Inference on Winners*

Isaiah Andrews† Toru Kitagawa‡ Adam McCloskey§

June 26, 2023

Abstract

Policymakers, firms, and researchers often choose among multiple options based on estimates. Sampling error in the estimates used to guide choice leads to a winner's curse, since we are more likely to select a given option precisely when we overestimate its effectiveness. This winner's curse biases our estimates for selected options upwards and can invalidate conventional confidence intervals. This paper develops estimators and confidence intervals that eliminate this winner's curse. We illustrate our results by studying selection of job training programs based on estimated earnings effects and selection of neighborhoods based on estimated economic opportunity. We find that our winner's curse corrections can make an economically significant difference to conclusions, but still allow informative inference.

Keywords: Winner's Curse, Selective Inference

JEL Codes: C12, C13

*We thank Tim Armstrong, Stéphane Bonhomme, Raj Chetty, Gregory Cox, Áureo de Paula, Nathaniel Hendren, Larry Katz, Patrick Kline, Hannes Leeb, Anna Mikusheva, Magne Mogstad, José Luis Montiel Olea, Mikkel Plagborg-Møller, Jack Porter, Adam Rosen, Frank Schoerfheide, Jesse Shapiro, anonymous referees, and participants at numerous seminars and conferences for helpful comments. We also thank Raj Chetty and Nathaniel Hendren for extremely generous assistance on the application using data from Chetty et al. (2020), and thank Jeff Rowley, Peter Ruhm, Bas Sanders, and Nicolaj Thor for outstanding research assistance. Andrews gratefully acknowledges financial support from the NSF under grant number 1654234. Kitagawa gratefully acknowledges financial support from the ESRC through the ESRC Centre for Microdata Methods and Practice (CeMMAP) (grant number RES-589-28-0001) and the European Research Council (Starting grant No. 715940). Initial version posted May 10, 2018.

†Department of Economics, Harvard University, iandrews@fas.harvard.edu

‡Department of Economics, Brown University, toru_kitagawa@brown.edu

§Department of Economics, University of Colorado, adam.mccloskey@colorado.edu

1 Introduction

Policymakers, researchers, and firms frequently select among multiple options (e.g. treatments, policies, or strategies) based on their estimated effects, picking the option that appears "best" according to some criterion. When the estimates used to guide our choices are uncertain, data-driven selection gives rise to a winner's curse and the selected option will systematically underperform on average relative to our initial estimate. This winner's curse arises because we are more likely to select a given option precisely when we overestimate its effectiveness. Hence, we encounter this winner's curse even in settings where the estimates used to guide our choice are unbiased, for example coming from a randomized trial. Problems related to the winner's curse have previously been discussed in a range of contexts including genome-wide association studies (e.g. Zhong and Prentice, 2009; Xu, Craiu, and Sun, 2011; Ferguson et al., 2013) and online A/B tests (Lee and Shen, 2018).

As an example, consider the JOBSTART demonstration, which was a randomized trial evaluating the effectiveness of different job-training and job-placement programs for high-school dropouts across 13 different sites in the US. The experiment found limited and statistically insignificant earnings effects at 12 of the 13 sites, but found that the remaining program generated large and statistically significant effects (see Cave et al., 1993, for a complete description). A subsequent replication study attempted to mimic the successful program at a further 12 sites, but found disappointing results. Miller et al. (2005) describe this replication study in detail and discuss multiple factors that may have led to the disappointing outcome including implementation flaws, differences in the demographics of participants across sites, and changing labor market conditions. Note, however, that the apparently successful JOBSTART site was selected for replication based on noisy estimates. Do we need to appeal to implementation issues and other challenges to explain the disappointing outcomes in the replication, or should we have expected as much based purely on the winner's curse?

In this paper we answer this and other questions by developing estimators and confidence intervals that correct for the winner's curse. Specifically, we develop estimators with controlled median bias (e.g. which are equally likely to over- and under-estimate the effectiveness of the selected option) and confidence intervals with guaranteed coverage (e.g. which cover the true effectiveness of the selected option with probability at least 95%). In developing these corrections we consider two different notions of "correct" inference: conditional inference that holds fixed the option selected (e.g. conditioning on the identity of the best-performing site in the JOBSTART experiment), and unconditional inference

that considers performance on average across options selected.

Our analysis of conditional inference builds on the rapidly growing literature on selective inference (e.g. Fithian, Sun, and Taylor, 2017,Tian and Taylor, 2018), which derives optimal conditional confidence intervals in a range of settings, as well as classical results from the statistics literature (Pfanzagl, 1979, 1994). Similarly, our analysis for the unconditional case is related to the large literature on post-selection inference (e.g. Romano and Wolf, 2005; Berk et al., 2013). For the unconditional case we recommend a new hybrid approach, which combines conditional and unconditional methods and which we find performs quite well in simulations. Conditional inference gives stronger statistical guarantees but tends to produce noisier point estimates and wider confidence intervals relative to the hybrid approach. We consequently recommend hybrid inference as the default approach except in cases where there is a specific need for conditional guarantees.

For simplicity, we focus on the case where our initial estimates, e.g. the site-specific estimates in the JOBSTART experiment, are normally distributed with known variance. While exact normality rarely holds in practice, standard approaches to inference, e.g. based on t-statistics, rely on an assumption of approximate normality. Our finite-sample results for the normal model translate to approximate results for feasible versions of our procedures, based on asymptotically normal estimators and consistent variance estimates, and we show in the appendix that the resulting asymptotic approximations are uniformly valid over a large class of data generating processes. By contrast, procedures which either ignore the winner's curse or are not shown to be uniformly valid over appropriate data generating processes can yield unreliable inferences, even in large samples.

In the next section we introduce both the winner's curse we study and our corrections in a simplified setting based on the JOBSTART demonstration. Simulations calibrated to the JOBSTART estimates show that there is scope for a winner's curse in this setting, with conventional estimators overestimating the average treatment effect of the selected site about 85% of the time, and conventional 95% confidence intervals covering the true effect only about 80% of the time. Interestingly, however, when we apply our corrected inference approach to the actual JOBSTART data, we find that both the conditional and hybrid approaches yield results similar to conventional methods, and strongly suggest that the differences between the findings in Cave et al. (1993) and Miller et al. (2005) cannot be explained by the winner's curse alone. By contrast, projection inference, which is a type of unconditional inference applied elsewhere in the literature (e.g. Berk et al., 2013), yields substantially less precise conclusions.

For our second application, we consider the problem of targeting neighborhoods based on estimated economic mobility. In cooperation with the Seattle and King County public housing authorities, Bergman et al. (2023) conduct an experiment encouraging housing voucher recipients to move to high-opportunity neighborhoods, which are selected based on census-tract level estimates of economic mobility from Chetty et al. (2020). We consider an analogous exercise in the 50 largest commuting zones (CZs) in the US, selecting top tracts based on estimated economic mobility and examining conventional and corrected inference on the average mobility in selected tracts, relative to the average tract where a voucher-recipient household with children lived in 2018.

Calibrating simulations to the Chetty et al. (2020) data, we again find that conventional approaches suffer from severe bias in many CZs, while our corrected inference procedures eliminate these biases. Applying our procedures to the original data we find lower mobility, and higher uncertainty, for selected tracts than conventional approaches, but our results nonetheless strongly indicate gains from moving to selected tracts. Specifically, across the 50 CZs the average conventional estimate implies that target tracts are associated with a 12.25 percentile-point higher income in adulthood (for children growing up in households at the 25th percentile of the income distribution) relative to the average tract in which a voucher-recipient household lived in 2018, while the average hybrid estimate is 10.27, and the average conditional estimate is 8.19. The average width of conventional confidence intervals is 1.13 percentile points, while the average width of hybrid confidence intervals is 3.58, and the average width of conditional confidence intervals is 21.46, highlighting the price of conditional guarantees in this setting.

An alternative route to correct the winner's curse is sample splitting. Split-sample inference divides the data into two parts, where the "winning" option is selected using the first part of the data, and inference is based on the second part of the data. Since this approach uses separate data for selection and inference it eliminates the winner's curse, but will also result in worse selections on average. In our simulations calibrated to the JOBSTART demonstration, for instance, we find that split-sample selection of the target site (using half of the data for selection and the other half for inference) reduces the average treatment effect from the selected site by over 25%. Moreover, since only part of the data is used for inference, split-sample inference is also statistically inefficient.

A final option for correcting the winner's curse is to apply Bayesian methods. One can show that Bayesian methods eliminate the winners curse on average under the researcher's prior distribution. This result is, however, sensitive to the prior: for instance, the posterior

median under a given prior will have positive bias under some values for the true effects and negative bias under others, so if we care about performance at a particular true effect value, or average performance under a different prior than the one used to form the posterior, Bayesian approaches can yield invalid inferences even without the winner's curse. In settings with data on many parallel units (e.g. an experiment run at a large number of different sites) one response is to adopt an empirical Bayes approach and estimate the prior from the data. Empirical Bayes approaches that assume a normal prior (e.g. a normal distribution for tract-level economic mobility conditional on tract-level covariates) are widely used in applications, including by Chetty et al. (2020).

As with other Bayesian approaches, empirical Bayes based on a normal prior will not in general correct for the winner's curse when the true distribution of effects is non-normal, and we find that the normal approximation is an imperfect fit to the distribution of effects in Chetty et al. (2020).[1] Consequently, while empirical Bayes methods reduce the winner's curse in this setting they do not fully correct it, and the coverage of empirical Bayes credible sets in our simulations ranges between 1% and over 80% across different CZs, with lower coverage on average in CZs where the normal approximation is worse. One potential response is to relax the normality assumption, and Empirical Bayes has been shown to correct for the impact of certain forms of selection in situations where we either treat the prior nonparametrically (Efron, 2011) or can correct for misspecification of the prior (Armstrong, Kolesar, and Plagborg-Moller, 2022). We are unaware, however, of results showing that empirical Bayes approaches correct the winner's curse in general settings.

The problem we consider, inference on the true effect of the estimated "best" option, is distinct from and complementary to several other problems considered in the recent literature. Gu and Koenker (2023) study the problem of optimal selection, and more generally optimal ranking, from a decision-theoretic perspective, examining potential loss functions and recommending a nonparametric empirical Bayes approach. By contrast, our analysis takes the rule used to select the "winner" as given and conducts inference on the true effect of the selected option. Similarly, Mogstad et al. (2022) consider inference on the *ranking* of different units, proposing valid confidence intervals for e.g. the rank of a particular experimental site. This is again distinct from our analysis, which conducts

[1]The connection between empirical Bayes approaches using normal priors and shrinkage estimators (Efron and Morris, 1975) implies that particular forms of empirical Bayes can yield improved point estimates even when the normal prior is incorrect. However, these results do not imply that empirical Bayes yields correct inference, even in settings without a winner's curse.

inference on the effect of the option estimated to have a given rank.[2]

In the next section we illustrate the winner's curse and our corrected inference techniques in the context of the JOBSTART example. Section 3 looks beyond this example to introduce the setting for our general results, shows how our setting nests many problems of interest, motivates the question of inference after selection, and discusses the distinction between conditional and unconditional inference. Sections 4 and 5 state our conditional and unconditional inference results, respectively. Section 6 discusses the practical implementation of our procedure, and recaps the steps needed to apply our approach in practice. Finally, Section 7 presents our application to neighborhood effects based on Chetty et al. (2020) and Bergman et al. (2023). The online appendix presents proofs, supporting theoretical results, additional details and results for the empirical applications, and an additional empirical application based on Karlan and List (2007).

2 Revisiting the JOBSTART Demonstration

We begin by revisiting the results of the JOBSTART demonstration, which was a randomized controlled trial investigating the effectiveness of a combination of basic skills education, occupational training, support services, and job placement assistance for low-skilled high school dropouts. Implemented between 1985 and 1988 in the 13 sites listed in Table I, experimental participants were randomized into either a treatment group, who received access to JOBSTART services, or a control group, who did not. The experimental sites differed in their program structures, in their local labor market and recruiting demographics and, presumably, in unmeasured staff and center competencies. Full details of the demonstration are available in Cave et al. (1993).

Table I presents estimates of the average treatment effect (ATE) on cumulative earnings over the third and fourth years of the study at each of the 13 sites, alongside sample sizes, imputed standard errors, and the average cumulative earnings for the control group.[3] The overall effects of the demonstration on earnings were muted.[4] The one exception was the

[2]While in this paper we focus on inference on the "winning" or first-ranked option, Andrews et al. (2022) extends our results to cover inference on options ranked two or lower.

[3]While Cave et al. (1993) report point estimates for each of the 13 sites (see Table 5.13 of Cave et al., 1993), they do not report standard errors for these estimates, instead reporting only statistical significance at the 1%-, 5%- or 10%-levels. In personal correspondence Fred Doolittle, one of the authors of Cave et al. (1993), indicated that the standard errors and microdata from this study are no longer accessible. To conduct our analysis we thus impute standard errors for the site-specific ATE estimates based on other results reported in Cave et al. (1993). See Appendix A for the (restrictive) assumptions that justify this imputation.

[4]Cave et al. (1993) report that the overall cost of the demonstration was not repaid through increases in earnings or other quantified benefits to individuals in treated groups by the end of the follow-up period.

Center for Employment Training (CET) in San Jose, CA, where per-capita earnings for the treatment group in months 25-48 of the experiment exceeded those for the control group by more then $6,500, and this difference was significant at the 1% level.

Based on the success of the CET in the JOBSTART experiment, as well as in another multi-site randomized trial of job-training and related services called the Minority Female Single Parent Demonstration (Burghardt et al., 1992), which again found large positive effects at the CET, the CET program was promoted as a possible model for non-residential federal assistance. To investigate if the CET model could be successfully replicated elsewhere, the US Department of Labor launched the Evaluation of the Center for Employment Training Replication Sites in 1992 (Miller et al., 2005). The evaluation (which we henceforth refer to as the replication study) recruited individuals at 12 sites (different from the original 13 JOBSTART sites), over a period from 1995 to 1999. Full details of the replication study are provided in Miller et al. (2005). The results of the replication study were disappointing relative to those observed at the CET in the JOBSTART experiment: across the 12 replication sites the total earnings effect for the third and fourth years of the study period was -$1135, with an imputed standard error of $1315.[5] A t-test for equality of the effect in the replication study and the initial CET estimate yields a t-statistic of 3.86, so we strongly reject equality of the initial and replication effects at conventional significance levels.

A possible explanation for the disappointing results in the replication study, discussed extensively by Miller et al. (2005), is that not all of the sites in the replication study adhered closely to the design of the CET program. Specifically, Miller et al. (2005) review the key elements of the CET model and conclude that of the 12 replication sites only four achieved high fidelity to the CET program. Across these four high-fidelity sites the total earnings effect for the third and fourth years of the study was -$1556, with an imputed standard error of $2607. While the standard error is larger in this case, this estimate is still significantly lower than the CET estimate: a t-test for equality of the two coefficients yields a t-statistic of 2.7.

Miller et al. (2005) discuss a number of factors beyond program differences that may have led to disappointing results in the replication study, including differences in the pool

However, there were clearer effects on some other outcomes, particularly the likelihood of passing the General Educational Development (GED) examination or completing high school.

