The Annals of Applied Statistics
2023, Vol. 17, No. 4, 2843–2864
https://doi.org/10.1214/23-AOAS1740

VARYING IMPACTS OF LETTERS OF RECOMMENDATION ON COLLEGE ADMISSIONS

BY ELI BEN-MICHAEL[1,a], AVI FELLER[2,b] AND JESSE ROTHSTEIN[3,c]

[1]*Department of Statistics and Heinz College, Carnegie Mellon University,* [a]*ebenmichael@cmu.edu*
[2]*Goldman School of Public Policy and Department of Statistics, University of California, Berkeley,* [b]*afeller@berkeley.edu*
[3]*Goldman School of Public Policy and Department of Economics, University of California, Berkeley,* [c]*rothstein@berkeley.edu*

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. This proved controversial within the university, with concerns that this change would further disadvantage applicants from disadvantaged groups. To inform this debate, 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 predefined subgroups. Understanding this variation is challenging in an observational setting 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. Applying this approach to the UC Berkeley pilot study 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. However, we find little average difference for applicants from disadvantaged groups, although this result is more mixed. In the end we conclude that soliciting letters of recommendation from a broader pool of applicants would not meaningfully change the composition of admitted undergraduates.

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 overreliance 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 nontraditional 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 letters of recommendation on subsequent admission. Our goal is to assess how impacts vary across predefined subgroups in order to inform the debate over the Berkeley policy and similar debates at other universities.

Assessing such heterogeneity is difficult in nonrandomized studies like this because variation in estimated impacts reflects both actual treatment effect variation and differences in

Received October 2022; revised February 2023.
Key words and phrases. Higher education, observational study, subgroup analysis, balancing weights.

covariate balance across groups. Existing approaches, such as balancing weights and traditional inverse propensity score weighting (IPW), face a curse of dimensionality when estimating subgroup effects: balancing all covariate-by-subgroup interactions is difficult, and fully interacted models can over-fit.

To address this, we develop a balancing weights approach tailored to estimating heterogeneous treatment effects in 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. The resulting weights control the estimation error of subgroup-specific effects, allowing us to better isolate treatment effect variation. This proposal also has a dual representation as inverse propensity weighting with a hierarchical propensity score model. Finally, we propose combining weighting with an outcome model to adjust for any remaining imbalance, analogous to bias correction for matching.

We then use this approach to assess heterogeneity in 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 by Berkeley's preferred markers of student disadvantage (such as being low income or from a low-scoring high school) and by applicant strength, estimated using data from the prior year's admissions cycle. We first 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 applicant's predicted 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 nondisadvantaged 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 nondisadvantaged 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 nondisadvantaged applicants. Since student disadvantage is correlated with applicant strength, this leads to a Simpson's Paradox-type pattern for subgroup effects (Bickel, Hammel and O'connell (1975), VanderWeele and Knol (2011)): there is a slightly larger point estimate for nondisadvantaged applicants pooled *across* applicant strength but larger point estimates for disadvantaged applicants *within* most levels of applicant strength.

We also 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 assumption of no unmeasured confounding, adapting a 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. Finally, we conduct a simple simulation exercise to project the impact of a policy requiring letters of recommendation for all applicants, finding minimal effects on the demographic composition of admitted students. 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 studies that also examine the content of the letters (Rothstein (2022)).

The paper proceeds as follows. In the next section we introduce the letter of recommendation pilot program at UC Berkeley. Section 3 introduces the problem setup and notation and discusses related work. Section 4 proposes and analyzes the approximate balancing weights approach. Section 5 presents empirical results on the effect of letters of recommendation.

Section 6 concludes with a discussion about possible extensions. The Supplementary Material includes additional analyses and theoretical discussion as well as an extensive simulation study (Ben-Michael, Feller and Rothstein (2023)).

2. A pilot program for letters of recommendation. There is a longstanding policy debate around the relative roles of quantitative and qualitative measures, including letters of recommendation, in selective undergraduate admissions; see, for example, Bowen and Bok (1996), Rothstein (2004), Karabel (2005), Bleemer (2022). 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, Kochevar and Ones (2014)). At the same time, letters from applicants from disadvantaged backgrounds or underresourced high schools may be less informative or prejudicial against the applicant, due, for example, to poor writing or grammar or to lower status of the letter writer (Schmader, Whitehead and Wysocki (2007)). A related concern arises in admissions essays: Alvero et al. (2021) find that essay text is strongly predictive of family income.

2.1. *Letters of recommendation at UC Berkeley.* Historically, undergraduate admissions at UC Berkeley were largely quantitative and mechanical, often determined by SAT scores and high school GPA alone (see, e.g., Bleemer (2022)). This began to change in the mid-2000s when Berkeley adopted a "holistic review" in which two separate reviewers read and scored each application (Hout (2005)). This shifted further in the mid-2010s with a push to consider LORs in admission, with the explicit goal of identifying students who were strong enough to admit but were unlikely to be admitted without the additional context that LORs provide (UC Berkeley (2017)). To explore this potential, UC Berkeley solicited LORs from a small number of applicants in the Fall 2015 admissions cycle, expanding to a larger number in the Fall 2016 admissions cycle.

