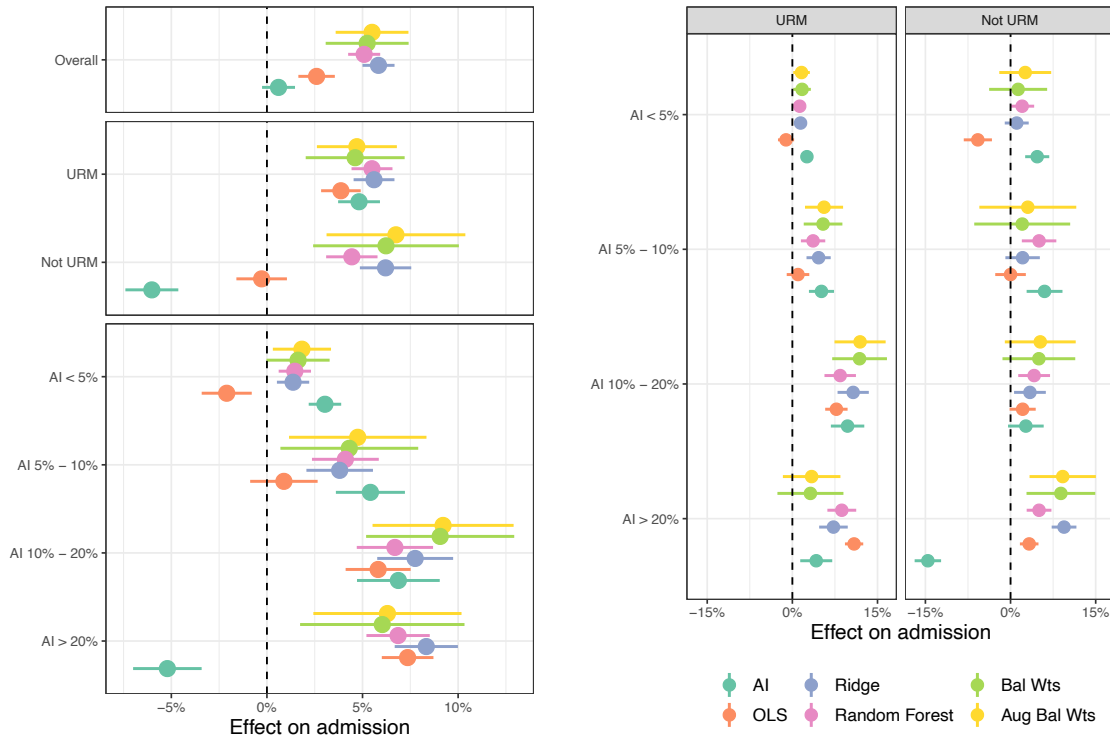


(a) Overall and by URM status and AI.



(b) By URM status interacted with AI.

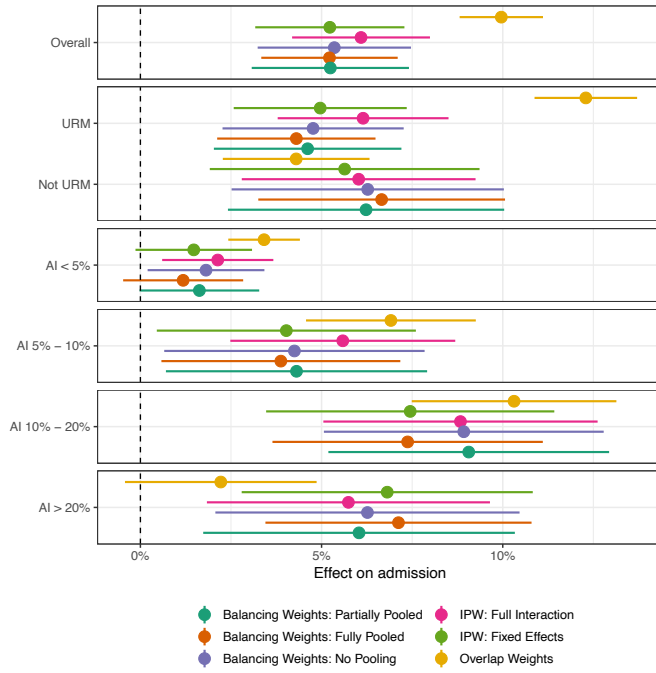
Figure D.10: Estimated effect of letters of recommendation on admission rates with and without augmentation via ridge regression with 5-fold cross validation.