[5]Miller et al. (2005) report p-values for earnings effects in each year separately, but do not report a p-value or standard error for the combined earnings effect in years 3 and 4. The correct standard error for the combined effect depends on the correlation of the single-year estimates, which is not reported by Miller et al. (2005). To obtain a standard error, we thus use the reported estimates and p-values to infer standard errors for years 3 and 4 separately, and define our imputed standard error as the upper bound on the standard error for the sum, corresponding to the case of perfect positive correlation between the single-year estimates.

of experimental participants, stronger labor market conditions, and more availability of training opportunities besides the experimental treatment. There is another potentially important factor, however: the CET program was selected for replication in part based on its promising performance in the JOBSTART experiment. Since the JOBSTART estimates were themselves noisy, we should expect treatment effect estimates from the "best" site to be biased upwards entirely apart from any implementation differences or changes in the economic environment. It is thus natural to ask if the replication results are truly indicative of changes in ATEs (whether due to implementation challenges or other factors) or if we can explain the disappointing performance in the replication experiment purely based on the winner's curse. Going a step further, could such disappointing performance have been predicted even without running the replication study?

We next explain why selection of the "best" site based on noisy data will lead to winner's curse bias, and then explore the scope for bias using simulations calibrated to the JOBSTART data. We then introduce our methods for correcting the winner's curse and apply these methods to the JOBSTART and replication results, where we find that the winner's curse cannot explain the differences between the findings in Cave et al. (1993) and Miller et al. (2005), corroborating the conclusion of Miller et al. (2005) that the treatment effects differed between the original and replication experiments due to other factors. In particular, the independent success of the CET in the Minority Female Single Parent Demonstration suggests that specific aspects of the CET, for instance connections with employers in San Jose which were not replicated at other sites, may have played a role.

2.1 Winner's Curse in the JOBSTART Demonstration

As in most economic applications, inference in the JOBSTART demonstration was based on the assumption (justified by the central limit theorem) that point estimates were approximately normally distributed. If we index the 13 sites by $\theta \in \Theta$ and write $X(\theta)$ for the estimate at site θ, this corresponds to an assumption that $X(\theta) \approx^d N(\mu_X(\theta), \Sigma_X(\theta))$, where $\approx^d$ denotes approximate equality in distribution, $\mu_X(\theta)$ is the ATE at site θ, and $\Sigma_X(\theta)$ is the variance of the estimator at this site. Let us assume for simplicity that this normal approximation holds exactly, $X(\theta) \sim N(\mu_X(\theta), \Sigma_X(\theta))$, and that $\Sigma_X(\theta)$ is known. As we discuss in Section 6 below, our results for the finite-sample normal model translate to asymptotic results under minimal regularity conditions, so to build intuition for both winner's curse bias and our proposed solutions it suffices to consider the case where estimates are normally distributed with known variance.

The JOBSTART estimates are statistically independent across the 13 sites, so if we let X and μ_X denote the 13-dimensional vectors collecting estimates and ATEs across sites, respectively, we have $X \sim N(\mu_X, \Sigma_X)$ for Σ_X the diagonal matrix with diagonal elements $\Sigma_X(\theta)$. The CET delivered the largest estimate in the JOBSTART experiment, and was selected for replication. To model this situation formally, suppose that after observing estimates X we are interested in the effect at the site with the largest estimate, $\hat{\theta} = \text{argmax}_{\theta \in \Theta} X(\theta)$.[6] Our quantity of interest is thus $\mu_X(\hat{\theta})$, the *true* effect associated with the *estimated* best site.[7]

Inference on $\mu_X(\hat{\theta})$ raises an immediate challenge: while $X(\theta)$ unbiasedly estimates $\mu_X(\theta)$ at each site, $X(\hat{\theta})$ systematically over-estimates $\mu_X(\hat{\theta})$. To see why, suppose we select a specific site $\tilde{\theta} \in \Theta$. By the definition of $\hat{\theta}$ we select this site only when $X(\tilde{\theta}) \geq X(\theta)$ for all $\theta \neq \tilde{\theta}$. This implies that once we condition on selecting site $\tilde{\theta}$, the distribution of $X(\tilde{\theta})$ is shifted upwards and $X(\tilde{\theta})$ has positive median bias as an estimator for $\mu_X(\tilde{\theta})$:[8]

$$Pr_{\mu_X}\left\{X(\tilde{\theta}) \geq \mu_X(\tilde{\theta}) | \hat{\theta} = \tilde{\theta}\right\} > \frac{1}{2} \text{ for all } \mu_X.$$

The same holds for all $\tilde{\theta} \in \Theta$, so $X(\hat{\theta})$ is also biased upwards unconditionally:

$$Pr_{\mu_X}\left\{X(\hat{\theta}) \geq \mu_X(\hat{\theta})\right\} > \frac{1}{2} \text{ for all } \mu_X.$$

Similarly, conventional t-statistic-based confidence intervals may undercover.

Selection of the "winning" site thus implies a sharp theoretical prediction for the direction of bias. The magnitude of the bias depends on the data generating process, however, and the scope for bias is reduced when there is a clear best site in the sense that one site $\tilde{\theta}$ has $\mu_X(\tilde{\theta}) \gg \mu_X(\theta)$ for all $\theta \neq \tilde{\theta}$ relative to the size of the standard errors. In this case we will almost always select $\hat{\theta} = \tilde{\theta}$, so the effect of selection will be minimal. Since the variance of our

[6] Since the CET also had the most statistically significant estimate we could alternatively define $\hat{\theta}$ to select the largest t-statistic. Selection based on t-statistics generates qualitatively similar biases to those we discuss below and, as shown in Section 3, our corrections also apply in that case. Our analysis also abstracts from the success of the CET in the Minority Female Single Parent Demonstration, which contributed to its selection as a model for replication (Miller et al., 2005). The participants in the replication study (16-21 year old out-of-school youth) much more closely reflect those in JOBSTART (17 to 21 year-old high school dropouts) than those in the MFSPD (single mothers belonging to an ethnic minority group, with an average age of 28).

[7] This is distinct from the problem of inference on the effect of the *true* best site, $\mu_X(\theta^*)$ for $\theta^* \in \text{argmax}_{\theta \in \Theta} \mu_X(\theta)$. Inference on $\mu_X(\theta^*)$ would allow us to make statements about the effect of the "best" program in the JOBSTART demonstration, but would not in general indicate which program generated this effect. See Dawid (1994) for further discussion of this distinction and an argument in favor of inference on $\mu_X(\hat{\theta})$.

[8] It also has positive mean bias, but we focus on median bias for consistency with our later results.

ATE estimates is decreasing in the sample size, this might suggest that the winner's curse is a purely "small sample" issue: if we hold the site-specific ATEs fixed and increase the sample size, so long as there is not an exact tie for the "best" site (i.e. there is a unique value θ^* that maximizes $\mu_X(\theta)$) there will eventually be a clear winner, and winner's curse bias will be negligible. Hence, we might be tempted to conclude that the winner's curse is a non-issue so long as sample sizes are not too small or, equivalently, the standard errors are not too large.

This intuition is incomplete at best. First, for a given sample size near-ties in the site-specific ATEs yield very similar behavior to exact ties, and the fact that the winner's curse would eventually go away if we had more data is not especially consoling. Moreover, no matter how large the sample size or how small the standard errors, there exist near-ties sufficiently close that inference ignoring selection remains unreliable. Hence, what matters for inference is neither whether there are exact ties, nor the sample size or standard errors as such, but instead how close the best-performing treatments are to each other *relative* to the degree of sampling uncertainty. So long as the gaps between the site-specific treatment effects are modest relative to sampling uncertainty, as is often the case in practice, there is scope for winner's curse bias.

2.2 JOBSTART Simulations

To explore the quantitative importance of the winner's curse, we calibrate simulations based on the JOBSTART data. Specifically, we draw $X \sim N(\mu_X, \Sigma_X)$ where Σ_X is the diagonal matrix with the squared JOBSTART standard errors (i.e. the square of the fifth column in Table I) along the diagonal, and $\mu_X = s \cdot \hat{\mu}_X$ for $\hat{\mu}_X$ the JOBSTART point estimates (i.e. the fourth column in Table I) and s a scaling factor. For each data realization we select the "winning" site $\hat{\theta} = \text{argmax}_{\theta \in \Theta} X(\theta)$, and conduct inference on the site-specific ATE $\mu_X(\hat{\theta})$.

Figure I examines the performance of conventional estimators and confidence intervals in this setting. The first panel shows the coverage of the conventional point estimate $\pm$ 1.96 standard error confidence intervals, while the second plots the difference between the over-estimation probability and one half, $Pr_{\mu_X}\{X(\hat{\theta}) \geq \mu_X(\hat{\theta})\} - \frac{1}{2}$, and the third plots the median bias in dollars for the ATE on cumulative earnings in years three and four, $Med_{\mu_X}(X(\hat{\theta}) - \mu_X(\hat{\theta}))$. On the horizontal axis we vary the scaling factor s. The scaling $s = 1$ corresponds to the JOBSTART point estimate $\mu_X = \hat{\mu}_X$, while $s > 1$ increases the difference between the site-specific ATEs and $s < 1$ decreases the differences between the site-specific ATEs.

Due to estimation error the JOBSTART point estimates will tend to overstate the

differences of the site-specific ATEs, so it is not clear that the results for $s=1$ are necessarily the best reflection of the underlying data generating process in this setting. In particular, if we imagine sampling sites independently from a population of potential sites, the average variance of the site-specific estimates will correspond to the variance of the site-specific ATEs, plus the average sampling variance. To offset this effect, we compute the value of s such that the variance of $s \cdot \hat{\mu}_X(\theta)$ across $\theta \in \Theta$ matches an unbiased estimator for the variance of the site-specific ATEs. This yields a scaling s^* slightly above $s=0.5$, which we focus on in our discussion and plot as a vertical line in all figures.

The results in Figure I show that conventional inference procedures can suffer from substantial distortions, where the severity of these distortions is larger for smaller scaling factors s. In the case where the treatment effect is zero at all sites (corresponding to $s=0$), we have a more than 99.9% probability of overestimating the effect at the selected site, the point estimate has a median bias of more than \$2,750, and the conventional confidence interval has coverage below 75%. As s increases these issues grow less severe. At our preferred scaling s^*, we still have a nearly 90% chance of overestimating the true effect, while the conventional estimator is biased upwards by nearly \$2,000, and the conventional 95% confidence interval has true coverage probability below 82%. As we make s still larger these distortions further attenuate, and standard approaches appear quite reliable for $s \geq 1.5$.

2.3 Corrected Inference Procedures

Our goal in this paper is to develop corrections that eliminate the winner's curse bias. This section briefly describes our corrected inference procedures in the context of the JOBSTART example, while Sections 4 and 5 below develop them in full generality.

For $x \in \mathbb{R}^{13}$, let $x(-\theta)$ denote the vector x excluding the element corresponding to θ, and let $F_{TN}(\cdot;\mu,\tilde{\theta},x(-\tilde{\theta}))$ be the cumulative distribution function for a $N(\mu,\Sigma_X(\tilde{\theta}))$ distribution truncated to the interval $[\max_{\theta \in \Theta \setminus \{\tilde{\theta}\}} x(\theta),\infty]$.[9] One can show that $F_{TN}(x(\tilde{\theta});\mu,\tilde{\theta},x(-\tilde{\theta}))$ is strictly decreasing in μ. For $\hat{\mu}_\alpha$ the unique solution to $F_{TN}(X(\hat{\theta});\mu,\hat{\theta},X(-\hat{\theta}))=1-\alpha$ in μ, Proposition 2 below shows that

$$Pr_{\mu_X}\left\{\hat{\mu}_\alpha \geq \mu_X(\hat{\theta}) | \hat{\theta}=\tilde{\theta}\right\}=\alpha \quad \text{for all } \tilde{\theta} \in \Theta \text{ and all } \mu_X.$$

Hence, $\hat{\mu}_\alpha$ is α-quantile unbiased for the ATE at the estimated best site conditional on its location. That is, among those draws of the data where this particular site "wins," we

[9] Recall that for Φ the cumulative distribution function of a standard normal distribution, the distribution function of a $N(\nu,\sigma^2)$ distribution truncated to the interval $[a,b]$ is $\frac{\Phi((x-\nu)/\sigma)-\Phi((a-\nu)/\sigma)}{\Phi((b-\nu)/\sigma)-\Phi((a-\nu)/\sigma)}$.

over-estimate the ATE at this this site with probability exactly α.

Using this result, we can eliminate the biases discussed above. The estimator $\hat{\mu}_{1/2}$ is median unbiased and the equal-tailed confidence interval $CI_{ET} = \left[\hat{\mu}_{\alpha/2}, \hat{\mu}_{1-\alpha/2}\right]$ has conditional coverage $1-\alpha$, where we say that an interval CI has conditional coverage $1-\alpha$ if it covers the ATE at the estimated best site with at least this probability conditional on its location, regardless of the value of the ATEs across sites:

$$Pr_{\mu_X}\left\{\mu_X(\hat{\theta}) \in CI | \hat{\theta} = \tilde{\theta}\right\} \geq 1-\alpha \text{ for } \tilde{\theta} \in \Theta \text{ and all } \mu_X. \tag{1}$$

By the law of iterated expectations, CI_{ET} also has unconditional coverage $1-\alpha$:

$$Pr_{\mu_X}\left\{\mu_X(\hat{\theta}) \in CI\right\} \geq 1-\alpha \text{ for all } \mu_X. \tag{2}$$

Unconditional coverage is easier to attain, however, so relaxing the coverage requirement from (1) to (2) allows shorter confidence intervals in some cases.

Conditional and unconditional coverage requirements address different questions, and which is more appropriate depends on the problem at hand. If we only need to ensure that our confidence intervals cover the true ATE with probability at least $1-\alpha$ on average across the realizations of the estimated best site, it suffices to require unconditional coverage. If we instead care about performance only in a subset of instances, for instance only for when we estimate a particular site to be best, then conditional coverage may be more appropriate. We discuss the choice between conditional and unconditional inference methods at length in Section 3 below. Since conditional coverage is more demanding and can sometimes result in much wider confidence intervals, we recommend unconditional inference as the default approach when there is not a clear reason to require conditional coverage.

We are unaware of alternatives in the literature that ensure conditional coverage (1). For unconditional coverage (2), however, one can form an unconditional confidence interval by projecting a simultaneous confidence set for the ATEs at all sites, μ_X. In particular, let c_α denote the $1-\alpha$ quantile of $\max_\theta |\xi(\theta)| / \sqrt{\Sigma_X(\theta)}$ for $\xi \sim N(0, \Sigma_X)$. If we define CI_P as

$$CI_P = \left[X(\hat{\theta}) - c_\alpha \sqrt{\Sigma_X(\hat{\theta})}, X(\hat{\theta}) + c_\alpha \sqrt{\Sigma_X(\hat{\theta})}\right],$$

then one can show that this interval has correct unconditional coverage for the ATE at the estimated best site - see Section 5 below.

Figure II plots the median length in dollars of 95% confidence intervals CI_{ET} and CI_P,

along with the conventional confidence interval.[10] As Figure II illustrates, the median length of CI_{ET} is shorter than that of CI_P once s is sufficiently large, and eventually converges to the length of the conventional interval. When s is small, on the other hand, CI_{ET} can be substantially wider than CI_P. This reflects that in these cases the estimated ATE at the winning site is frequently close to the estimated ATE at the next-best site. For the truncated normal distribution used to compute CI_{ET}, an observation close to the lower endpoint provides evidence of a small mean but little precision about the exact value, leading to long confidence intervals. These features become still more pronounced if we consider higher quantiles of the length distribution: to illustrate, Figure III plots the 95th percentile of the distribution of length.[11] One can show (see Proposition 7 in Appendix C) that the endpoints of CI_{ET} are optimal quantile unbiased estimators. So long as we impose correct conditional coverage, there is hence little scope to improve conditional performance. If we instead focus on unconditional performance, by contrast, improved performance is possible.