The pilot LOR policy led to significant debate within the university. 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 was concerned 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)). Ultimately, the faculty senate limited the use of LORs following the pilot—though before any results were available.

Our goal is to conduct an impact analysis for the effect of LORs on undergraduate admissions from the UC Berkeley pilot study, especially regarding how impacts vary across key student subgroups. To the best of our knowledge, this type of impact estimate is the first of its kind: the policy change at UC Berkeley is unique. In a companion paper, Rothstein (2022) uses natural language processing methods to understand the role of letter *content* in admissions. Unlike in our paper, Rothstein (2022) restricts his analysis to a subset of 10,000 applications that received additional review after admissions decisions were made; we discuss this alternative approach in Appendix A. Finally, in an internal UC Berkeley report, Rothstein (2017) discusses key implementation details and explores alternative research designs.

2.2. *UC Berkeley pilot study.* Our analysis focuses on applicants for undergraduate admissions to UC Berkeley in the 2016 admissions cycle. Specifically, we restrict the pool of applicants to nonathlete California residents who applied for freshman admission to either

the College of Letters and Science or the College of Engineering. 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 5 and Appendix A.

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

2.3. *Selection into treatment.* Selection into submitting letters was a two-step process: A subset of 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."[2] 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.[3] When submitted, letters were made available to the second reader for possible consideration, with the 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 letters, and 11,143 (76% of those invited) eventually submitted them. 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 3453 applicants who were invited but did not submit LORs as part of the possible comparison group; we consider alternative definitions in Appendix A.2.

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 standard deviation) for applicants who submitted LORs vs. those who did not.[4] We see that there are large imbalances in observable applicant characteristics, most notably average school income, GPA, the number of honors

[1]26% are low-income, 37% from low-performing schools; 23% are first generation, and 28% are from underrepresented racial or ethnic groups; there is substantial overlap among these categories.

[2]Application 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.

[3]The 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 (vs. 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.

[4]The 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

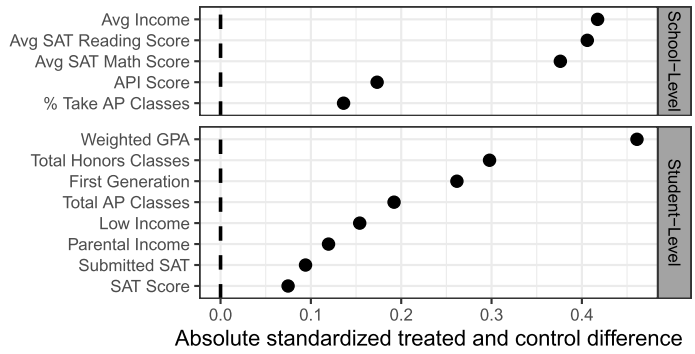


FIG. 1. *Absolute difference in means, standardized by the pooled standard deviation, between applicants submitting and not submitting LORs for several key covariates. By design, applicants submitting LORs disproportionately have a "Possible" score from the first reader* (70% *of treated applicants vs.*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

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 vs. 53% of those who did not submit. Our statistical goal is to adjust for these differences in observable characteristics between applicants who do and do not submit LORs. However, differences in *unobservable* characteristics across applicants, for example, 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.

2.4. *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.[5] Overall, we view the AI as a useful, albeit simple, a priori measure of applicant strength and seek to adjust for it in our main analysis.[6]

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 important 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

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.

[5]These 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. About 15% of students in the data used to train the AI model submitted LORs in an earlier iteration of the UC Berkeley pilot; the AI does not include any information on whether students submitted LORs.

[6]Unfortunately, the AI cannot be used to impute counterfactuals on its own. The model used to generate the AI did not include interactions and, in particular, did not account for differences in admissions outcomes across applicants to different colleges within the university. As we show in Appendix Figure D.1, while the AI has decent calibration for the overall sample, it is miscalibrated for engineering applicants and high admissibility letters and science applicants, both URM and non-URM.

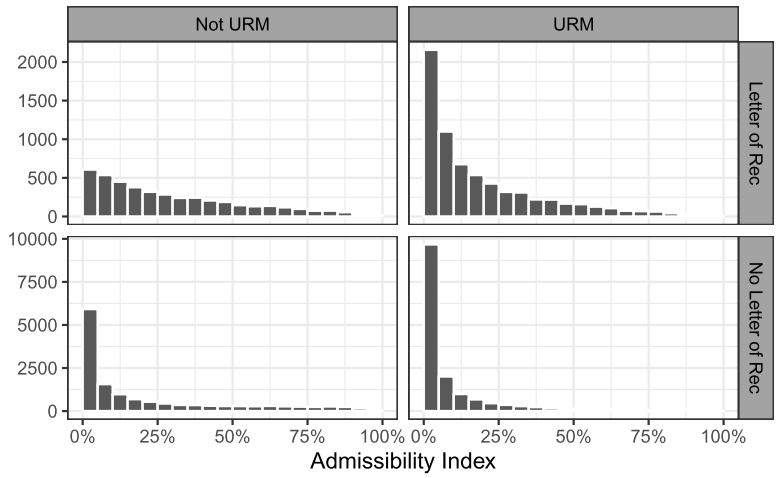


FIG. 2. *Distribution of the "admissibility index"—an a priori estimate of applicant strength—for the* 2016 *UC Berkeley application cohort, separated into URM and non-URM, and those that submitted letters vs. those that did not.*

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 (nonequally-sized) strata of the AI: $<5\%$, 5%–10%, 10%–20%, and $>20\%$. Interacting with URM status, this leads to eight nonoverlapping 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 5, we will further divide each of these subgroups by first-reader score and college to ensure exact balance on these important covariates.