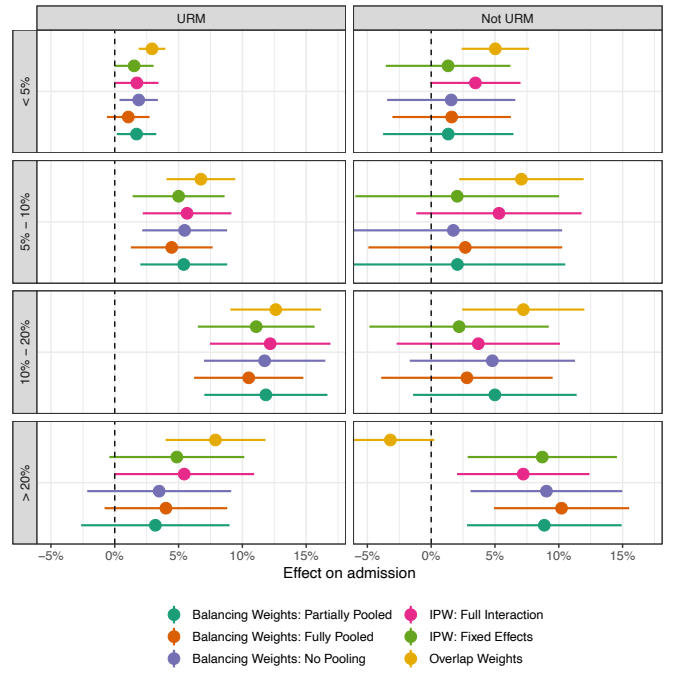
(a) Overall and by URM status and AI

(b) By URM status interacted with AI

Figure D.11: Estimated effect of letters of recommendation on admission rates *without* adjusting for selection, and comparing to the expected admission rate from the AI score and regression.

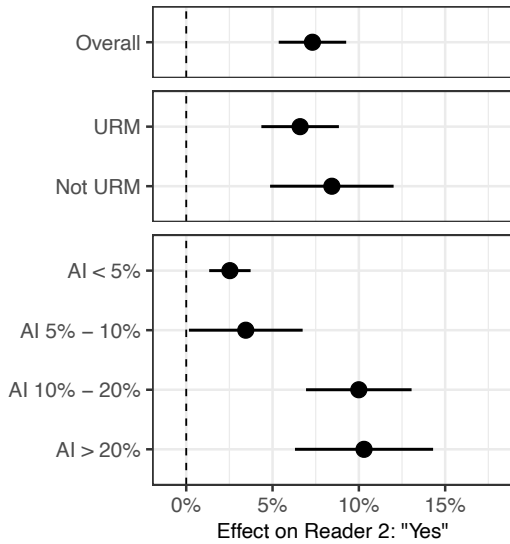


(a) Overall and by URM status and AI.

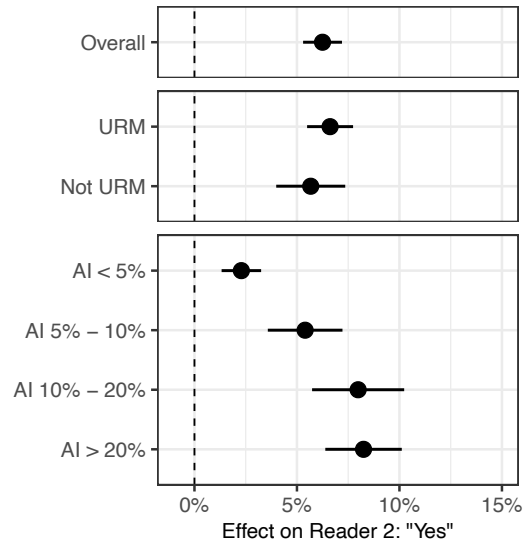


(b) By URM status interacted with AI.