To improve performance we propose a hybrid inference approach, which combines the conditional and projection approaches. Hybrid inference first computes a level $\beta<\alpha$ projection interval CI_P^β, and then considers conditional inference given the location of the winning site and that its ATE lies within the projection interval CI_P^β. For $F_{TN}^H(x(\tilde{\theta});\mu,\tilde{\theta},x(-\tilde{\theta}))$ the cumulative distribution function for a $N(\mu,\Sigma_X(\tilde{\theta}))$ distribution truncated to the interval

$$\left[\max\left\{\max_{\theta\in\Theta\setminus\{\tilde{\theta}\}} x(\theta),x(\tilde{\theta})-c_\beta\sqrt{\Sigma_X(\tilde{\theta})}\right\},x(\tilde{\theta})+c_\beta\sqrt{\Sigma_X(\tilde{\theta})}\right]$$

and evaluated at $x(\tilde{\theta})$, one can show that this function is again strictly decreasing in μ. For $\hat{\mu}_\alpha^H$ the unique solution to $F_{TN}^H(X(\hat{\theta});\mu,\hat{\theta},X(-\hat{\theta}))=1-\alpha$ in μ, Proposition 5 below shows that $\hat{\mu}_\alpha^H$ is α-quantile unbiased conditional on the event $\left\{\mu_X(\hat{\theta})\in CI_P^\beta\right\}$. Since $Pr_{\mu_X}\left\{\mu_X(\hat{\theta})\in CI_P^\beta\right\}\geq 1-\beta$ one can further show that $\hat{\mu}_\alpha^H$ is nearly α-quantile unbiased for the ATE at the estimated best site,

$$\left|Pr_{\mu_X}\left\{\hat{\mu}_\alpha^H\geq\mu_X(\hat{\theta})\right\}-\alpha\right|\leq\beta\cdot\max\{\alpha,1-\alpha\}\text{ for all }\mu_X.$$

[10] We focus on median length, rather than mean length, because the results of Kivaranovic and Leeb (2021) imply that CI_{ET} has infinite expected length.

[11] In both Figures II and III the lengths of conditional confidence intervals feature steep declines over a particular range of s values (slightly below 0.5 in Figure II, and slightly above 1 in Figure III). These steep declines reflect changes in the frequency with which sites with particularly noisy estimates are picked as we vary s.

We again form level $1-\alpha$ equal-tailed confidence intervals based on these estimates, where to account for the dependence on the projection interval we adjust the quantile considered and take $CI^H_{ET} = \left[\hat{\mu}^H_{\frac{\alpha-\beta}{2(1-\beta)}}, \hat{\mu}^H_{1-\frac{\alpha-\beta}{2(1-\beta)}}\right]$. See Section 5.2 for details on this adjustment. By construction, hybrid intervals are never longer than the level $1-\beta$ projection interval CI^β_P. For all results reported in this paper we set $\beta = \frac{\alpha}{10} = 0.5\%$.

Due to their dependence on the projection interval, hybrid intervals do not in general have correct conditional coverage (1). By relaxing the conditional coverage requirement, however, we obtain improvements in unconditional performance, as illustrated in Figure II, where we see that the hybrid confidence intervals have shorter median length than the unconditional interval CI_P for all parameter values considered.[12] The gains relative to conditional confidence intervals CI_{ET} are large for many values of true ATEs, and Figure III shows that these gains are even more pronounced for higher quantiles of the length distribution.

The improved unconditional performance of the hybrid confidence intervals is achieved by requiring only unconditional, rather than conditional, coverage. To illustrate, Figure 7 in Appendix A shows the conditional coverage of our conditional, hybrid, and projection intervals. As expected, only the conditional interval ensures correct conditional coverage.

2.4 Winner's Curse Corrections in the JOBSTART Demonstration

Having introduced our corrected inference procedures, we now return to the data from the JOBSTART demonstration, and ask how accounting for the winner's curse affects our conclusions. Table II reports point estimates and 95% confidence intervals for the ATE at the CET/San Jose using (i) the conventional approach, (ii) our conditional median unbiased estimator $\hat{\mu}_{\frac{1}{2}}$ and conditional interval CI_{ET}, (iii) the projection confidence interval CI_P and (iv) the hybrid estimator $\hat{\mu}^H_{\frac{1}{2}}$ and hybrid confidence interval CI^H_{ET}. We see that our conditional and hybrid adjustments make only a minimal difference in this case: the point estimates differ by at most \$3, while the length of the hybrid and conditional intervals is within about 4% of the length of the conventional interval. The one exception is the projection interval, which is over 47% longer than the conventional interval. Hence, correcting for the winner's curse in the JOBSTART data changes our conclusions very little, unless we focus on the projection approach, which necessarily implies longer confidence intervals.[13]

Given that adjusting for the winner's curse has little effect on our conclusions from the

[12] One can also compare the median absolute error of conventional, conditional, and hybrid point estimators. We find that the performance differences in this application are limited.

[13] Intuitively, the gap between estimates for the CET/San Jose and the other sites is sufficiently large to indicate a "clear winner," similar to our simulation results with s large.

JOBSTART experiment, one would expect it to also have little impact on our interpretation of the replication study. To explore this formally, we extend our theoretical results to derive winner's-curse-adjusted forecasts for the estimates in the replication study. Specifically, given the JOBSTART results and a standard error for the replication study, Section 4.1 and Appendix E show how to compute intervals which are guaranteed to cover the result in a follow-up study with a given probability (e.g. 95%) either conditional on $\hat{\theta}$ (for our conditional approach) or unconditionally (for our hybrid approach), under the assumption that the effects in the original and replication studies are the same. Table III reports forecast intervals based on the JOBSTART data, reporting separate intervals for the set of all 12 sites and the 4 high-fidelity sites. Comparing these forecast intervals to the point estimates from the replication study, we see that the forecast intervals include only positive values and so exclude the replication point estimates. This can be interpreted as a rejection of the hypothesis that the effects in the JOBSTART and replication studies are the same at the 5% significance level, and the third column of Table III computes the associated p-values of this test.

Hence, even after correcting for the winner's curse, we find strong evidence that the ATEs in the initial and replication studies were different, which suggests a role for other explanations such as those discussed by Miller et al. (2005). Specifically, Miller et al. (2005) argue that the high effectiveness of the CET found in JOBSTART and the Minority Female Single Parent Demonstration was due to unique features of the CET, including close connections with local employers, a clear organizational focus on employment as the goal, little upfront screening of applicants, training in occupations demanded by the local labor market, relatively intensive services concentrated over a short period of time, strong job placement efforts, and a high-wage labor market. Our empirical results suggest that these or other features of the CET program could not be adequately reproduced at the sites in the replication study.

The use of our conditional and hybrid approaches is important for the result obtained in this application. While we do not know of a method to produce a non-conservative forecast interval using the projection approach, the 95% projection confidence interval based on the Cave et al. (1993) results overlaps with the conventional 95% confidence interval for the average effect in the high-fidelity sites, but not for the full 12 sites in Miller et al. (2005). Hence, if we relied on the projection approach to correct for the winner's curse, our conclusions about the comparison between JOBSTART and the replication experiment would depend on the set of sites considered.

An important limitation of our analysis is that since Cave et al. (1993) do not report standard errors for the site-specific ATE estimates in the JOBSTART demonstration, our analy-

sis relies on standard errors imputed (under restrictive assumptions) using other information in Cave et al. (1993). To examine the sensitivity of our conclusions to these imputations, we consider proportionately scaling up the standard errors at all sites, and ask how much we would need to scale up the standard errors in order to *not* reject equality of the effects in Cave et al. (1993) and Miller et al. (2005) at the 5% level. The resulting scalings are reported in the last column of Table III, and range from 1.72 to 1.84 depending on the method and the set of replication sites considered. Focusing on the smallest of these, if we scaled up the standard errors in Cave et al. (1993) by a factor of 1.72, this would imply a standard error at the CET/San Jose of approximately $2573. This would imply, however, that the CET/San Jose estimate in Cave et al. (1993) is insignificantly different from zero at the 1% significance level, inconsistent with the reported significance levels in Cave et al. (1993). Hence, even correcting for the winner's curse and possible inaccuracy in our imputed standard errors, it seems likely that the treatment effects in Cave et al. (1993) and Miller et al. (2005) were different.

Alternative Correction: Split-Sample Approaches An alternative approach to correct for the winner's curse is sample splitting. In the context of the JOBSTART example, sample splitting would entail dividing the data at each site into two parts, where we would use the first part to select among the sites and the second to conduct inference. This ensures that selection and inference are based on independent observations and so eliminates the winner's curse.

While sample splitting avoids the winner's curse, it comes at a cost on multiple dimensions. First, and most importantly, selecting a site based on only part of the data leads to worse selections on average. To illustrate this point, Appendix A explores the performance of split-sample approaches in our JOBSTART simulations, focusing on the case where we use half the data for selection and the other half for inference. Using just half the data for selection substantially reduces the ATEs for selected sites, and we find that (at our preferred scaling s^*) selecting the target site based on half of the data means that the selected site has an ATE $514 lower, on average, relative to the case where we use the full data. This is over 26% of the ATE under full-data selection at this scaling. Second, even if we are comfortable with the poorer targeting that results from sample splitting, conventional split-sample inference is statistically inefficient, yielding wider confidence intervals and noisier point estimates than necessary (Fithian, Sun, and Taylor, 2017).

Alternative Correction: Bayesian Methods One may also correct for the winner's curse using Bayesian methods. Under a given prior distribution π for μ_X, the posterior

distribution given X, $\pi(\mu_X|X)$, fully summarizes our beliefs given the observed data. Since the identity of the "winner" is just a function of X, further conditioning on the winner does not change our posterior, $\pi(\mu_X|X,\hat{\theta})=\pi(\mu_X|X)$.

Using the law of iterated expectations one can show that if CR is a level $1-\alpha$ credible set for $\mu_X(\hat{\theta})$ (that is, a set which contains $\mu_X(\hat{\theta})$ with probability $1-\alpha$ under our posterior distribution), then $Pr_\pi\left\{\mu_X(\hat{\theta})\in CR|\hat{\theta}\right\}=1-\alpha$ so CR has conditional coverage $1-\alpha$ *under the prior*. It is critical for this calculation, however, that the probability is computed under the prior π. If the data are drawn in some other way (for instance under a fixed parameter value μ_X), Bayes credible sets do not in general have correct conditional coverage (1) or unconditional coverage (2).[14] One response to this sensitivity to the prior is to adopt an empirical Bayes perspective and try to estimate the prior (e.g. the distribution of site-specific ATEs) from the data. Such estimation seems likely to be challenging given data from only 13 sites as in the current example, but we discuss the performance of empirical Bayes approaches in another context in Section 7 below.

3 Setting and Inference Problem

We now move beyond the JOBSTART example to introduce the general class of problems we study, which covers many other settings of potential interest. We then frame and motivate the question of inference-after-selection and discuss the choice between conditional and unconditional inference.

Let Θ be a finite set of options (e.g. treatments or policies). For each option $\theta\in\Theta$ we observe a two-dimensional vector of estimates $(X(\theta),Y(\theta))'\in\mathbb{R}^2$, where $X(\theta)$ will be used to select among options while $Y(\theta)$ estimates a quantity of interest associated with option θ. For $|\Theta|$-dimensional vectors $X=\left(X(\theta_1),...,X\left(\theta_{|\Theta|}\right)\right)'$ and $Y=\left(Y(\theta_1),...,Y\left(\theta_{|\Theta|}\right)\right)'$ that collect these estimates, we assume that (X,Y) follow a joint normal distribution,

$$\begin{pmatrix} X \\ Y \end{pmatrix}\sim N(\mu,\Sigma) \tag{3}$$

for

$$E\left[\begin{pmatrix} X(\theta) \\ Y(\theta) \end{pmatrix}\right]=\mu(\theta)=\begin{pmatrix} \mu_X(\theta) \\ \mu_Y(\theta) \end{pmatrix},$$

[14] Indeed, conventional confidence sets correspond to Bayes credible sets under a flat prior, and we have already observed that they can undercover.

$$\Sigma(\theta,\tilde{\theta}) = \begin{pmatrix} \Sigma_X(\theta,\tilde{\theta}) & \Sigma_{XY}(\theta,\tilde{\theta}) \\ \Sigma_{YX}(\theta,\tilde{\theta}) & \Sigma_Y(\theta,\tilde{\theta}) \end{pmatrix} = Cov\left(\begin{pmatrix} X(\theta) \\ Y(\theta) \end{pmatrix}, \begin{pmatrix} X(\tilde{\theta}) \\ Y(\tilde{\theta}) \end{pmatrix} \right),$$

where Σ is known while μ is unknown. We abbreviate $\Sigma(\theta,\theta)$ to $\Sigma(\theta)$.

This model arises naturally as an asymptotic approximation in settings where we have asymptotically normal vectors of estimates $(\tilde{X}_n,\tilde{Y}_n)$ and a consistent estimator $\widetilde{\Sigma}_n$ for their variance matrix. Section 6 and Appendix F discuss the implementation of our approach using non-normal estimates and estimated variances, and show that in that case our procedures are uniformly asymptotically valid over large classes of data generating processes.

We assume that an option $\hat{\theta}$ is selected by picking the "winner" based on X,

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} X(\theta), \tag{4}$$

where we further assume that $\hat{\theta}$ is unique unless otherwise noted. We are interested in inference (e.g. estimators and confidence sets) for $\mu_Y(\hat{\theta})$, the mean of the element of $Y(\theta)$ associated with the selected option. Before turning to our formal inference goals and results, we discuss the motivation and interpretation of this setup, where one first selects a target by maximizing $X(\theta)$ and then conducts inference on an associated target parameter.

Motivation for "Picking the Winner" Our analysis takes as given that $\hat{\theta}$ is chosen to maximize $X(\theta)$. We view this as a reasonable point of departure since selection of this form arises in many different contexts. Moreover, while we do not study the problem of optimal selection in this paper, many previous recommendations from the optimal selection literature give rise to selections that can be written in the form (4).

Selecting $\hat{\theta}$ to maximize $X(\theta)$ seems particularly natural when our goal is to maximize $\mu_X(\theta)$. In the JOBSTART example of the last section, for instance, $X(\theta)$ corresponded to the estimated ATE at site θ, while $\mu_X(\theta)$ was the associated site-specific ATE. If our goal is to select the site where treatment is most effective it thus seems natural to pick $\hat{\theta}$ corresponding to the largest estimate. A number of recent papers in the econometrics literature propose selection rules of this form and prove that they are optimal in various senses, including Manski (2004), Hirano and Porter (2009), and Kitagawa and Tetenov (2018b).[15] There is also a large statistics literature which considers the problem of optimal

[15]Manski (2004) and Hirano and Porter (2009) study the problem of assigning a binary treatment based on a discrete covariate, which can be cast into our setting by letting θ index the possible treatment allocations (e.g. if there are three covariate values we could treat people with the first only, or with the first and second but not the third, and so on). Manski (2004) and Hirano and Porter (2009) establish

assignment. Lehmann (1966) and Eaton (1967) prove that $\hat{\theta}$ defined as in (4) corresponds to an optimal selection under a variety of optimality criteria when $\Sigma_X = Var(X)$ is proportional to the identity matrix, while Gupta and Miescke (1988) refer to $\hat{\theta}$ as the "natural rule" and discuss criteria under which this rule is optimal for general Σ_X.