3. Treatment effect variation in observational studies.

3.1. *Setup and estimands.* We now describe the letter of recommendation study as an observational study where for each applicant $i = 1, \ldots, n$, we observe applicant and school level-covariates $X_i \in \mathcal{X}$, a group indicator $G_i \in \{1, \ldots, K\}$ denoting a predefined 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 assumption rules out interference between applicants; that is, we assume that the availability of LORs for one student does not affect any other student's admission probability. While this assumption cannot strictly hold in our setting—there are more applicants than admissions slots—we view this as a reasonable working assumption. UC Berkeley admitted nearly 19,000 students for Fall 2017 in a relatively mechanistic way with minimal coordination across application readers. We discuss this further in Appendix B and also formalize the relevant no interference assumption.

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 letters,

$$\tau = \mathbb{E}\big[Y(1) - Y(0) \mid W = 1\big] = \mu_1 - \mu_0,$$

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),

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

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,

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

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

$$e(X, G) < 1. \tag{3}$$

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 that we describe in Section 2.3, 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 5.1 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,

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

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, $\frac{e(\cdot,\cdot)}{1-e(\cdot,\cdot)}$. In our setting the large number of groups and relatively small number of observations per group means that we cannot estimate any of these quantities precisely at the group level. Thus, our methodological approach will focus on ways of borrowing information across groups to improve precision while maintaining validity.

3.2. *Related work.* There is an extensive literature on estimating varying treatment effects in observational studies; see Anoke, Normand and Zigler (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 predefined, 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,

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

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 (2021), Hahn, Murray and Carvalho (2020) for discussions of approaches that instead focus on outcome modeling.

3.2.1. *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), Griffin et al. (2022)). Traditionally, this proceeds in two steps: first, estimate the propensity score $\hat{e}(x, g)$, for example, via logistic regression; second, estimate μ_{0g}, as in equation (5), with weights $\hat{\gamma}_i = \frac{\hat{e}(X_i,G_i)}{1-\hat{e}(X_i,G_i)}$,

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

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 $\phi(X_i) \in \mathbb{R}^p$ using some transformation function $\phi : \mathbb{R}^d \to \mathbb{R}^p$,

$$\operatorname{logit}\big(e(x, g)\big) = \alpha_g + \beta_g \cdot \phi(x). \tag{7}$$

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, Nguyen and Stuart (2021)), hierarchical Bayesian modeling (Li, Zaslavsky and Landrum (2013)), and generalized boosted models (McCaffrey, Ridgeway and Morral (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, Morgan and Zaslavsky (2018) 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 5.1, 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.

3.2.2. *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 $\phi(X_i) \in \mathbb{R}^p$,

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

for weights γ, typically constrained to the simplex as we have written, allowable covariate imbalance δ, and a transformation function $\phi : \mathbb{R}^d \to \mathbb{R}^p$ giving transformations of the covariates $\phi_j(\cdot)$, $j = 1, \ldots, p$. These methods have a long history in calibrated survey weighting (see, e.g., Deville, Särndal and Sautory (1993)) and have recently been extensively studied in the observational study context (e.g., Zubizarreta (2015), Athey, Imbens and Wager (2018), Hirshberg, Maleki and Zubizarreta (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 typically 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 5.1, 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.

4. Balancing weights for treatment effect variation. We now describe a specialization of balancing weights that minimizes the bias for estimating both the overall treatment effect and the subgroup-specific treatment effects. 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.

4.1. *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. (2021) 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

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

Since the weights $\hat{\gamma}$ are *design-based* (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 $\frac{1}{n_{1g}^2} \sum_{G_i=g} (1 - W_i) \hat{\gamma}_i^2$.[7] 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,[8] 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; that is, $\mathcal{M} = \{m_0(x, g) = \eta_g \cdot \phi(x) \mid \|\eta_g\|_2 \le C\}$. We can then bound the bias by the level of *local imbalance* within the subgroup via the Cauchy–Schwarz inequality,

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

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.

4.2. *Overall treatment effect.* We estimate aggregated effects by taking weighted averages of the subgroup-specific estimates; for example, we estimate μ_0 as $\hat{\mu}_0 = \sum_{g=1}^K \frac{n_{1g}}{n_1} \hat{\mu}_{0g} = \frac{1}{n_1} \sum_{W_i=0} \hat{\gamma}_i Y_i$. 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 local imbalance plays a part. The error is

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

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

[8] See 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, Maleki and Zubizarreta (2019) for a general nonparametric treatment.

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,

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

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, that is, $\eta_g = \bar{\eta}$ 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, for example, $\|\eta_g - \bar{\eta}\|_2$ 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," that is, $\|\eta_g - \bar{\eta}\|_2 < \|\bar{\eta}\|_2$ (see Feller and Gelman (2015)). As a result, our goal is to find weights that prioritize global balance while still achieving good local balance.