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

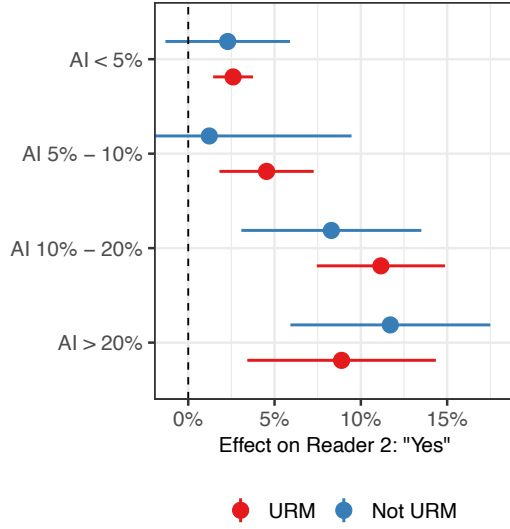


(a) Partially pooled balancing weights

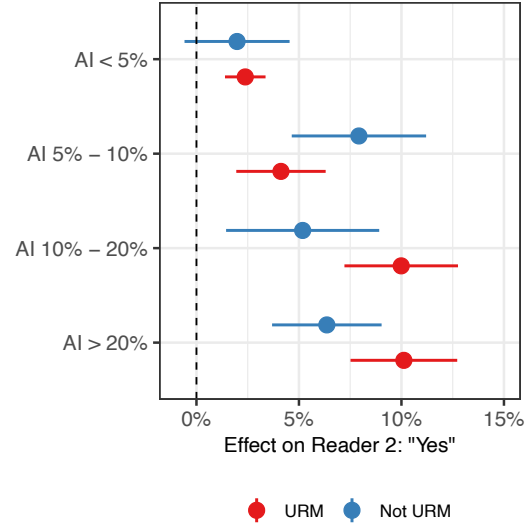


(b) Within-subject design

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.

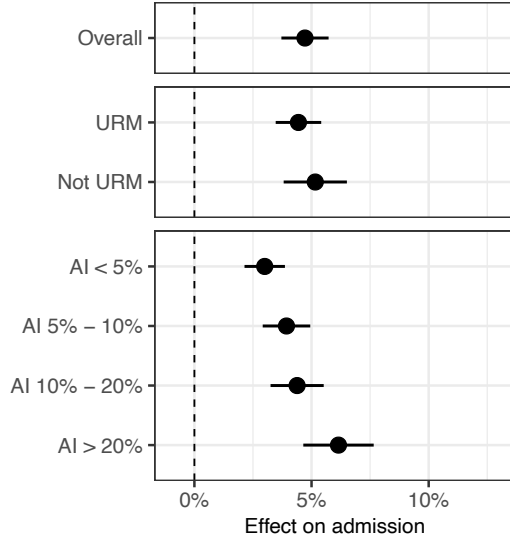


(a) Partially pooled balancing weights

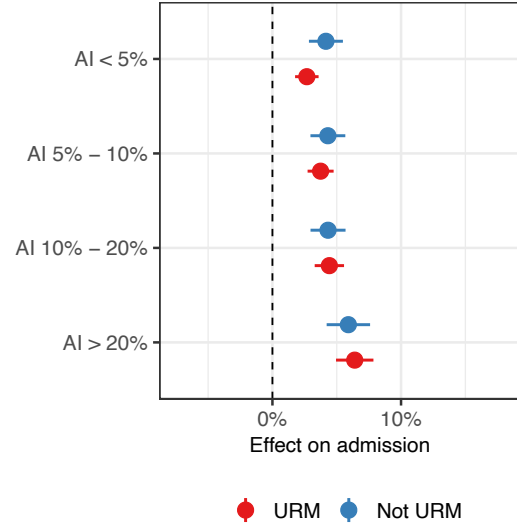


(b) Within-subject design

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.