By defining $X(\theta)$ appropriately, "picking the winner" also nests a number of additional cases that may not be immediately obvious:

- **Selection Based on Multiple Outcomes** In many contexts there will be multiple outcomes that matter for our choice of $\hat{\theta}$. In the JOBSTART example, for instance, we might care not just about earnings but also about educational attainment. So long as we can combine the outcomes of interest into a single index, for instance treating a completed GED as equivalent to a specific increase in earnings, we can cast this into our setting by defining $X(\theta)$ as the estimated effect on the index.

- **Selection Relative to a Fixed Threshold** Suppose that we are picking between two options, $\Theta = \{0,1\}$, and that we want to pick option 1 only if the associated estimate exceeds a threshold c. For instance, we might estimate the marginal value of public funds (MVPF) for some program (Hendren and Sprung-Keyser, 2020) and keep the program in place if and only if the estimated MVPF exceeds one. If we begin with a normally distributed estimate X^*, we can cast this into our setting by defining $X(0) = c$ and $X(1) = X^*$, so $X(1) > X(0)$ if and only if $X^* > c$.

- **Selection Based on Statistical Significance** Suppose that we want to pick the option that has the largest t-statistic against the null hypothesis of zero effect. If for each option θ we have a normally distributed estimate $X^*(\theta)$ with standard error $\sigma^*(\theta)$, we can cast this into our setting by defining $X(\theta) = X^*(\theta)/\sigma^*(\theta)$, so that largest element of X corresponds to the largest t-statistic.[16]

- **Selection Based on Posterior Means** As discussed in Gupta and Miescke (1988), a Bayesian looking to maximize the value of $\mu_X(\theta)$ associated with their selection would pick the option with the largest posterior mean. If we observe normally distributed estimates $X^* \sim N(\mu_X^*, \Sigma_X^*)$ and have a normal prior $\mu_X^* \sim N(\eta, \Omega)$, however,

finite-sample and asymptotic optimality properties for $\hat{\theta}$ in this setting, respectively, while Kitagawa and Tetenov (2018b) prove rate-optimality for analogous assignment rules in settings with continuous covariates.

[16] We might also want to incorporate a fixed significance threshold, for instance if we only pick one of the initial options when we conclude it is significantly better than zero based on a two-sided t-test. To cast this into our setting we can add an extra element θ_{null} to Θ, and define $X(\theta_{null}) = 1.96$ so we select θ_{null} when none of our estimates is positive and significant at the 5% level.

the vector of posterior means $X=(\Sigma_X^{*-1}+\Omega^{-1})^{-1}\left(\Sigma_X^{*-1}X^*+\Omega^{-1}\eta\right)$ is also normally distributed, and selection based on the posterior mean fits our setting.[17] Since many forms of linear shrinkage (e.g. ridge regression) are numerically equivalent to Bayes posterior means, selection based on such estimates is also covered.

- **Selection Based on Model-Implied Estimates** While the our examples consider selection based on estimates which do not impose an explicit economic model, selection using model-implied estimates also fits our setting. To illustrate, suppose that in the JOBSTART context we had a model which implied that we could write the ATE in site θ as $\mu_X(\theta)=\mu_X^*(W_\theta,\gamma)$ for W_θ a vector of observed site-level characteristics and γ a vector of model parameters. If we have an estimator $\hat{\gamma}$ for γ, standard regularity conditions (e.g. asymptotic normality of $\hat{\gamma}$, differentiability of μ_X^* in γ) will imply that the plug-in estimates $\mu_X^*(W_\theta,\hat{\gamma})$ are normally distributed in large samples. Hence we can cast this example into our setting by taking $X(\theta)=\mu_X^*(W_\theta,\hat{\gamma})$.

While selection of the form (4) covers many cases of interest, there are some situations that do not fit this model of selection. For instance, we might want to select the site with the largest estimated earnings increase, but restrict ourselves to those sites with a non-negative estimated effect on GED completion, which cannot naturally be cast into the form (4). In Appendix C we state theoretical results that allow for general conditioning events (specifically, we develop results that condition on $\gamma(X)=\tilde{\gamma}$ for $\gamma(\cdot)$ a user-selected function). The generality of these results means, however, that to apply them in practice some details would have to worked out on a case-by-case basis. We have worked out results for two alternative forms of selection in companion papers, considering selection on the absolute value of $X(\theta)$ (or more generally on $\|X(\theta)\|$ when $X(\theta)$ may be vector-valued) in Andrews, Kitagawa, and McCloskey (2021), and considering inference on the kth-best option in Andrews et al. (2022). In the present paper we focus on selection of the form (4) because it allows us to give fully worked out results that cover many cases of practical interest.

Motivation for Inference After Selection Taking as given the rule used to select $\hat{\theta}$, our goal is to construct estimators and confidence intervals for $\mu_Y(\hat{\theta})$. In many cases, as in Section 2 above, we are interested in the mean of the same variable that drives selection so $X=Y$ and $\mu_X=\mu_Y$. In other settings, however, we may select on one variable but want to do inference on the mean of another. Continuing with the JOBSTART example, we

[17] Note that in this case we use the prior only to inform the definition of X, and our inference results will not rely on the prior being correct.

might select $\hat{\theta}$ based on outcomes for all individuals, but want to conduct inference on average outcomes for some subgroup defined using covariates, for instance focusing on the effect for women. In this case we would define $Y(\theta)$ to be the estimated average outcome for women at site θ. As with the definition of X, our framework can incorporate a wide range of different target parameters by defining Y appropriately.

It is important to emphasize that since we take the rule generating $\hat{\theta}$ as given, the goal of inference on $\mu_Y(\hat{\theta})$ is *not* to guide the choice among the options Θ.[18] Instead we aim to conduct inference on a quantity of interest associated with the choice that was already made. Inference of this sort may be of interest for a variety of reasons. First, we may be interested in $\mu_Y(\theta)$ for the same sorts of scientific reasons that motivate other ex-post program evaluations not linked to an explicit prospective treatment choice problem. Second, in cases where a treatment or policy corresponding to θ has already been implemented and the results were not as hoped (as in the JOBSTART example) we may be interested in inference on $\mu_Y(\theta)$ in order to understand whether the disappointing results could be explained solely based on the winner's curse or whether there seem to be other factors at play. And third, in cases where some treatment or policy is going to be implemented in the future, we may be interested in forecasting the effect that it is going to have.

Unconditional and Conditional Inference For all of these purposes, we need reliable estimates and confidence sets for $\mu_Y(\hat{\theta})$. In particular, we will say that an estimator $\hat{\mu}_Y$ of $\mu_Y(\hat{\theta})$ is unconditionally median unbiased if

$$Pr\Big\{\hat{\mu}_Y \geq \mu_Y(\hat{\theta})\Big\} = \frac{1}{2} \text{ for all } \mu, \tag{5}$$

while a confidence interval CI has unconditional coverage $1-\alpha$ if

$$Pr\Big\{\mu_Y(\hat{\theta}) \in CI\Big\} \geq 1-\alpha \text{ for all } \mu. \tag{6}$$

Note that these probability statements integrate over the distribution of both X and Y, so the selection $\hat{\theta}$ and thus the target parameter $\mu_Y(\hat{\theta})$ are random variables. Hence, these notions of unbiasedness and coverage correspond to the case where we are interested in average performance across all realizations of $\hat{\theta}$.

In some cases, however, all realizations of $\hat{\theta}$ may not be of equal interest. For instance,

[18] Indeed, if we were to change our choice based on our winner's curse-corrected estimates and confidence intervals, this would effectively change the definition of $\hat{\theta}$, and so would necessitate further corrections to ensure valid inference.

it may be that only some of the options in Θ represent treatments that could plausibly be implemented, and we might only be interested in performance conditional on recommending one of these. Or if $\hat{\theta}$ is selected based on statistical significance, we might only be interested in results in the case where at least one treatment yields a significant result, as otherwise we think the findings are unlikely to be circulated or published. To formally discuss such cases, let us introduce a binary (latent) variable $S \in \{0,1\}$, where we are interested in performance when $S=1$ but not when $S=0$. The corresponding performance measures are thus

$$Pr\Big\{\hat{\mu}_Y \geq \mu_Y(\hat{\theta}) | S=1\Big\}, \quad Pr\Big\{\mu_Y(\hat{\theta}) \in CI | S=1\Big\}.$$

We assume that S is conditionally independent of (X,Y) given $\hat{\theta}$, so selection S depends on (X,Y) only through $\hat{\theta}$.[19] If we are willing to explicitly model the distribution of S conditional on $\hat{\theta}$ then we can use this model to correct for selection, as in the publication bias corrections of Andrews and Kasy (2019). In many contexts, however, the appropriate model for $S|\hat{\theta}$ is unclear. To ensure median unbiasedness and correct coverage for all possible conditional distributions of $S|\hat{\theta}$ it is necessary and sufficient to ensure conditional median unbiasedness and conditional coverage given each possible realization of $\hat{\theta}$:

$$Pr\Big\{\hat{\mu}_Y \geq \mu_Y(\hat{\theta}) | \hat{\theta}=\tilde{\theta}\Big\} = \frac{1}{2} \text{ for all } \tilde{\theta} \in \Theta \text{ and all } \mu, \tag{7}$$

$$Pr\Big\{\mu_Y(\hat{\theta}) \in CI | \hat{\theta}=\tilde{\theta}\Big\} \geq 1-\alpha \text{ for all } \tilde{\theta} \in \Theta \text{ and all } \mu. \tag{8}$$

Conditional median unbiasedness and coverage (7) and (8) are strictly stronger guarantees than their unconditional analogs (5) and (6). Hence they restrict the set of procedures we consider and can come at a substantial cost in terms of other performance criteria (e.g. the precision of estimators or the length of confidence intervals). We thus recommend that researchers view the unconditional criteria (5) and (6) as the "default" option, and enforce the more restrictive conditional criteria (7) and (8) only in settings where they need to guard against specific selection concerns.

Decision-Theoretic Motivation One can also relate our inference-after-selection problem to a formal decision-theoretic model, detailed in Appendix B. This model has two stages, where in the first stage a decisionmaker selects an option $\theta \in \Theta$ to maximize some

[19] This assumption is restrictive, and in Appendix C we extend our conditional inference results to cover the more general case where there is an additional variable $\hat{\gamma}=\gamma(X)$ such that S is independent of (X,Y) conditional on the pair $(\hat{\theta},\hat{\gamma})$.

objective, while in the second stage they report an interval that trades off (i) the probability of covering $\mu_Y(\hat{\theta})$ and (ii) the length of the interval. Decisionmakers have lexicographic preferences and strictly prioritize performance in the first stage, for instance because they value a better treatment recommendation more than precise inference, or because there are in fact different decisionmakers in the two stages and the first-stage decisionmaker is indifferent to the second-stage loss. We show that for a class of second-stage loss functions, minimax decision rules for the second stage necessarily ensure either unconditional or conditional coverage, depending on whether or not the second stage includes a selection problem. Hence, our coverage criteria emerge as necessary (although not in general sufficient) conditions for minimaxity in the second-stage problem. Importantly, both criteria imply that we need to cover $\mu_Y(\hat{\theta})$, the parameter associated with the (known) selected option, not the parameter $\mu_Y(\theta^*)$ associated with the (unknown) best option $\theta^* = \text{argmax}_{\theta\in\Theta}\mu_X(\theta)$.

4 Conditional Inference

While we recommend unconditional inference as the default option, our recommended unconditional inference procedures build on our conditional inference approach. Hence, we start by discussing conditional inference.

Our goal in this section is to develop estimators that are conditionally median-unbiased in the sense of (7), and confidence intervals that have conditional coverage $1-\alpha$ in the sense of (8). Our approach will be based on conditionally quantile unbiased estimators, where we say that $\hat{\mu}_\alpha$ is an α-quantile conditionally unbiased estimator if its overestimation probability conditional on $\hat{\theta}$ is exactly α:

$$Pr_\mu\left\{\hat{\mu}_\alpha \geq \mu_Y(\hat{\theta}) | \hat{\theta}=\tilde{\theta}\right\} = \alpha \text{ for all } \tilde{\theta}\in\Theta \text{ and all } \mu. \tag{9}$$

Since $\hat{\theta}$ is chosen as a function of X, conditioning on $\left\{\hat{\theta}=\tilde{\theta}\right\}$ is the same as conditioning on X falling in the set $\mathcal{X}(\tilde{\theta}) = \left\{X : \hat{\theta}=\tilde{\theta}\right\}$.[20] Hence, we are interested in inference on $\mu_Y(\tilde{\theta})$ conditional on $\left\{X\in\mathcal{X}(\tilde{\theta})\right\}$. Note, however, that since (X,Y) are unconditionally normally distributed, their joint distribution conditional on $\left\{X\in\mathcal{X}(\tilde{\theta})\right\}$ is multivariate truncated normal, and correlation between X and $Y(\tilde{\theta})$ implies that conditional on $\left\{\hat{\theta}=\tilde{\theta}\right\}$, $Y(\tilde{\theta})$ is no longer $N(\mu_Y(\tilde{\theta}),\Sigma_Y(\tilde{\theta}))$ distributed. To develop conditional inference procedures we

[20] If $\hat{\theta}$ is non-unique with positive probability, we change the conditioning event from $\hat{\theta}=\tilde{\theta}$ to $\tilde{\theta}\in\text{argmax}\,X(\theta)$.

thus need to understand the conditional distribution of $Y(\tilde{\theta})$ given $\left\{X \in \mathcal{X}(\tilde{\theta})\right\}$.

To account for the effect of conditioning, let

$$Z_{\tilde{\theta}} = X - \left(\Sigma_{XY}(\cdot,\tilde{\theta})/\Sigma_Y(\tilde{\theta})\right)Y(\tilde{\theta}). \tag{10}$$

This corresponds to the residual from the regression of X on $Y(\tilde{\theta})$ under their joint (unconditional) distribution. One can show that $Z_{\tilde{\theta}}$ is a minimal sufficient statistic for μ_X relative to the distribution of $(X,Y(\tilde{\theta}))$, so the conditional distribution of $(X,Y(\tilde{\theta}))|Z_{\tilde{\theta}}$ depends only on the parameter of interest $\mu_Y(\tilde{\theta})$. This remains true when we condition on $\left\{X \in \mathcal{X}(\tilde{\theta})\right\}$, and the conditional distribution of $Y(\tilde{\theta})$ given $\left\{\hat{\theta} = \tilde{\theta}, Z_{\tilde{\theta}} = z\right\}$ is a $N\left(\mu_Y(\tilde{\theta}),\Sigma_Y(\tilde{\theta})\right)$ distribution truncated to the set

$$\mathcal{Y}(\tilde{\theta},z) = \left\{y : z + \left(\Sigma_{XY}(\cdot,\tilde{\theta})/\Sigma_Y(\tilde{\theta})\right)y \in \mathcal{X}(\tilde{\theta})\right\}. \tag{11}$$

To derive estimators and confidence intervals based on this result, we need a tractable characterization for $\mathcal{Y}(\tilde{\theta},z)$. The following proposition, based on Lemma 5.1 of Lee et al. (2016), provides one such characterization.