4.3. *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:

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

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 transformed covariates globally over the entire sample. Equivalently, this finds weights that achieve exact balance marginally on the transformed 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.[9] Taken together, this ensures that we are minimizing the overall bias as

[9]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

well as the 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 targeting local balance alone without the global balance constraint can fail to achieve good global balance. On the other hand, including the exact global balance constraint guarantees global balance. As we discuss in Section 5.1, in our study this global balance constraint improves global balance at a small cost to local balance.

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. By 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 5 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.

In Appendix B 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. 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.

The optimization problem in equation (13) also includes 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 hyperparameter 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; 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—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 5.1.

Next, equation (13) incorporates two additional constraints on the weights. We include a fine balance constraint (Rosenbaum, Ross and Silber (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 nonnegative, which stops the estimates from extrapolating outside of the support of the control units.[10] 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 $(X_1, G_1, W_1), \ldots, (X_n, G_n, W_n)$ or, equivalently, conditioned on the weights. The conditional variance is

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

in practice, due to the proliferation of hyperparameters. Instead, we can expand the set of transformed covariates $\phi(x)$ to include additional interaction terms between covariates and levels of the hierarchy. We discuss this choice in the letters of recommendation study in Section 5.

[10]Without this constraint the optimization problem is equivalent to fitting a hierarchical ridge regression outcome model. For additional discussion, see Ben-Michael, Feller and Hartman (2021).

Using the ith residual to estimate $\mathrm{Var}(Y_i)$ yields the empirical sandwich estimator for the treatment effect

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

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, Maleki and Zubizarreta (2019) for discussion on asymptotic normality and semiparametric efficiency for estimators of this form.

4.4. *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 $\phi(X)$,[11]

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

where $[x]_+ = \max\{0, x\}$, and the coefficients β_g are *partially pooled* toward 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 g,

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

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*

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

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

$$\hat{\gamma}_i = \big[\hat{\alpha}_{G_i} + \hat{\beta}_{G_i} \cdot \phi(X_i)\big]_+. \tag{19}$$

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),

[11] The truncation arises from constraining weights to be nonnegative, 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. (2021) for a review of the different choices.

which corresponds to a fully interacted propensity score model where the coefficients on observed 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 toward 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, Zaslavsky and Landrum (2013)), as in Section 3.2, 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 toward zero without any pooling. In contrast, ignoring the local balance measure by setting $\lambda_g \to \infty$ 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 toward the global model; 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.

4.5. *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, Imbens and Wager (2018), Hirshberg and Wager (2021), Ben-Michael, Feller and Rothstein (2021)). Analogous to bias correction for matching (Rubin (1973)) or model-assisted estimation in survey sampling (Särndal, Swensson and Wretman (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 $\hat{m}_0(x, g)$, for example, with a flexible regression model. An estimate of the bias in group g is then,

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

This is the estimated bias due to imbalance in the 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

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

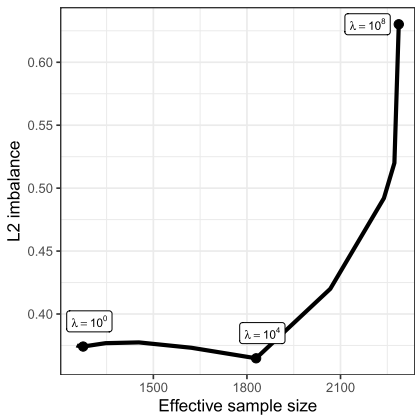
Thus, if the balancing weights fail to achieve good covariate balance in a given subgroup, the working outcome model, $\hat{m}_0(X_i, g)$, can further adjust for any differences; see Ben-Michael et al. (2021) for further discussion.

5. 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 by the interaction between URM status (two levels) and admissibility index (four levels); see Appendix Table D.1. Due to the selection mechanism described in

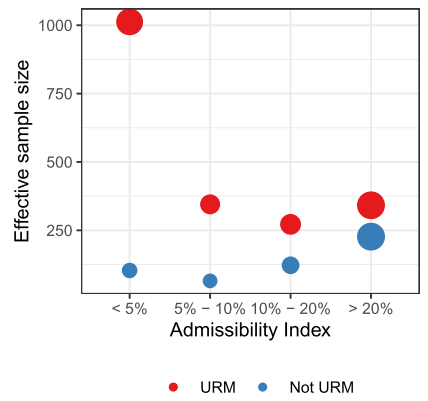
Section 2, however, it is useful to create even more fine-grained strata and then aggregate to these eight subgroups. Specifically, we define $G = 41$ fine-grained strata based on URM status, AI grouping, first reader score, and college applied to.[12] While we are not necessarily interested in treatment effect heterogeneity across all 41 strata, this allows us to exactly match on key covariates and then aggregate to obtain the 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. Because of the importance of the admissibility index, we also include a natural cubic spline for AI with knots at the sample quantiles. We next include the probability of a "Possible" score predicted by the admissions model, interacted with a binary indicator for whether it 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)).