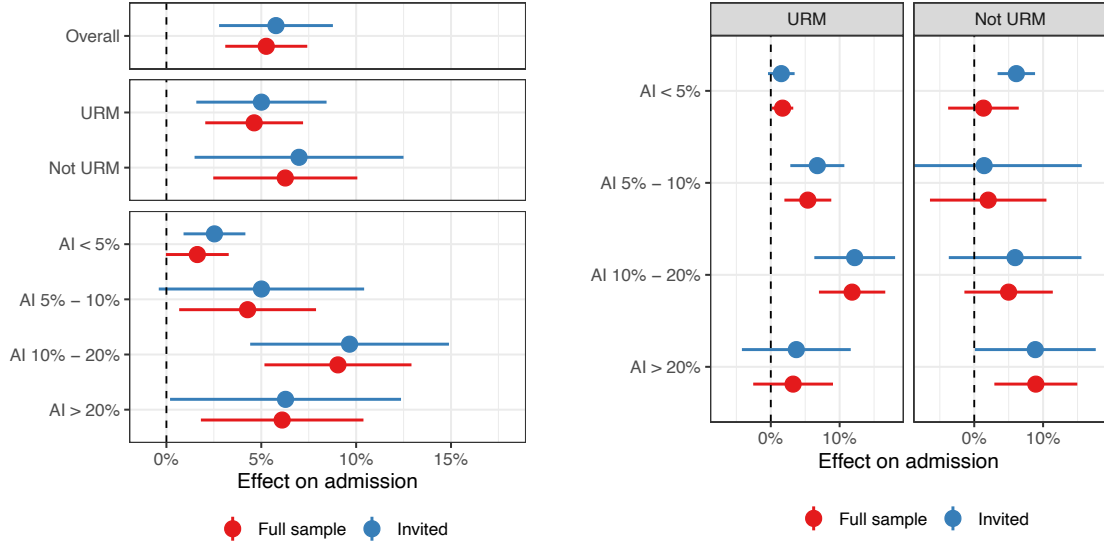


(a) Overall and by URM status and AI.



(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

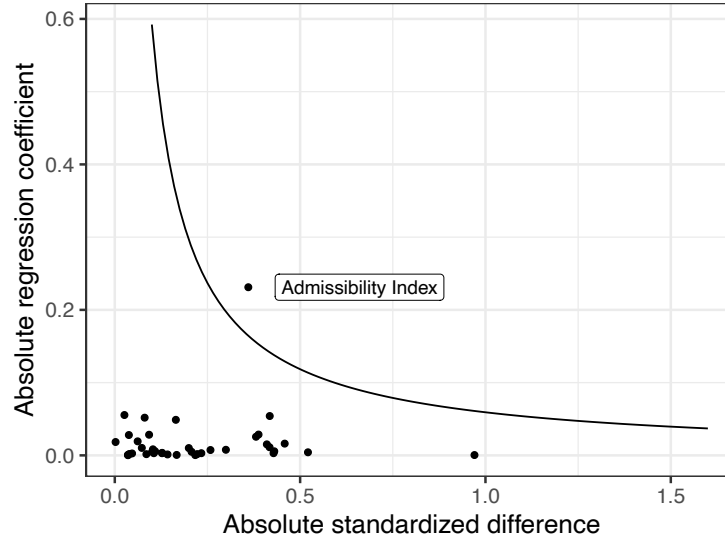


Figure D.17: Amplification of a sensitivity analysis. The line shows the magnitude of the regression coefficient and the magnitude of the imbalance in an unmeasured standardized covariate required to produce enough bias to remove the effect. Points correspond to the regression coefficients and imbalance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

E Proofs

Proof of Proposition 1. First, we will augment the primal optimization problem in Equation (13) with auxiliary covariates $\mathcal{E}_1, \dots, \mathcal{E}_J$ 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:

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

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 $\delta_1, \dots, \delta_J$, and the sum-to-one constraints induce Lagrange multipliers $\alpha_1, \dots, \alpha_J$. Then the Lagrangian is

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

The dual objective is:

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

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

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

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. \square

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.
- 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.