Proposition 1
Let $\Sigma_{XY}(\tilde{\theta}) = Cov(X(\tilde{\theta}), Y(\tilde{\theta}))$. *For* $Z_{\tilde{\theta}}(\theta) = X(\theta) - \frac{\Sigma_{XY}(\theta,\tilde{\theta})}{\Sigma_Y(\tilde{\theta})}Y(\tilde{\theta})$ *the element of* $Z_{\tilde{\theta}}$ *corresponding to* θ, *define*

$$\mathcal{L}(\tilde{\theta},Z_{\tilde{\theta}}) = \max_{\theta \in \Theta : \Sigma_{XY}(\tilde{\theta}) > \Sigma_{XY}(\tilde{\theta},\theta)} \frac{\Sigma_Y(\tilde{\theta})\left(Z_{\tilde{\theta}}(\theta) - Z_{\tilde{\theta}}(\tilde{\theta})\right)}{\Sigma_{XY}(\tilde{\theta}) - \Sigma_{XY}(\tilde{\theta},\theta)},$$

$$\mathcal{U}(\tilde{\theta},Z_{\tilde{\theta}}) = \min_{\theta \in \Theta : \Sigma_{XY}(\tilde{\theta}) < \Sigma_{XY}(\tilde{\theta},\theta)} \frac{\Sigma_Y(\tilde{\theta})\left(Z_{\tilde{\theta}}(\theta) - Z_{\tilde{\theta}}(\tilde{\theta})\right)}{\Sigma_{XY}(\tilde{\theta}) - \Sigma_{XY}(\tilde{\theta},\theta)},$$

and

$$\mathcal{V}(\tilde{\theta},Z_{\tilde{\theta}}) = \min_{\theta \in \Theta : \Sigma_{XY}(\tilde{\theta}) = \Sigma_{XY}(\tilde{\theta},\theta)} -\left(Z_{\tilde{\theta}}(\theta) - Z_{\tilde{\theta}}(\tilde{\theta})\right).$$

If $\mathcal{V}(\tilde{\theta},z) \geq 0$, *then* $\mathcal{Y}(\tilde{\theta},z) = \left[\mathcal{L}(\tilde{\theta},z),\mathcal{U}(\tilde{\theta},z)\right]$. *If* $\mathcal{V}(\tilde{\theta},z) < 0$, *then* $\mathcal{Y}(\tilde{\theta},z) = \emptyset$.

Thus, $\mathcal{Y}(\tilde{\theta},z)$ is an interval bounded above and below by functions of z. To understand the form of these bounds, consider any $\tilde{\theta} \in \Theta$ and any z and y. For any $\theta \in \Theta$ we can use (10) to solve for the implied $X(\theta)$, $x(\theta;y,z) = z(\theta) + \Sigma_{XY}(\theta,\tilde{\theta})/\Sigma_Y(\tilde{\theta})y$. To have $y \in \mathcal{Y}(\tilde{\theta},z)$,

however, we must have $x(\theta;y,z) \leq x(\tilde{\theta};y,z)$ for all $\theta \in \Theta$. Collecting and rearranging these inequalities yields the result. One can further show that the requirement that $\mathcal{V}(\tilde{\theta},Z_{\tilde{\theta}}) \geq 0$ holds whenever $\hat{\theta} = \tilde{\theta}$. Hence, in applications we can safely ignore this constraint and calculate only $\mathcal{L}(\hat{\theta},Z_{\hat{\theta}})$ and $\mathcal{U}(\hat{\theta},Z_{\hat{\theta}})$.

Using Proposition 1 it is straightforward to construct quantile-unbiased estimators for $\mu_Y(\hat{\theta})$. Let $F_{TN}(y;\mu_Y(\tilde{\theta}),\tilde{\theta},z)$ denote the distribution function for a $N\Big(\mu_Y(\tilde{\theta}),\Sigma_Y(\tilde{\theta})\Big)$ distribution truncated to $\mathcal{Y}(\tilde{\theta},z)$. This function is again strictly decreasing in $\mu_Y(\tilde{\theta})$, so we can define $\hat{\mu}_\alpha$ as the unique solution to $F_{TN}(Y(\hat{\theta});\mu,\tilde{\theta},Z_{\tilde{\theta}}) = 1-\alpha$ in μ. This estimator is conditionally α-quantile unbiased for any $\alpha \in (0,1)$.

Proposition 2
For any $\alpha \in (0,1)$, $\hat{\mu}_\alpha$ is conditionally α-quantile-unbiased in the sense of (9).

Hence, $\hat{\mu}_{\frac{1}{2}}$ is conditionally median-unbiased in the sense of (7), while the equal-tailed interval $CI_{ET} = \left[\hat{\mu}_{\alpha/2},\hat{\mu}_{1-\alpha/2}\right]$ has conditional coverage $1-\alpha$ in the sense of (8). We show in Proposition 7 in the online appendix that under mild conditions, results in Pfanzagl (1979) and Pfanzagl (1994) imply that $\hat{\mu}_\alpha$ is optimal in the class of quantile-unbiased estimators.

One can further show that in the case where the selection problem is "easy" in the sense that $\hat{\theta}$ takes a given value with high probability, the median unbiased estimator and equal-tailed confidence interval reduce to the usual ones. To state this result, let CI_N denote the conventional confidence interval which ignores selection, $CI_N = \left[Y(\hat{\theta}) - c_{\alpha/2,N}\sqrt{\Sigma_Y(\hat{\theta})}, Y(\hat{\theta}) + c_{\alpha/2,N}\sqrt{\Sigma_Y(\hat{\theta})}\right]$, where $c_{\alpha,N}$ is the $1-\alpha$-quantile of the standard normal distribution. As we already saw in the simulation results of Section 2, this interval has approximately correct coverage when $Pr_\mu\left\{\hat{\theta} = \tilde{\theta}\right\}$ is close to one, so one might worry that our conditional inference procedures will be unnecessarily conservative in this case. As also previewed in Section 2 this problem does not arise, since the conditional and conventional approaches agree in this case.

Proposition 3
Consider any sequence of values μ_m such that $Pr_{\mu_m}\left\{\hat{\theta} = \tilde{\theta}\right\} \to 1$ as $m \to \infty$. Then under μ_m, $CI_{ET} \to_p CI_N$ and $\hat{\mu}_{\frac{1}{2}} \to_p Y(\tilde{\theta})$ both conditional on $\hat{\theta} = \tilde{\theta}$ and unconditionally, where for confidence intervals $\to_p$ denotes convergence in probability of the endpoints.

Additional Theoretical Results Appendix C generalizes our conditional inference results in two directions. First, we consider the case where the selection S depends not only on $\hat{\theta}$ but also on an additional variable $\hat{\gamma} = \gamma(X)$, and show that our results extend

immediately to inference conditional on $\left\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\right\}$. Second, we show how to construct an alternative type of optimal confidence interval, uniformly most accurate unbiased confidence intervals, which ensure that no incorrect parameter value is covered with probability higher than $1-\alpha$. These unbiased intervals are somewhat more computationally demanding to construct than the equal-tailed intervals.

In the remainder of this section we briefly discuss two other points related to conditional inference, first providing forecast intervals and then discussing split-sample inference.

4.1 Forecast Intervals

Rather than simply conducting inference on $\mu_Y(\tilde{\theta})$ we might be interested in forecasting subsequent outcomes, for instance results in a follow-up experiment as in the JOBSTART example discussed in Section 2. This is a somewhat different problem than inference on $\mu_Y(\tilde{\theta})$ since we must also account for the randomness in the subsequent outcome.

To pose this forecasting problem, suppose that in addition to observing (X,Y) as in (3), in the future we will observe an independent normal draw $Y_2 \sim N(\mu_Y,\Sigma_{Y_2})$ where we assume that the effect in the two stages is the same, $E[Y]=E[Y_2]=\mu_Y$, and Σ_{Y_2} is known (for instance because we know the sample size in the follow-up experiment). The assumption that Y and Y_2 have the same mean implies that for $Y_{1-2}=Y-Y_2$,

$$\begin{pmatrix} X \\ Y_{1-2} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ 0 \end{pmatrix},\begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y+\Sigma_{Y_2} \end{pmatrix}\right).$$

Hence the distribution of (X,Y_{1-2}) has the same form as that of (X,Y) except that we know the mean of Y_{1-2} is equal to zero.

We can use the same arguments as above to construct a confidence interval for $\mu_{Y_{1-2}}(\tilde{\theta})=E[Y(\tilde{\theta})-Y_2(\tilde{\theta})]$ conditional on $\hat{\theta}=\tilde{\theta}$. Denote this confidence interval by CI_{ET}^{1-2}, and note that since $\mu_{Y_{1-2}}(\tilde{\theta})=0$ by assumption, this interval covers zero with probability $1-\alpha$,

$$Pr_\mu\left\{0\in CI_{ET}^{1-2}|\hat{\theta}=\tilde{\theta}\right\}=1-\alpha \text{ for all } \tilde{\theta}\in\Theta \text{ and all } \mu. \tag{12}$$

Once we have observed $(X,Y(\hat{\theta}))$, however, we can solve for the range of values for $Y_2(\hat{\theta})$ such that 0 lies in the implied interval CI_{ET}^{1-2}, $FI=\left\{Y_2(\hat{\theta}):0\in CI_{ET}^{1-2}\right\}$. Equation (12) immediately implies that this forecast interval covers $Y_2(\hat{\theta})$ with probability $1-\alpha$,

$$Pr_\mu\left\{Y_2(\hat{\theta})\in FI|\hat{\theta}=\tilde{\theta}\right\}=1-\alpha \text{ for all } \tilde{\theta}\in\Theta \text{ and all } \mu,$$

so we have guaranteed coverage.[21] Appendix E provides further discussion.

4.2 Sample Splitting

An alternative remedy for winner's curse bias is to split the sample. If we have independent and identically distributed observations and select $\hat{\theta}_1$ based on the first half of the data, conventional estimates and confidence intervals for $\mu_Y(\hat{\theta}_1)$ constructed using the second half of the data will be conditionally valid given $\hat{\theta}_1$. In large samples, 50-50 sample splits yield a pair of independent and identically distributed normal draws (X_1,Y_1) and (X_2,Y_2), both of which follow the normal model (3), albeit with a different scaling for (μ,Σ) than in the full-sample case.[22] Conventional sample splitting procedures calculate $\hat{\theta}_1$ as in (4), replacing X by X_1, and use Y_2 for inference. Independence of X_1 and Y_2 implies that the conventional 95% sample-splitting confidence interval for $\mu_Y(\hat{\theta}_1)$, $\left[Y_2(\hat{\theta}_1)-1.96\sqrt{\Sigma_Y(\hat{\theta}_1)},Y_2(\hat{\theta}_1)+1.96\sqrt{\Sigma_Y(\hat{\theta}_1)}\right]$, has correct conditional coverage given $\hat{\theta}_1$, and $Y_2(\hat{\theta}_1)$ is conditionally median-unbiased for $\mu_Y(\hat{\theta}_1)$.

Sample splitting resolves the winner's curse but comes at multiple costs. First, and most importantly, $\hat{\theta}_1$ is based on less data than in the full-sample case. As discussed in the JOBSTART example, this will result in noisier choices of the target $\hat{\theta}_1$ and so will be undesirable in contexts where we care about the quality of the selection made (e.g. treatment choice problems). Second, split-sample inference effectively throws away the first half of the data after using it to pick $\hat{\theta}_1$, and so is inefficient – see Fithian, Sun, and Taylor (2017).

5 Unconditional Inference

We next turn to unconditional inference. As a first result, note that conditional median unbiasedness and conditional coverage imply their unconditional analogs provided $\hat{\theta}$ is unique with probability one.

Proposition 4

Suppose that $\hat{\theta}$ is unique with probability one for all μ. Then conditional median unbiasedness (7) implies unconditional median unbiasedness (5), and correct conditional coverage (8) implies correct unconditional coverage (6).

For uniqueness of $\hat{\theta}$ it suffices that the elements of X are not perfectly positively correlated.

[21] FI is a predictive interval based on a similar test: see e.g. Chapter 10 in Young and Smith (2005).

[22] An analogous statement is also true for uneven sample splits.

Lemma 1

Suppose that at most one $\theta \in \Theta$ *has* $\Sigma_X(\theta) = 0$ *and for all other* θ, $\tilde{\theta} \in \Theta$ *such that* $\theta \neq \tilde{\theta}$, $X(\theta)$ *and* $X(\tilde{\theta})$ *are not perfectly positively correlated. Then* $\hat{\theta}$ *is unique with probability one for all* μ.

Hence, to ensure unconditional median unbiasedness and unconditional coverage we may continue to use the conditional procedures developed in the last section. Relaxing our requirements to unconditional median unbiasedness and unconditional coverage may, however, allow us to improve performance in some settings. This section explores this possibility.

5.1 Projection Confidence Intervals

One approach to obtain an unconditional confidence interval for $\mu_Y(\hat{\theta})$ is to start with a joint confidence set for the vector $\mu_Y \in \mathbb{R}^{|\Theta|}$ and then report the implied interval for $\mu_Y(\hat{\theta})$. To formally describe this approach, let c_α denote the $1-\alpha$ quantile of $\max_\theta |\xi(\theta)|/\sqrt{\Sigma_Y(\theta)}$ for $\xi \sim N(0, \Sigma_Y)$. Note that this corresponds to the $1-\alpha$ quantile of the maximum absolute studentized estimation error for μ_Y, $\max_{\theta \in \Theta} |Y(\theta) - \mu_Y(\theta)|/\sqrt{\Sigma_Y(\theta)}$, from which it follows that

$$CS_{\mu_Y} = \left\{ \mu_Y : |Y(\theta) - \mu_Y(\theta)| \leq c_\alpha \sqrt{\Sigma_Y(\theta)} \text{ for all } \theta \in \Theta \right\}$$

is a level $1-\alpha$ confidence set for μ_Y, $Pr_\mu\{\mu_Y \in CS_{\mu_Y}\} \geq 1-\alpha$. If we then define

$$CI_P = \left\{ \mu_Y(\hat{\theta}) : \exists \tilde{\mu} \in CS_{\mu_Y} \text{ such that } \mu_Y(\hat{\theta}) = \tilde{\mu}_Y(\hat{\theta}) \right\} = \left[Y(\hat{\theta}) - c_\alpha \sqrt{\Sigma_Y(\hat{\theta})}, Y(\hat{\theta}) + c_\alpha \sqrt{\Sigma_Y(\hat{\theta})} \right]$$

as the projection of CS_{μ_Y} on the dimension corresponding to $\hat{\theta}$, then since $\mu_Y \in CS_{\mu_Y}$ implies $\mu_Y(\hat{\theta}) \in CI_P$, CI_P satisfies the unconditional coverage requirement (6).[23]

The width of the projection interval CI_P depends on the variance $\Sigma_Y(\hat{\theta})$ but does not otherwise depend on the data.[24] To account for the randomness of $\hat{\theta}$, the critical value c_α is

[23] Similar projection approaches were used by Romano and Wolf (2005) in the context of multiple testing, by Kitagawa and Tetenov (2018a) for inference on welfare at an estimated optimal policy, and by a large and growing statistics literature on post-selection inference including Berk et al. (2013), Bachoc, Preinerstorfer, and Steinberger (2020) and Kuchibhotla et al. (2020). An advantage of the projection method, not shared by the conditional or hybrid approaches, is that the projection method is valid without any restriction on how selection is performed, that is, we can construct projection intervals without specifying how $\hat{\theta}$ depends on X.

[24] The projection interval described here is "balanced" in the same sense as a "balanced" simultaneous confidence band/set: it adds and subtracts the same multiple of the standard deviation from the estimate of $\mu_Y(\hat{\theta})$ regardless of the value $\hat{\theta}$ takes. With little modification to our analysis, one could consider alternative projection intervals, for instance optimized to have shorter length at some $\hat{\theta}$ values in exchange for greater length at others. See Freyberger and Rai (2018), Olea and Plagborg-Moller (2019), and Frandsen (2020).

typically larger than the conventional two-sided normal critical value. For instance, if Σ_Y is diagonal (so the elements of Y are independent), c_α is approximately equal to 2.24 when $|\Theta|=2$, 2.8 when $|\Theta|=10$, and 3.28 when $|\Theta|=50$. Hence, as we already saw in Section 2, CI_P will be conservative in cases where $\hat{\theta}$ takes a given value $\tilde{\theta}$ with high probability.[25] To improve performance in such cases, we propose a hybrid inference approach.