5.1. *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 3(a) shows this



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



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

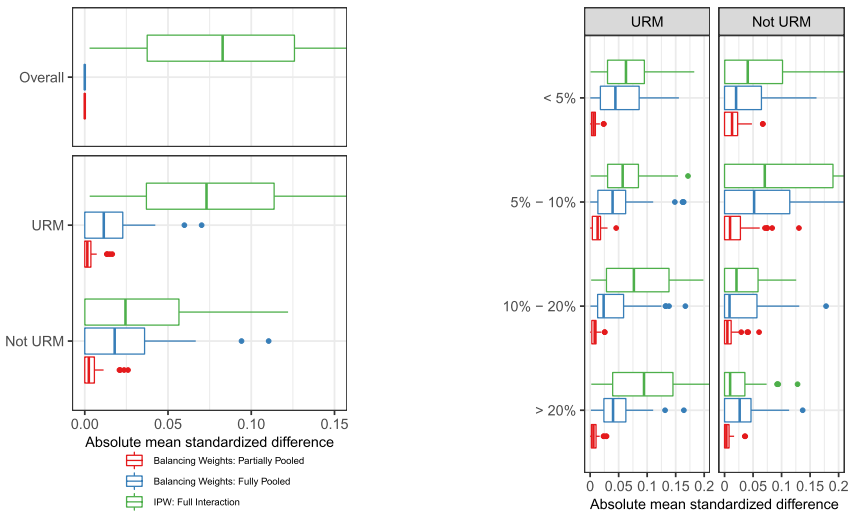
FIG. 3. (*a*) *Imbalance measured as the square root of the objective in* (13) *plotted against the effective sample size of the overall control group*. (*b*) *Effective sample size of the control group for each subgroup, with weights solving equation* (13) *with* $\lambda_g = \frac{10^4}{n_g}$.

[12]Of the 48 possible strata, we drop seven 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, p. 25: 195, median: 987, p. 75: 1038, max: 8000.

tradeoff. We plot the square root of the local balance measure in (13) against the *effective sample size* for the reweighted 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 3(b) shows the effective control group sample size for each of the primary URM and 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 4.1. We focus on three estimators: fully- and partially-pooled balancing weights, which solve equation (13) with $\lambda_g \to \infty$ 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 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.



(a) Overall and by URM status and AI. (b) By URM status interacted with AI.

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

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 9834 (34%) receive a weight larger than 0.001. This is indicative of a lack of "left-sided" overlap: 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 nontrivial number of comparison applicants that "count for" over 1% of the reweighted 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 3(b).

5.2. *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 estimates, $\hat{\mu}_{1g} - \hat{\mu}_{0g}$; 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 five percentage points (pp). We estimate a larger effect for non-URM applicants (6.2 pp) than URM applicants (4.5 pp), though there is insufficient evidence to distinguish between the two effects. We also 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 unlikely to be admitted (AI <5%), the point estimates are nearly

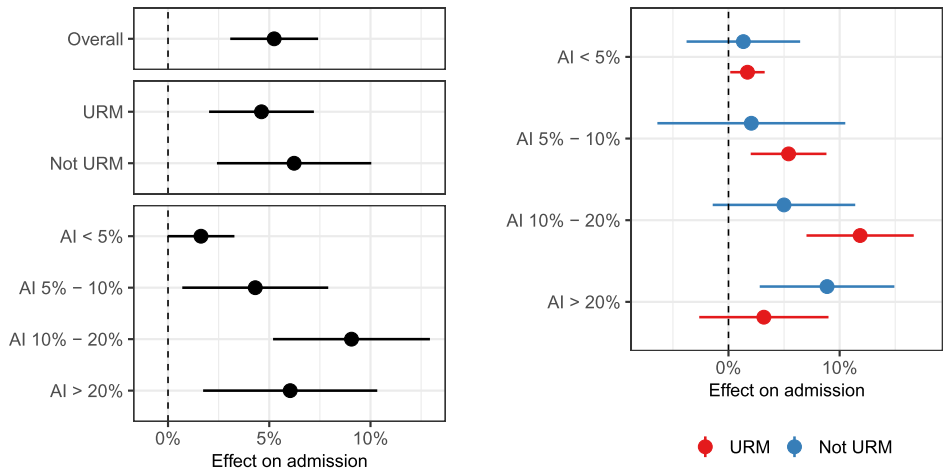


FIG. 5. *Estimated treatment effects of letters of recommendation on admission ± two standard errors*: *overall, by URM status, by Admissibility Index, and by URM × AI.*

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%), 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. We view this as a form of Simpson's Paradox (Bickel, Hammel and O'connell (1975), VanderWeele and Knol (2011)): the *prima facie* difference between the point estimates for URM and non-URM applicants is a result of the correlation between AI and URM status and masks differences in estimates within admissibility groups. 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. However, we emphasize that there is insufficient precision to make strong claims about any of these differences between effects for URM and non-URM applicants at any resolution.

We also estimate effects separately by which college the applicant applied to. Engineering admissions are more competitive than letters and science (L&S) admissions, and so the availability of additional context through LORs might have different effects. Appendix Figure D.9 shows the treatment effects overall, by URM status, and by AI for the two schools. We find that effects for applicants to L&S broadly follow the same pattern we see overall. In contrast, the effects for applicants to engineering are either substantively small or statistical indistinguishable from zero. This shows that the positive effects we see across the two schools are driven by positive effects for applicants to L&S. However, for both colleges we fail to find differential effects by URM status; although URM applicants to L&S have a higher point estimate than non-URM applicants, the opposite is true in engineering. This is primarily because there is a higher degree of uncertainty about the effects for non-URM engineering applicants.