5.2 Hybrid Inference

As shown in Section 2, the conditional and projection approaches each have good unconditional performance in some cases, but neither is fully satisfactory. Hybrid inference combines the approaches to obtain good performance over a wide range of parameter values.

As with our conditional approach, hybrid inference will be based on conditionally quantile-unbiased estimators. Hybrid inference changes the conditioning event, however, and conditions both on $\hat{\theta}=\tilde{\theta}$ and on the event that $\mu_Y(\hat{\theta})$ lies in the level $1-\beta$ projection confidence interval CI_P^β for $0\leq\beta<\alpha$. The implied set of values for $Y(\tilde{\theta})$ becomes

$$\mathcal{Y}^H(\tilde{\theta},\mu_Y(\tilde{\theta}),z)=\mathcal{Y}(\tilde{\theta},z)\cap\left[\mu_Y(\tilde{\theta})-c_\beta\sqrt{\Sigma_Y(\tilde{\theta})},\mu_Y(\tilde{\theta})+c_\beta\sqrt{\Sigma_Y(\tilde{\theta})}\right].$$

Let $F_{TN}^H(y;\mu_Y(\tilde{\theta}),\tilde{\theta},z)$ denote the distribution function for a $N(\mu_Y(\tilde{\theta}),\Sigma_Y(\tilde{\theta}))$ distribution truncated to $\mathcal{Y}^H(\tilde{\theta},\mu_Y(\tilde{\theta}),z)$ and define $\hat{\mu}_\alpha^H$ to solve $F_{TN}^H(Y(\hat{\theta});\mu,\hat{\theta},Z_{\tilde{\theta}})=1-\alpha$ in μ. The hybrid estimator $\hat{\mu}_\alpha^H$ is α-quantile unbiased conditional on $\mu(\hat{\theta})\in CI_P^\beta$.

Proposition 5
For $\alpha\in(0,1)$, $\hat{\mu}_\alpha^H$ is unique and $\hat{\mu}_\alpha^H\in CI_P^\beta$. If $\hat{\theta}$ is unique almost surely for all μ, $\hat{\mu}_\alpha^H$ is α-quantile unbiased conditional on $\mu_Y(\hat{\theta})\in CI_P^\beta$:

$$Pr_\mu\left\{\hat{\mu}_\alpha^H\geq\mu_Y(\hat{\theta})|\mu_Y(\hat{\theta})\in CI_P^\beta\right\}=\alpha \text{ for all } \mu.$$

Proposition 5 implies several notable properties for the hybrid quantile-unbiased estimator $\hat{\mu}_\alpha^H$. First, since $Pr_\mu\left\{\mu_Y(\hat{\theta})\in CI_P^\beta\right\}\geq 1-\beta$, one can show that

$$\left|Pr_\mu\left\{\hat{\mu}_\alpha^H\geq\mu_Y(\hat{\theta})\right\}-\alpha\right|\leq\beta\cdot\max\{\alpha,1-\alpha\} \text{ for all } \mu.$$

Indeed, rather than using such "balanced" projection intervals for the application in Section 7, we instead use fixed-length projection intervals for computational reasons. See Appendix G for details.

[25] Zrnic and Fithian (2022) propose a novel unconditional inference approach, building on projection ideas, that allows the length of the interval to adjust when there is a clear winner.

This implies that the absolute median bias of $\hat{\mu}^H_{\frac{1}{2}}$ (measured as the deviation of the overestimation probability from 1/2) is bounded above by $\beta/2$. On the other hand, since $\hat{\mu}^H_{\frac{1}{2}} \in CI^\beta_P$ we have $\left|\hat{\mu}^H_{\frac{1}{2}} - Y(\hat{\theta})\right| \leq c_\beta \sqrt{\Sigma_Y(\tilde{\theta})}$, so the difference between $\hat{\mu}^H_{\frac{1}{2}}$ and the conventional estimator $Y(\hat{\theta})$ is bounded above by half the width of CI^β_P.

As with the quantile-unbiased estimator $\hat{\mu}_\alpha$, we can form confidence intervals based on hybrid estimators. In particular, the interval $[\hat{\mu}^H_{\alpha/2}, \hat{\mu}^H_{1-\alpha/2}]$ has coverage $1-\alpha$ conditional on $\mu_Y(\hat{\theta}) \in CI^\beta_P$. This is not fully satisfactory, however, as $Pr_\mu\{\mu_Y(\hat{\theta}) \in CI^\beta_P\} < 1$. Hence, to ensure correct coverage, we define the level $1-\alpha$ hybrid confidence interval as $CI^H_{ET} = \left[\hat{\mu}^H_{\frac{\alpha-\beta}{2(1-\beta)}}, \hat{\mu}^H_{1-\frac{\alpha-\beta}{2(1-\beta)}}\right]$. With this adjustment, hybrid confidence intervals have coverage at least $1-\alpha$ both conditional on $\mu_Y(\hat{\theta}) \in CI^\beta_P$ and unconditionally.

Proposition 6

Provided $\hat{\theta}$ is unique with probability one for all μ, the hybrid confidence interval CI^H_{ET} has coverage $\frac{1-\alpha}{1-\beta}$ conditional on $\mu_Y(\hat{\theta}) \in CI^\beta_P$:

$$Pr_\mu\left\{\mu_Y(\hat{\theta}) \in CI^H_{ET} | \mu_Y(\hat{\theta}) \in CI^\beta_P\right\} = \frac{1-\alpha}{1-\beta} \text{ for all } \mu.$$

Moreover, its unconditional coverage is between $1-\alpha$ and $\frac{1-\alpha}{1-\beta}$:

$$\inf_\mu Pr_\mu\left\{\mu_Y(\hat{\theta}) \in CI^H_{ET}\right\} \geq 1-\alpha, \quad \sup_\mu Pr_\mu\left\{\mu_Y(\hat{\theta}) \in CI^H_{ET}\right\} \leq \frac{1-\alpha}{1-\beta}.$$

Hybrid confidence intervals strike a balance between the conditional and projection approaches. The maximal length of hybrid confidence intervals is bounded above by the length of CI^β_P. For small β, hybrid confidence intervals will be close to conditional confidence intervals, and thus to conventional confidence intervals, when $\hat{\theta} = \tilde{\theta}$ with high probability. For $\beta > 0$, however, hybrid confidence intervals do not fully converge to conventional confidence intervals as $Pr_\mu\left\{\hat{\theta} = \tilde{\theta}\right\} \to 1$, which is a disadvantage of hybrid intervals relative to the conditional approach. Nevertheless, our simulations in Section 2 find similar performance for the hybrid and conditional approaches in cases with a clear winner.

While hybrid confidence intervals combine the conditional and projection approaches, they can yield overall performance more appealing than either. In Section 2 we found that hybrid confidence intervals had a shorter median length for a wide range of parameter values than did either the conditional or projection approaches used in isolation. Our simulation

results in Section 7 below provide further evidence of outperformance in realistic settings. One can also use our hybrid approach to develop unconditional forecast intervals analogous to the conditional forecast intervals we discussed in Section 4.1. See Appendix E for details.

Choice of β To use the hybrid approach we must select the coverage β of the initial projection interval CI_P^β. Intuitively this choice trades off the length of CI_P^β, which bounds the worst-case length of CI_{ET}^H in the poorly-separated case, against the length of CI_{ET}^H in the well-separated case. The length of CI_{ET}^H in the well-separated case is bounded above by the length of the level $\frac{1-\alpha}{1-\beta}$ conventional confidence interval. For the standard choice of $\alpha=5\%$, choosing $\beta=\frac{\alpha}{10}=0.5\%$ implies that the CI_{ET}^H has half-length no more than 2.0025 standard errors in the well-separated case, compared to a half-length of 1.96 standard errors for the conventional 95% interval. Hence, we suggest $\beta=\frac{\alpha}{10}$ as a default choice, and use this value of β in our simulations and applications.[26]

6 Feasible Inference and Large-Sample Results

Our results have so far assumed that (X,Y) are jointly normal with known variance Σ. While exact normality is rare in practice, standard approaches to inference rely on the assumption that estimators are approximately normally distributed with a variance that we can estimate well, typically justified by appealing to large-sample asymptotic results. Our results for the finite-sample normal model translate to asymptotic results under the same conditions. In this section we summarize how to implement our estimators and confidence intervals in practice and briefly describe the translation of our finite-sample results to asymptotic results. Appendix F provides formal theoretical results demonstrating the uniform asymptotic validity of our approach over large classes of data generating processes.

Suppose that for sample size n we have vectors of estimates $(\tilde{X}_n,\tilde{Y}_n)$ where we define the option of interest as $\hat{\theta}_n=\text{argmax}_{\theta\in\Theta}\tilde{X}_n(\theta)$. We are interested in the mean of $\tilde{Y}_n(\hat{\theta}_n)$. We write $(\tilde{X}_n,\tilde{Y}_n)$ rather than (X,Y) to emphasize that (i) $(\tilde{X}_n,\tilde{Y}_n)$ may be non-normal, unlike (X,Y) and (ii) $(\tilde{X}_n,\tilde{Y}_n)$ are associated with the sample of size n. In the JOBSTART example discussed in Section 2, for instance, θ indexes treatments, $\tilde{X}_n(\theta)$ is the ATE

[26] Romano, Shaikh, and Wolf (2014) and McCloskey (2017) likewise find this choice to perform well in two different settings when using a Bonferroni correction. A simple Bonferroni approach for our setting intersects a level $1-\beta$ projection confidence interval CI_P^β with a level $1-\alpha+\beta$ conditional interval that conditions only on $\hat{\theta}=\tilde{\theta}$. Bonferroni intervals differ from our hybrid approach in two respects. First, they use a level $1-\alpha+\beta$ conditional confidence interval, while the hybrid approach uses a level $\frac{1-\alpha}{1-\beta}$ conditional interval, where $\frac{1-\alpha}{1-\beta}\leq 1-\alpha+\beta$. Second, the conditional interval used by the Bonferroni approach does not condition on $\mu_Y(\tilde{\theta})\in CI_P^\beta$, while that used by the hybrid approach does. Consequently, one can show that hybrid confidence intervals exclude the endpoints of CI_P^β almost surely, while the same is not true of Bonferroni intervals.

estimate at site θ, and $\tilde{Y}_n(\theta)=\tilde{X}_n(\theta)$. While $(\tilde{X}_n,\tilde{Y}_n)$ may be non-normally distributed, we assume that they are asymptotically normal in the usual sense: once recentered by vectors $(\tilde{\mu}_{X,n},\tilde{\mu}_{Y,n})$ and scaled by $\sqrt{n}$, their joint distribution is approximately normal for large n,

$$\sqrt{n}\begin{pmatrix} \tilde{X}_n-\tilde{\mu}_{X,n} \\ \tilde{Y}_n-\tilde{\mu}_{Y,n} \end{pmatrix} \Rightarrow N(0,\Sigma), \tag{13}$$

where $\Rightarrow$ denotes convergence in distribution. In the JOBSTART example, $\tilde{\mu}_{X,n}(\theta)=\tilde{\mu}_{Y,n}(\theta)$ is the ATE at site θ. We further assume that we have a variance estimator $\widetilde{\Sigma}_n$ such that $n\cdot\widetilde{\Sigma}_n$ is consistent for the asymptotic variance Σ. In the JOBSTART example, since $\tilde{Y}_n(\theta)=\tilde{X}_n(\theta)$ and the estimates are independent across sites, $\widetilde{\Sigma}_n$ consists of four copies of a diagonal matrix $\widetilde{\Sigma}_{X,n}$, where the diagonal entry of $\widetilde{\Sigma}_{X,n}$ corresponding to site θ is simply the squared standard error for that site.

More broadly, $(\tilde{X}_n,\tilde{Y}_n)$ can be any vectors of asymptotically normal estimators, and we should calculate $\widetilde{\Sigma}_n$ however we usually would for inference on $(\tilde{\mu}_{X,n},\tilde{\mu}_{Y,n})$, including corrections for clustering, serial correlation, and so on in the usual way. Feasible inference based on our approach simply substitutes $(\tilde{X}_n,\tilde{Y}_n)$ and $\widetilde{\Sigma}_n$ in place of (X,Y) and Σ in all expressions.

Alternatively, to implement our approach directly, one should compute

$$\tilde{Z}_{\hat{\theta},n}=\tilde{X}_n-\left(\widetilde{\Sigma}_{XY,n}(\cdot,\hat{\theta}_n)/\widetilde{\Sigma}_{Y,n}(\hat{\theta}_n)\right)\tilde{Y}_n(\hat{\theta}_n),$$

and then calculate $\mathcal{Y}(\hat{\theta}_n,\tilde{Z}_{\hat{\theta},n})=\left[\mathcal{L}(\hat{\theta}_n,\tilde{Z}_{\hat{\theta},n}),\mathcal{U}(\hat{\theta}_n,\tilde{Z}_{\hat{\theta},n})\right]$, where

$$\mathcal{L}(\hat{\theta}_n,\tilde{Z}_{\hat{\theta},n})=\max_{\theta\in\Theta:\widetilde{\Sigma}_{XY,n}(\hat{\theta}_n)>\widetilde{\Sigma}_{XY,n}(\hat{\theta}_n,\theta)}\frac{\widetilde{\Sigma}_{Y,n}(\hat{\theta}_n)\left(\tilde{Z}_{\hat{\theta},n}(\theta)-\tilde{Z}_{\hat{\theta},n}(\hat{\theta}_n)\right)}{\widetilde{\Sigma}_{XY,n}(\hat{\theta}_n)-\widetilde{\Sigma}_{XY,n}(\hat{\theta}_n,\theta)},$$

$$\mathcal{U}(\hat{\theta}_n,\tilde{Z}_{\hat{\theta},n})=\min_{\theta\in\Theta:\widetilde{\Sigma}_{XY,n}(\hat{\theta}_n)<\widetilde{\Sigma}_{XY,n}(\hat{\theta}_n,\theta)}\frac{\widetilde{\Sigma}_{Y,n}(\hat{\theta}_n)\left(\tilde{Z}_{\hat{\theta},n}(\theta)-\tilde{Z}_{\hat{\theta},n}(\hat{\theta}_n)\right)}{\widetilde{\Sigma}_{XY,n}(\hat{\theta}_n)-\widetilde{\Sigma}_{XY,n}(\hat{\theta}_n,\theta)}.$$

For $F_{TN}(y;\mu_Y(\tilde{\theta}),\tilde{Z}_{\tilde{\theta},n})$ the distribution function for a $N\left(\mu_Y(\tilde{\theta}),\widetilde{\Sigma}_{Y,n}(\tilde{\theta})\right)$ distribution truncated to $\mathcal{Y}(\tilde{\theta},\tilde{Z}_{\tilde{\theta},n})$, the conditional estimator $\hat{\mu}_{\alpha,n}$ is then the unique solution to $F_{TN}(\tilde{Y}_n(\hat{\theta}_n);\mu,\tilde{Z}_{\hat{\theta},n})=1-\alpha$ in μ. Our conditionally median unbiased estimator is $\hat{\mu}_{\frac{1}{2},n}$, while our conditional equal-tailed confidence interval is $CI_{ET,n}=\left[\hat{\mu}_{\frac{\alpha}{2},n},\hat{\mu}_{1-\frac{\alpha}{2},n}\right]$.