The Appendix includes extensive robustness checks and additional analyses. We find that the overall pattern of results is consistent across a wide range of estimators and data definitions. We also conduct 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.

5.3. *Conclusions and policy implications*. First, our overall finding that submitting LORs indeed increases the probability of undergraduate admissions to UC Berkeley is largely unsurprising: readers were given explicit instructions that letters should only help applicants. That said, the point estimate of five percentage points is large relative to the expectations that we heard from university policymakers.

More relevant for the policy debate are our estimates of treatment effect variation. Our clearest results are for the 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—consistent with the goals of the admissions office—high for applicants on the margin, for whom letters provide useful context, with some evidence of a dip for the highest admissibility applicants.

At the same time, 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. Several studies have found evidence that letter writers use different language when describing different types of students with evidence, in particular, that letters written for female applicants are weaker (Trix and Psenka (2003), Madera, Hebl and Martin (2009), Schmader, Whitehead and Wysocki (2007)). However, Rothstein (2022) does not find large systematic

differences in the strength of the language used in letters written for URM and non-URM students in the UCB setting. Our point estimates of effects of letters on admissions outcomes 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. Thus, while we do not find evidence of detrimental impacts on URM applicants, we also do not find a "measurable impact on increasing diversity in undergraduate admissions," as desired by the academic senate committee (UC Berkeley (2017)).

We assess this question further in Appendix A.4 by conducting a simple policy simulation to evaluate 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)). This addresses how the overall cap on undergraduate admissions would combine with the differential effects of LORs. We find that a universal LOR requirement would raise the number of admits with strong applications but would have a negligible effect on the URM composition of admitted students.

6. Discussion. Letters of recommendation and other qualitative inputs play important roles in selective undergraduate admissions. Using a pilot study from the 2016 undergraduate admissions cycle at UC Berkeley, we find that submitting LORs increase the overall probability of admissions by five percentage points, relative to an estimated baseline of 17 percent. We find strong evidence of treatment effect variation across a baseline measure of applicant strength with larger impacts for stronger applicants. At the same time, we find no evidence of differential effects by URM status, although this is much noisier.

Taken together, our results are mixed on the UC Berkeley LOR policy debate. Those in favor of LORs can find some evidence in their favor, especially the larger impacts for students at the margins of admissions. Similarly, those opposed to the policy can point to results in support of their position, especially the possibly adverse impacts for URM applicants with the highest baseline probability of admission. In the end, however, it is unlikely that—at least based on this study—expanding LORs would meaningfully change the proportion of admitted URM students. Given the small estimated impacts, it is instead likely that other parts of the UC Berkeley admissions process, such as decisions around requiring standardized tests, will be more important factors.

Subsequent to the period that we study, the University of California system set new rules governing campus admissions that provided for the regular but limited use of LORs (University of California Board of Regents (2022)). Specifically, admissions offices may identify a small group of applicants for "Augmented Review" based on a judgment that the initial application yields an incomplete picture of their qualifications or presents extraordinary circumstances that invite further comment. Only for these applicants, no more than 15% of the overall pool, can LORs be considered. Our results suggest that these applicants may be helped by the inclusion of LORs, but the impact on the composition of the admitted pool will depend importantly on who is selected for Augmented Review. These criteria suggest that AR candidates are likely to come from middle AI ranges, suggesting that within the AR pool, LORs will be more beneficial to URM than to non-URM applicants.

Methodologically, there are several directions for future 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. (2022) 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 a key 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 hyperparameter changes. However, cross-validation approaches, such as that proposed by Wang and Zubizarreta (2020), may have better properties. This an an important avenue for future work.

Third, analogous to the issue of local balance is the issue of local overlap: if there are very many fine-grained subgroups, it may become impossible to find weights that achieve adequate local balance, even if we can achieve exact global balance. In extreme cases all individuals in a subgroup might receive either treatment or control. This occurs in the LOR pilot study, where seven strata have no applicants that submit LORs. In our analysis we can drop these strata because we target the average treatment effect on the treated. However, we could not do this if there were any strata where all applicants submitted LORs. In such settings with a lack of "right-sided" local overlap within subgroups, we may consider changing the estimand. For instance, we could trim the sample by dropping subgroups with limited overlap or by changing the target estimand to the average treatment effect on the overlapping population (Li, Morgan and Zaslavsky (2018)). We could even consider mixed estimands that shift from the treated to overlapping populations only within subgroups with poor overlap. The weighting procedure we propose can be adapted to target such estimands by incorporating weights on the treated units as well. We leave an investigation of this to future work.

Finally, while we have developed this procedure specifically for the context of the LOR pilot study, it can be applied more broadly. In particular, many observational studies exhibit a grouped structure where individual units are part of groups, for example, patients belonging to hospitals or students belonging to schools. In fact, the LOR pilot study has such a structure, with each individual applicant enrolled in one of over 1000 high schools in California. Even though measuring treatment effect heterogeneity across groups is not always the primary aim for such studies, controlling both global balance across groups and local balance within groups is key to controlling the bias for the overall treatment effect, as we have shown. We anticipate that the weighting procedure we have developed will be readily applicable to such settings; however, the exact form of the procedure may depend on study-specific factors.

Acknowledgments. The authors would like to 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.

Funding. This work was supported in part by 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.

SUPPLEMENTARY MATERIAL

Supplementary materials (DOI: 10.1214/23-AOAS1740SUPP; .pdf). Additional results, simulation details, and proofs.

REFERENCES

ALVERO, A., GIEBEL, S., GEBRE-MEDHIN, B., ANTONIO, A. L., STEVENS, M. L. and DOMINGUE, B. W. (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., NORMAND, S.-L. and ZIGLER, C. M. (2019). Approaches to treatment effect heterogeneity in the presence of confounding. *Stat. Med.* **38** 2797–2815. MR3962143 https://doi.org/10.1002/sim.8143

ATHEY, S., IMBENS, G. W. and WAGER, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 597–623. MR3849336 https://doi.org/10.1111/rssb.12268

BEN-MICHAEL, E., FELLER, A. and HARTMAN, E. (2021). Multilevel calibration weighting for survey data. Preprint. Available at arXiv:2102.09052.

BEN-MICHAEL, E., FELLER, A. and ROTHSTEIN, J. (2021). The augmented synthetic control method. *J. Amer. Statist. Assoc.* **116** 1789–1803. MR4353714 https://doi.org/10.1080/01621459.2021.1929245

BEN-MICHAEL, E., FELLER, A. and ROTHSTEIN, J. (2023). Supplement to "Varying impacts of letters of recommendation on college admissions." https://doi.org/10.1214/23-AOAS1740SUPP

BEN-MICHAEL, E., HIRSCHBERG, D., FELLER, A. and ZUBIZARRETA, J. (2021). The balancing act for causal inference. Preprint. Available at arXiv:2110.14831.

BICKEL, P. J., HAMMEL, E. A. and O'CONNELL, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science* **187** 398–404. https://doi.org/10.1126/science.187.4175.398

BLEEMER, Z. (2022). Affirmative action, mismatch, and economic mobility after California's Proposition 209. *Q. J. Econ.* **137** 115–160.

BOWEN, W. G. and BOK, D. (1996). The shape of the river: Long-term consequences of considering race in college and university admissions. In *The Shape of the River*. Princeton University Press, Princeton.

CARVALHO, C., FELLER, A., MURRAY, J., WOODY, S. and YEAGER, D. (2019). Assessing treatment effect variation in observational studies: Results from a data challenge. *Obs. Stud.* **5** 21–35.

CHALFANT, J. (2017). Letter from Jim Chalfant, Chair of the Academic Senate, to Janet Napolitano. June 20, 2017.

DEVILLE, J. C., SÄRNDAL, C. E. and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *J. Amer. Statist. Assoc.* **88** 1013–1020. https://doi.org/10.1080/01621459.1993.10476369

DONG, J., ZHANG, J. L., ZENG, S. and LI, F. (2020). Subgroup balancing propensity score. *Stat. Methods Med. Res.* **29** 659–676. MR4078241 https://doi.org/10.1177/0962280219870836

FELLER, A. and GELMAN, A. (2015). Hierarchical models for causal effects. In *Emerging Trends in the Social and Behavioral Sciences*: *An Interdisciplinary*, *Searchable*, *and Linkable Resource* 1–16.

GREEN, K. M. and STUART, E. A. (2014). Examining moderation analyses in propensity score methods: Application to depression and substance use. *J. Consult. Clin. Psychol.* **82** 773–783. https://doi.org/10.1037/a0036515

GRIFFIN, B. A., SCHULER, M. S., CEFALU, M., AYER, L., GODLEY, M., GREIFER, N., COFFMAN, D. L. and MCCAFFREY, D. (2022). A tutorial for using propensity score weighting for moderation analysis: An application to smoking disparities among LGB adults. Preprint. Available at arXiv:2204.03345.

HAHN, P. R., MURRAY, J. S. and CARVALHO, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Anal.* **15** 965–1056. MR4154846 https://doi.org/10.1214/19-BA1195

HAZLETT, C. (2020). Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects. *Statist. Sinica* **30** 1155–1189. MR4257528 https://doi.org/10.5705/ss.20

HILL, J. L. (2011). Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Statist.* **20** 217–240. MR2816546 https://doi.org/10.1198/jcgs.2010.08162

HIRSHBERG, D. A., MALEKI, A. and ZUBIZARRETA, J. (2019). Minimax linear estimation of the retargeted mean. Preprint. Available at arXiv:1901.10296.

HIRSHBERG, D. A. and WAGER, S. (2021). Augmented minimax linear estimation. *Ann. Statist.* **49** 3206–3227. MR4352528 https://doi.org/10.1214/21-aos2080

HOUT, M. (2005). Berkeley's comprehensive review method for making freshman admissions decisions: An assessment. Technical report, Univ. California, Berkeley.

KARABEL, J. (2005). *The Chosen*: *The Hidden History of Admission and Exclusion at Harvard*, *Yale*, *and Princeton*. Houghton Mifflin Harcourt, Boston.