To implement our hybrid approach, one must also approximate the projection critical value c_β, which can be done by simulation. Specifically, for S a large number (e.g. $S=10^4$),

independently draw S normal random vectors $\xi_1,...,\xi_S$ where for each s, $\xi_s \sim N(0,\widetilde{\Sigma}_{Y,n})$. For each S compute the maximum absolute studentized deviation $\max_{\theta\in\Theta}|\xi_s(\theta)|/\sqrt{\widetilde{\Sigma}_{Y,n}(\theta)}$, and define $\hat{c}_\beta$ as the $1-\beta$ quantile of these maximum absoluate deviations across the simulation draws. Define

$$\mathcal{Y}^H(\tilde{\theta},\mu_Y(\tilde{\theta}),\tilde{Z}_{\tilde{\theta},n}) = \mathcal{Y}(\tilde{\theta},\tilde{Z}_{\tilde{\theta},n}) \cap \left[\mu_Y(\tilde{\theta}) - \hat{c}_\beta\sqrt{\widetilde{\Sigma}_{Y,n}(\tilde{\theta})}, \mu_Y(\tilde{\theta}) + \hat{c}_\beta\sqrt{\widetilde{\Sigma}_{Y,n}(\tilde{\theta})}\right],$$

and let $F^H_{TN}(y;\mu_Y(\tilde{\theta}),\tilde{Z}_{\tilde{\theta},n})$ denote the distribution function for a $N\left(\mu_Y(\tilde{\theta}),\widetilde{\Sigma}_{Y,n}(\tilde{\theta})\right)$ distribution truncated to $\mathcal{Y}^H(\tilde{\theta},\mu_Y(\tilde{\theta}),\tilde{Z}_{\tilde{\theta},n})$. The conditional estimator $\hat{\mu}^H_{\alpha,n}$ is then the unique solution to $F^H_{TN}(\tilde{Y}_n(\hat{\theta}_n);\hat{\mu},\tilde{Z}_{\hat{\theta},n}) = 1-\alpha$ in μ. Our approximately median unbiased hybrid estimator is $\hat{\mu}^H_{\frac{1}{2},n}$, while our hybrid equal-tailed confidence interval is $CI^H_{ET,n} = \left[\hat{\mu}^H_{\frac{\alpha-\beta}{2(1-\beta)},n}, \hat{\mu}^H_{1-\frac{\alpha-\beta}{2(1-\beta)},n}\right]$.

Appendix F shows that this plug-in approach yields uniformly asymptotically valid inference on $\mu_{Y,n}(\hat{\theta}_n)$. Loosely speaking, we suppose that the data in the sample of size n can be generated from any distribution P in a class $\mathcal{P}_n$. We show that if the classes $\mathcal{P}_n$ are such that as $n\to\infty$ (i) the convergence in distribution (13) holds uniformly over $P\in\mathcal{P}_n$, where the variance Σ may depend on P, (ii) $n\cdot\widetilde{\Sigma}$ is consistent for Σ, and (iii) the diagonal elements of Σ are bounded above and away from zero and the pairwise correlations in Σ are bounded away from one, then all of our finite-sample results for the normal model translate to asymptotic results. These conditions are quite weak, and the asymptotic normality and consistent variance estimation that we require is precisely the same as that used to justify standard t-statistic-based inference. Hence our approach may be applied when, absent winner's curse concerns, we would usually apply standard large-sample inference methods.

The uniformity of our asymptotic approximations is important, since it ensures that our procedures remain reliable even in cases where there are ties or near-ties for the "best" option, as in the normal model when multiple choices θ imply nearly the same value of $\mu_X(\theta)$. In this sense, uniform asymptotic validity is the asymptotic analog of our requirement in the normal model that our procedures be valid no matter the value of μ. By contrast, for procedures that are not uniformly valid, even for arbitrarily large samples there exist data generating processes where the procedure yields unreliable results, and these distortions (e.g. undercoverage for confidence sets) can be quantitatively large. Consequently, results from winner's curse corrections that are not uniformly asymptotically valid, or for which uniform asymptotic validity has not been established, should be treated with caution.

One limitation of our uniformity results is that we treat the dimension of $(\tilde{X}_n, \tilde{Y}_n)$ (i.e. $|\Theta|$) as fixed when $n \to \infty$, which rules out settings where the number of options considered grows with the sample size. Extension of our results to high-dimensional settings where $|\Theta|$ grows with n is an interesting topic for future work.

7 Application: Neighborhood Effects

We next discuss simulation and empirical results based on Chetty et al. (2020) and Bergman et al. (2023). Earlier work, including Chetty and Hendren (2018a) and Chetty and Hendren (2018b) argues that the neighborhood in which a child grows up has a long-term causal impact on income in adulthood, as well as other outcomes. Moreover, they show that the causal impact of moving to a given neighborhood is closely related to the average outcome for children already living there, and that these causal effects explain much of the observed difference in average outcomes across neighborhoods.

Motivated by these findings, Bergman et al. (2023) partnered with the public housing authorities in Seattle and King County in Washington State in an experiment aiming to help housing voucher recipients with children move to a set of higher-opportunity target neighborhoods. Bergman et al. (2023) choose target neighborhoods based on the Chetty et al. (2020) "Opportunity Atlas." This atlas compiles census-tract level estimates of economic mobility for communities across the United States. Bergman et al. (2023) define target neighborhoods by selecting approximately the top third of tracts in Seattle and King County based on estimated economic mobility, where their measure for economic mobility is the average household income rank in adulthood for children growing up at the 25th percentile of the income distribution (see Chetty et al., 2020). They then make "relatively minor" adjustments to the set of target tracts based on other criteria (Bergman et al., 2023, Appendix A).

A central empirical question in this setting is whether families moving to the target tracts will in fact experience the positive outcomes predicted based on the Opportunity Atlas estimates and the hypothesis of neighborhood effects. Once long-term outcomes for the experimental sample are available, researchers will be able to answer this question by comparing outcomes for children in treated families to the Opportunity Atlas estimates used to select the target tracts in the first place. Such a comparison is complicated by the winner's curse, however: the Atlas estimates were already used to select the target tracts, so the conventional estimate for the causal effect of the selected tracts will be systematically biased upwards. Our winner's curse corrections address precisely this bias.

Motivated by related concerns, Chetty et al. (2020) and Bergman et al. (2023) adopt a

shrinkage or empirical Bayes approach. Their estimates correspond to Bayesian posterior means under a prior that takes tract-level economic mobility to be normally distributed conditional on a vector of observable tract characteristics, and then estimates mean and variance hyperparameters from the data. One can show (see Appendix G.3) that if the normal prior correctly describes the distribution of economic mobility, then the posterior median for average economic mobility over selected tracts will be median-unbiased under the prior, and Bayesian credible sets will have correct coverage. This guarantee for the empirical Bayes approach depends critically on the correct specification of the prior, however: if the distribution across tracts is non-normal, then these empirical Bayes estimates and credible sets do not in general solve the winner's curse. By contrast, our results ensure correct coverage and controlled median bias for all possible distributions of economic mobility across tracts. Given the widespread use of normality-based empirical Bayes approaches in applications, we include empirical Bayes procedures in our analysis.

Simulation Results To examine the extent of winner's curse bias and the performance of different corrections, we calibrate simulations to the Opportunity Atlas data. For each of the 50 largest CZs in the United States we treat the (un-shrunk) tract-level Opportunity Atlas estimates as the true values.[27] We then simulate estimates by adding normal noise with standard deviation equal to the Opportunity Atlas standard error.[28] Since these simulations impose normality of the estimation errors by construction, they are not informative about the quality of the normal approximation for these errors.[29]

We are interested in understanding the extent to which programs like the one studied in Bergman et al. (2023), which target tracts with higher estimated mobility, succeed in picking higher-mobility tracts relative to the tracts in which voucher-recipent households currently live. Specifically, we define target tracts in each CZ as the top third of tracts for estimated economic mobility.[30] Our parameter of interest is then the average economic mobility across

[27] We use un-shrunk estimates here, rather than e.g. the normal distribution implicit in empirical Bayes, since the distribution of estimates in many CZs is strongly suggestive of non-normality. See Appendix G.6 for details. Using the realized estimates will tend to overstate the variance of mobility across tracts, but since the winner's curse is more severe when true effects are close together this should bias us against finding distortions from the winner's curse.

[28] We base our estimates in this setting on the public Opportunity Atlas estimates and standard errors since we do not have access to the underlying microdata. We also do not have access to the correlation structure of the estimates across tracts. Such correlations arise from individuals who move across tracts, and there are few movers between most pairs of tracts, so we expect that these omitted correlations are small.

[29] Appendix H reports additional simulation and empirical results, based on an experiment by Karlan and List (2007) studying charitable giving, where we we have access to the microdata and so do not have to impose normality. We find that our approaches continue to perform well in that setting.

[30] We select the target tracts based on the un-shrunk estimates, rather than shrunk estimates as in

targeted tracts, minus the weighted-average mobility across all tracts in the CZ, where we weight based on the number of voucher-recipient households with children in each tract.[31]

Formally, let $\mathcal{T}$ be the set of tracts in a given CZ and Θ the set of selections from $\mathcal{T}$ containing one third of tracts, rounded down, $\Theta = \{\theta \subset \mathcal{T} : |\theta| = \lfloor |\mathcal{T}|/3 \rfloor\}$. Let μ_t be the true economic mobility for tract t (that is, the average household income rank in adulthood for children growing up in this tract and in households at the 25th percentile of the income distribution). Define $X(\theta)$ as the average estimate over tracts in θ, $X(\theta) = \frac{1}{|\theta|}\sum_{t\in\theta}\hat{\mu}_t$, and let $\hat{\theta}$ select the top third of tracts, $\hat{\theta} = \text{argmax}_{\theta\in\Theta} X(\theta)$. For c_t the number of voucher households with children in tract t in 2018, the year Bergman et al. (2023)'s experiment began, our quantity of interest is $\mu_Y(\hat{\theta}) = \frac{1}{|\hat{\theta}|}\sum_{t\in\hat{\theta}}\mu_t - \frac{\sum_{t\in\mathcal{T}} c_t\mu_t}{\sum_{t\in\mathcal{T}} c_t}$, and we correspondingly define $Y(\theta) = \frac{1}{|\hat{\theta}|}\sum_{t\in\hat{\theta}}\hat{\mu}_t - \frac{\sum_{t\in\mathcal{T}} c_t\hat{\mu}_t}{\sum_{t\in\mathcal{T}} c_t}$.[32] We study the performance of conventional estimates and confidence intervals, empirical Bayes estimates and credible sets based on a normal prior, and our corrected estimates and confidence intervals.

Figure IV reports results based on ten thousand simulation draws. Panel (a) plots the distribution of the mean of our target parameter, $E\left[\mu_Y(\hat{\theta})\right]$, across the 50 CZs considered. Targeted tracts are associated with higher than-average mobility (that is, $E\left[\mu_Y(\hat{\theta})\right] > 0$) across all 50 CZs, though the magnitude ranges from a 6.44 to 18.04 percentile-point difference in earnings between the target tracts and the weighted CZ average. Panel (b) plots the distribution of the standard deviation of mobility across the 50 CZs, which ranges from 4.65 percentile points to 8.98 percentile points. Panel (c) shows the distributions of the median bias for the estimators we consider. As expected the conventional estimator is biased upwards, with magnitude ranging from 0.72 percentile points to 1.88 percentile

Bergman et al. (2023). We do this because we find that selecting based on un-shrunk estimates yields slightly higher average mobility for selected tracts than selecting on shrunk estimates, and because selection based on shrunk estimates introduces nonlinearity (due to estimation of the degree of shrinkage to use) which complicates conditional and hybrid inference.

[31] Specifically, we compute weights based on data from US Department of Housing and Urban Development (2018).

[32] $\mu_Y(\hat{\theta})$ corresponds to the average change in tract-level mobility from moving a randomly selected voucher-recipient household with children from their initial location in the CZ to a randomly selected target tract. There are several reasons this need not correspond to the average treatment effect from the experiment in Bergman et al. (2023). First, even with additional support to move to one of the targeted tracts, some households may choose to locate elsewhere. If this is unrelated to baseline location and location choice conditional on moving to a target tract, the average effect in this case will simply be a scaled-down version of $\mu_Y(\hat{\theta})$. Second, households that do move to a targeted tract will not in general choose a tract uniformly at random. Given realized location choices for treatment and control households, one could re-define $\mu_Y(\hat{\theta})$ to address both of these issues. We do not pursue this extension, however, as data on location choice under treatment exists only for the Seattle CZ, where Bergman et al. (2023) conducted their experiment, and is not publicly available even there.

points, while the sign of the bias for empirical Bayes varies across CZs, with the bias ranging from -1.24 percentile points to 0.3 percentile points. Hence we see that both the conventional and empirical Bayes estimates exhibit bias in this application. The conditional estimator is median unbiased up to simulation error, while the hybrid estimator is very close to median unbiased. Panel (d) plots the distributions of the median absolute estimation error for the four estimators. The conventional estimator has the largest median absolute estimation error in most CZs, while the empirical Bayes estimator typically has the smallest. The conditional and hybrid estimators are in the middle, with quite similar median absolute estimation errors for this application. Finally, panels (e) and (f) plot the distributions of coverage and median length of confidence intervals. We see that the conventional confidence interval severely under-covers in all 50 CZs. The coverage of empirical Bayes intervals, credible sets for a normal prior, differs widely across CZs, ranging from less than 1% to over 80%.[33] Conditional confidence intervals have coverage equal to 95% up to simulation error in all CZs, while the hybrid intervals have coverage very close to 95% in all cases. Finally, projection intervals have coverage nearly equal to 100% in all CZs. Turning to median length, we see that hybrid intervals are longer than empirical Bayes and conventional confidence intervals (which both have incorrect coverage), but with a median length under 5 percentile points in all CZs, are considerably shorter than conditional and projection intervals.

Empirical Results We next apply the winner's curse corrections directly to the Opportunity Atlas data. As in the simulations we select the top third of census tracts in each CZ based on the conventional estimates. Our parameter of interest is the average mobility across the selected tracts, less the weighted average over the CZ. We report results for conventional, empirical Bayes, and hybrid estimates and intervals, as well as projection intervals in Figure V, while for visibility we defer the results for conditional intervals to Figure 9 in Appendix G. For comparison, we also plot the within-CZ standard deviation of mobility.[34] The conventional estimates range from 7.1 to 18.6 percentile points across CZs (or, dividing the estimate in each CZ by the within-CZ standard deviation, from 1.7 to 2.4 standard deviations). Both the empirical Bayes and hybrid corrections shift the estimates downward, with empirical Bayes estimates ranging from 5.2 to 17.4 percentile points (1.2 to

[33] In Appendix G.6, we show that the coverage of empirical Bayes intervals in a given CZ is correlated with the quality of the normal approximation to the true distribution of economic mobility in that CZ. This highlights the fragility of empirical Bayes corrections for the winner's curse when the normality assumption on the distribution of true effects, μ_t in this example, fails.

[34] We estimate this by the square root of the difference between the sample variance of the tract-level estimates and the average squared standard error.

2.3 standard deviations) and hybrid estimates ranging from 4 to 17.2 percentile points (0.9 to 2.3 standard deviations). Hence, while both the empirical Bayes and hybrid corrections somewhat deflate the estimates, they remain uniformly positive across CZs. Moreover, the effect sizes are large in both percentile point and standard deviation terms. Comparing the empirical Bayes and hybrid approaches, the hybrid estimate is slightly lower on average but the two estimators are not ordered: for instance, the hybrid estimate is smaller than empirical Bayes in Chicago, but larger in New York.