KUNCEL, N. R., KOCHEVAR, R. J. and ONES, D. S. (2014). A meta-analysis of letters of recommendation in college and graduate admissions: Reasons for hope. *Int. J. Sel. Assess.* **22** 101–107.

KÜNZEL, S. R., SEKHON, J. S., BICKEL, P. J. and YU, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl. Acad. Sci. USA* **116** 4156–4165. https://doi.org/10.1073/pnas.1804597116

LEE, Y., NGUYEN, T. Q. and STUART, E. A. (2021). Partially pooled propensity score models for average treatment effect estimation with multilevel data. *J. Roy. Statist. Soc. Ser. A* **184** 1578–1598. MR4344649 https://doi.org/10.1111/rssa.12741

LI, F., MORGAN, K. L. and ZASLAVSKY, A. M. (2018). Balancing covariates via propensity score weighting. *J. Amer. Statist. Assoc.* **113** 390–400. MR3803473 https://doi.org/10.1080/01621459.2016.1260466

LI, F., ZASLAVSKY, A. M. and LANDRUM, M. B. (2013). Propensity score weighting with multilevel data. *Stat. Med.* **32** 3373–3387. MR3074363 https://doi.org/10.1002/sim.5786

MADERA, J. M., HEBL, M. R. and MARTIN, R. C. (2009). Gender and letters of recommendation for academia: Agentic and communal differences. *J. Appl. Psychol.* **94** 1591–1599. https://doi.org/10.1037/a0016539

MCCAFFREY, D. F., RIDGEWAY, G. and MORRAL, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol. Methods* **9** 403–425. https://doi.org/10.1037/1082-989X.9.4.403

NIE, X. and WAGER, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108** 299–319. MR4259133 https://doi.org/10.1093/biomet/asaa076

ROSENBAUM, P. R., ROSS, R. N. and SILBER, J. H. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *J. Amer. Statist. Assoc.* **102** 75–83. MR2345534 https://doi.org/10.1198/016214506000001059

ROTHSTEIN, J. M. (2004). College performance predictions and the SAT. *J. Econometrics* **121** 297–317. MR2059876 https://doi.org/10.1016/j.jeconom.2003.10.003

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

ROTHSTEIN, J. (2022). Qualitative information in undergraduate admissions: A pilot study of letters of recommendation. *Econ. Educ. Rev.* **89** 102285.

RUBIN, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* **29** 185–203.

RUBIN, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* **2** 808–804. MR2516795 https://doi.org/10.1214/08-AOAS187

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (2003). *Model Assisted Survey Sampling. Springer Series in Statistics*. Springer, New York. MR1140409 https://doi.org/10.1007/978-1-4612-4378-6

SCHMADER, T., WHITEHEAD, J. and WYSOCKI, V. H. (2007). A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex Roles* **57** 509–514. https://doi.org/10.1007/s11199-007-9291-4

SORIANO, D., BEN-MICHAEL, E., BICKEL, P., FELLER, A. and PIMENTEL, S. (2020). Interpretable sensitivity analysis for balancing weights. Technical report.

TRIX, F. and PSENKA, C. (2003). Exploring the color of glass: Letters of recommendation for female and male medical faculty. *Discourse Soc.* **14** 191–220.

UC BERKELEY (2017). Notice of Meeting: May 11, 2017. University Committee on Affirmative Action, Diversity, and Equity.

UNIVERSITY OF CALIFORNIA BOARD OF REGENTS (2022). Regents Policy 2110: Policy on Augmented Review in Undergraduate Admissions.

VANDERWEELE, T. J. and KNOL, M. J. (2011). Interpretation of subgroup analyses in randomized trials: Heterogeneity versus secondary interventions. *Ann. Intern. Med.* **154** 680–683. https://doi.org/10.7326/0003-4819-154-10-201105170-00008

WANG, Y. and ZUBIZARRETA, J. R. (2020). Minimal dispersion approximately balancing weights: Asymptotic properties and practical considerations. *Biometrika* **107** 93–105. MR4064142 https://doi.org/10.1093/biomet/asz050

WANG, J., WONG, R. K. W., YANG, S. and CHAN, K. C. G. (2022). Estimation of partially conditional average treatment effect by double kernel-covariate balancing. *Electron. J. Stat.* **16** 4332–4378. MR4474576 https://doi.org/10.1214/22-ejs2000

YANG, S., LORENZI, E., PAPADOGEORGOU, G., WOJDYLA, D. M., LI, F. and THOMAS, L. E. (2021). Propensity score weighting for causal subgroup analysis. *Stat. Med.* **40** 4294–4309. MR4300087 https://doi.org/10.1002/sim.9029

ZUBIZARRETA, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *J. Amer. Statist. Assoc.* **110** 910–922. MR3420672 https://doi.org/10.1080/01621459.2015.1023805

ZUBIZARRETA, J. R. and KEELE, L. (2017). Optimal multilevel matching in clustered observational studies: A case study of the effectiveness of private schools under a large-scale voucher system. *J. Amer. Statist. Assoc.* **112** 547–560. MR3671751 https://doi.org/10.1080/01621459.2016.1240683