Turning to confidence intervals we see that, as expected given our simulation results, the coverage-maintaining hybrid intervals (with an average length of 3.3 percentile points, or 0.6 standard deviations) are wider than the under-covering empirical Bayes intervals (0.6 percentile points, or 0.1 standard deviations), but considerably shorter than projection intervals (6.2 percentile points, or 1.1 standard deviations). Hybrid and projection intervals exclude zero in all CZs, suggesting that, under the hypothesis of neighborhood effects, there is real scope for selecting higher-mobility neighborhoods based on the Opportunity Atlas, albeit less than the conventional estimates suggest. The results for conditional procedures in Figure 9 of Appendix G are qualitatively similar, but the conditional intervals are much longer on average (with a mean length of 19.1 percentile points, or 3.3 standard deviations). This length is heavily influenced by a small number of long intervals, and the median length is substantially lower (at 10.6 percentile points, or 1.6 standard deviations). Conditional intervals lie above zero in 28 of the 50 CZs, but include zero in the other 22. Hence, if we are satisfied with unconditional coverage we find strong evidence that selected tracts are better than average, while if we demand conditional coverage results are more mixed, and depend on which CZ we consider.

Overall, these results show that accounting for the winner's curse in this setting makes an economically significant difference: the average hybrid estimate (10.27 percentile points) is nearly 20% smaller than the average conventional estimate (12.25 percentile points). That said, even this smaller estimate is economically large. Correcting for the winner's curse also increases our degree of uncertainty, but provided we are satisfied with unconditional inference we are still able to draw highly informative inferences in this setting. Specifically, we conclude that targeting tracts based on estimated opportunity succeeds in selecting higher-opportunity tracts on average. Moreover, even after accounting for uncertainty the effect sizes are large: in all but two CZs the lower bounds of both the hybrid and projection intervals exceed the within-CZ standard deviation of mobility across tracts. The use of unconditional rather than conditional inference is important for this conclusion, since if we

instead consider conditional intervals we are sometimes unable to reject zero. This reflects the fact that conditional inference is highly demanding in this setting due to the enormous number of ways that we may select a third of tracts in a given CZ. In our view there is not an obvious reason to require conditional validity in this application: the question of primary interest is whether targeting tracts based on economic opportunity will succeed in selecting high-opportunity tracts, which is inherently an unconditional question since it concerns the method for selecting tracts rather than the specific set of tracts selected in a particular CZ.

It is useful to compare our results with those of Mogstad et al. (2022), who study the problem of inference on ranks and consider the Opportunity Atlas data for Seattle as an example. They show that if one forms simultaneous confidence sets for differences between individual tracts, one can say very little about which tracts are best. Hence, we can say little about the effect of moving an individual from an arbitrary non-target tract to an arbitrary target tract, and can likewise say little about the average treatment effect of shifting households from one group of tracts to the other if we allow arbitrary location choices within each group of tracts. We consider a complementary exercise, inference on the average mobility over the selected sets of tracts, corresponding to uniformly distributed location choices. For this problem we find strong evidence that selected tracts are, as a group, better than average. These exercises answer different questions, and the more positive result obtained in our case reflects that it is much easier statistically to distinguish average mobility across groups of selected tracts than it is to rank individual tracts.

8 Conclusion

This paper considers a form of the winner's curse that arises when we select one of several options based on noisy estimates. We propose corrected inference procedures that eliminate the winner's curse either conditional on the selection made or unconditionally. Since conditional inference is statistically more demanding and can yield less precise conclusions, we recommend a novel (unconditional) hybrid procedure as a default approach when there is not a clear reason for one to condition. Using data from Cave et al. (1993) and Chetty et al. (2020), we find that our corrected inference procedures can make an economically significant difference, but continue to allow precise inference. For other recent applications of these methods see Banerjee et al. (2022) and Bergeron et al. (2022)

Our results suggest possible directions for future work. While our inference results build on the statistics literatures on selective inference (e.g. Fithian, Sun, and Taylor, 2017) and post-selection inference (e.g. Berk et al., 2013), our hybrid approach is novel

relative to both, and the analysis of McCloskey (2023) shows that hybrid inference may be helpful in other settings considered by these literatures. Similarly, Andrews, Roth, and Pakes (2022) show that a version of our hybrid approach offers a useful tool in moment inequality settings. Finally, inference distortions due to data-driven selection are known to arise in many other contexts, including adaptive experiments (see e.g. Zhang, Janson, and Murphy, 2020). While some inference results are already available for adaptive experiments, the extension of our analysis to cover such settings, and more broadly the impact of the winner's curse on optimal experimental design, is an interesting question for future work.

References

Andrews, I., D. Bowen, T. Kitagawa, and A. McCloskey (2022). Inference for losers. *AEA Papers and Proceedings 112*, 635–42.

Andrews, I. and M. Kasy (2019). Identification of and correction for publication bias. *American Economic Review 109*(8), 2766 – 94.

Andrews, I., T. Kitagawa, and A. McCloskey (2021). Inference after estimation of breaks. *Journal of Econometrics 224*, 39–59.

Andrews, I., J. Roth, and A. Pakes (2022). Inference for linear conditional moment inequalities. *Review of Economic Studies Forthcoming*.

Armstrong, T., M. Kolesar, and M. Plagborg-Moller (2022). Robust empirical Bayes confidence intervals. *Econometrica 90*(6), 2567–2602.

Bachoc, F., D. Preinerstorfer, and L. Steinberger (2020). Uniformly valid confidence intervals post-model-selection. *Annals of Statistics 48*(1), 440–463.

Banerjee, A., A. G. Chandrasekhar, S. D. E. Duflo, J. Floretta, M. O. Jackson, H. Kannan, F. Loza, A. Sankar, A. Schrimpf, and M. Shrestha (2022). Selecting the most effective nudge: Evidence from a large-scale experiment on immunization. Unpublished Manuscript.

Bergeron, A., P. Bessone, J. K. Kabeya, G. Z. Tourek, and J. L. Weigel (2022). Optimal assignment of bureaucrats: Evidence from randomly assigned tax collectors in the DRC. Unpublished Manuscript.

Bergman, P., R. Chetty, S. DeLuca, N. Hendren, L. F. Katz, and C. Palmer (2023). Creating moves to opportunity: Experimental evidence on barriers to neighborhood choice. Unpublished Manuscript.

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao (2013). Valid post-selection inference. *Annals of Statistics 41*(2), 802–831.

Burghardt, J., A. Rangarajan, A. Gordon, and E. Kisker (1992). *Evaluation of the Minority Female Single Parent Demonstration: Volume I.* New York, NY: The Rockefeller Foundation.

Cave, G., H. Bos, F. Doolittle, and C. Toussaint (1993). JOBSTART – Final report on a program for school dropouts. Technical report, Manpower Demonstration Research Corporation (MDRC), New York, NY.

Chetty, R., J. Friedman, N. Hendren, M. R. Jones, and S. R. Porter (2020). The opportunity atlas: Mapping the childhood roots of social mobility. Unpublished Manuscript.

Chetty, R. and N. Hendren (2018a). The impacts of neighborhoods on intergenerational mobility i: Childhood exposure effects. *Quarterly Journal of Economics 133*(3), 1107–1162.

Chetty, R. and N. Hendren (2018b). The impacts of neighborhoods on intergenerational mobility ii: County-level estimates. *Quarterly Journal of Economics 133*(3), 1163–1228.

Dawid, A. P. (1994). Selection paradoxes in bayesian inference. *Institute of Mathematical Statistics Lecture Notes - Monograph Series 24*, 211–220.

Eaton, M. L. (1967). Some optimum properties of ranking procedures. *Annals of Mathematical Statistics 38*(1), 124–137.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association 106*(496), 1602–1614.

Efron, B. and C. Morris (1975). Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association 70*(350), 311–319.

Ferguson, J. P., J. H. Cho, C. Yang, and H. Zhao (2013). Empirical bayes correction for the winner's curse in genetic association studies. *Genetic Epidemiology 37*(1), 60–68.

Fithian, W., D. Sun, and J. Taylor (2017). Optimal inference after model selection. *arXiv*.

Frandsen, B. (2020). A rational approach to inference on multiple parameters. Unpublished Manuscript.

Freyberger, J. and Y. Rai (2018). Uniform confidence bands: Characterization and optimality. *Journal of Econometrics 1*(204), 119–130.

Gu, J. and R. Koenker (2023). Invidious comparisons: Ranking and selection as compound decisions. *Econometrica 91*, 1–41.

Gupta, S. S. and K. J. Miescke (1988). On the problem of finding the largest normal mean under heteroskedasticity. In S. S. Gupta and J. O. Berger (Eds.), *Statistical Decision Theory and Related Topics - IV*, Volume 2, pp. 37–49. Springer Verlag.

Hendren, N. and B. Sprung-Keyser (2020). A unified welfare analysis of government policies. *Quarterly Journal of Economics 135*(3), 1209–1318.

Hirano, K. and J. R. Porter (2009). Asymptotics for statistical treatment rules. *Econometrica 77*, 1683–1701.

Karlan, D. and J. A. List (2007). Does price matter in charitable giving? evidence from a large-scale natural field experiment. *American Economic Review 97*(5), 1774–1793.

Kitagawa, T. and A. Tetenov (2018a). Supplement to "who should be treated? empirical welfare maximization methods for treatment choice". *Econometrica*.

Kitagawa, T. and A. Tetenov (2018b). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica 86*(2), 591–616.

Kivaranovic, D. and H. Leeb (2021). On the length of post-model-selection confidence intervals conditional on polyhedral constraints. *Journal of the American Statistical Association 116*, 845–857.

Kuchibhotla, A. K., L. D. Brown, A. Buja, J. Cai, E. I. George, and L. Zhao (2020). Valid post-selection inference in model-free linear regression. *Annals of Statistics 48*, 2953–2981.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2016). Exact post-selection inference, with application to the LASSO. *Annals of Statistics 44*(3), 907–927.

Lee, M. R. and M. Shen (2018). Winner's curse: Bias estimation for total effects of features in online controlled experiments. In *KDD*.

Lehmann, E. L. (1966). On a theorem of bahadur and goodman. *Annals of Mathematical Statistics 37*(1), 1–6.

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica 72*(4), 1221–1246.

McCloskey, A. (2017). Bonferroni-based size-correction for nonstandard testing problems. *Journal of Econometrics 200*(1), 17–35.

McCloskey, A. (2023). Hybrid confidence intervals for informative uniform asymptotic inference after model selection. *Biometrika Forthcoming*.

Miller, C., J. M. Bos, K. E. Porter, F. M. Tseng, and Y. Abe (2005). The challenge of repeating success in a changing world – Final report on the Center for Employment Training replication sites. Technical report, Manpower Demonstration Research Corporation (MDRC), New York, NY.

Mogstad, M., J. P. Romano, A. Shaikh, and D. Wilhelm (2022). Inference for ranks with applications to mobility across neighborhoods and academic achievement across countries. *Review of Economic Studies Forthcoming*.

Olea, J. M. and M. Plagborg-Moller (2019). Simultaneous confidence bands: Theory, implementation, and an application to svars. *Journal of Applied Econometrics 34*(1), 1–17.

Pfanzagl, J. (1979). On optimal median unbiased estimators in the presence of nuisance parameters. *Annals of Statistics 7*(1), 187–193.

Pfanzagl, J. (1994). *Parametric Statistical Theory*. De Gruyter.

Romano, J. P., A. Shaikh, and M. Wolf (2014). A practical two-step method for testing moment inequalities. *Econometrica 82*(5), 1979–2002.

Romano, J. P. and M. Wolf (2005). Stepwise multiple testing as formalized data snooping. *Econometrica 73*(4), 1237–1282.

Tian, X. and J. Taylor (2018). Selective inference with a randomized response. *Annals of Statistics 46*(2), 679–710.

US Department of Housing and Urban Development (2018). Picture of subsidized households.

Xu, L., R. V. Craiu, and L. Sun (2011). Bayesian methods to overcome the winner's curse in genetic studies. *Annals of Applied Statistics 5*(201-231).

Young, G. A. and R. L. Smith (2005). *Essentials of Statistical Inference*. Cambridge University Press.

Zhang, K., L. Janson, and S. Murphy (2020). Inference for batched bandits. In *Conference on Neural Information Processing Systems*.

Zhong, H. and R. Prentice (2009). Correcting "winner's curse" in odds ratios from genomewide association findings for major complex human diseases. *Genetic Epidemiology 34*(1), 78–91.

Zrnic, T. and W. Fithian (2022). Locally simultaneous inference. Unpublished Manuscript.

Table I

Site	n_T	n_C	ATE Estimate	S.E.	Control Mean
Atlanta Job Corps	33	36	\$2093	\$2288.40	\$10112
CET/San Jose	84	83	\$6547	\$1496.17	\$12362
Chicago Commons	40	35	-\$1417	\$2168.21	\$11726
Connelley (Pittsburgh)	91	93	\$785	\$1681.92	\$6685
East LA Skills Center	50	56	\$1343	\$1735.51	\$13158
EGOS (Denver)	103	95	\$401	\$1329.05	\$10690
Phoenix Job Corps	70	64	-\$1325	\$1598.03	\$8198
SET/Corpus Christi	125	122	\$485	\$971.05	\$7992
El Centro (Dallas)	93	86	\$336	\$1523.33	\$11057
LA Job Corps	116	115	-\$121	\$1409.79	\$12757
Allentown (Buffalo)	71	64	\$904	\$1814.10	\$6577
BSA (New York City)	60	57	\$1424	\$1768.44	\$10499
CREC (Hartford)	52	47	-\$1370	\$1860.45	\$11124

Results for the 13 sites in the JOBSTART demonstration. The first column indicates the site, the second (n_T) reports the number of treated individuals at that site, the third (n_C) the number of control individuals, the fourth the estimated average treatment effect on cumulative earnings in years three and four following random assignment (months 25-48), the fifth an imputed standard error, and the sixth the mean cumulative earnings for the control group in years three and four. We use imputed standard errors because Cave et al. (1993) report statistical significance at the 1%, 5%, and 10% levels but not standard errors, t-statistics or precise p-values. See Appendix A for the (restrictive) assumptions that underlie this imputation.

Table II

Method	Point Estimate	CI
Conventional	\$6547	(\$3615, \$9479)
Conditional	\$6544	(\$3485, \$9478)
Projection	\$6547	(\$2232, \$10862)
Hybrid	\$6545	(\$3420, \$9538)

Point estimates and 95% confidence intervals for the ATE of the CET/San Jose program using the treatment-control differences and imputed standard errors from Table I.

Table III

Method	Forecast Interval	p-value for Equality	S.E. Scaling
	All 12 Sites		
Conditional	($2609,$10449)	2×10^{-4}	1.84
Hybrid	($2531,$10529)	2×10^{-4}	1.82
	High Fidelity Sites		
Conditional	($634,$12436)	8×10^{-3}	1.76
Hybrid	($520,$12556)	7×10^{-3}	1.72

Forecasting results for replication of the CET program using imputed standard errors for both the replication study across all 12 replication sites and the four high fidelity sites. The first column indicates whether the conditional or hybrid approach was used for forecasting. The second column reports 95% forecast intervals for the estimated ATE. The third column reports p-values for testing equality of the ATEs in the original and replication studies. The fourth column reports the amount by which the imputed standard errors for the Cave et al. (1993) estimates must be scaled for a test of the ATEs in the original and replication studies to fail to reject at the 5% significance level.

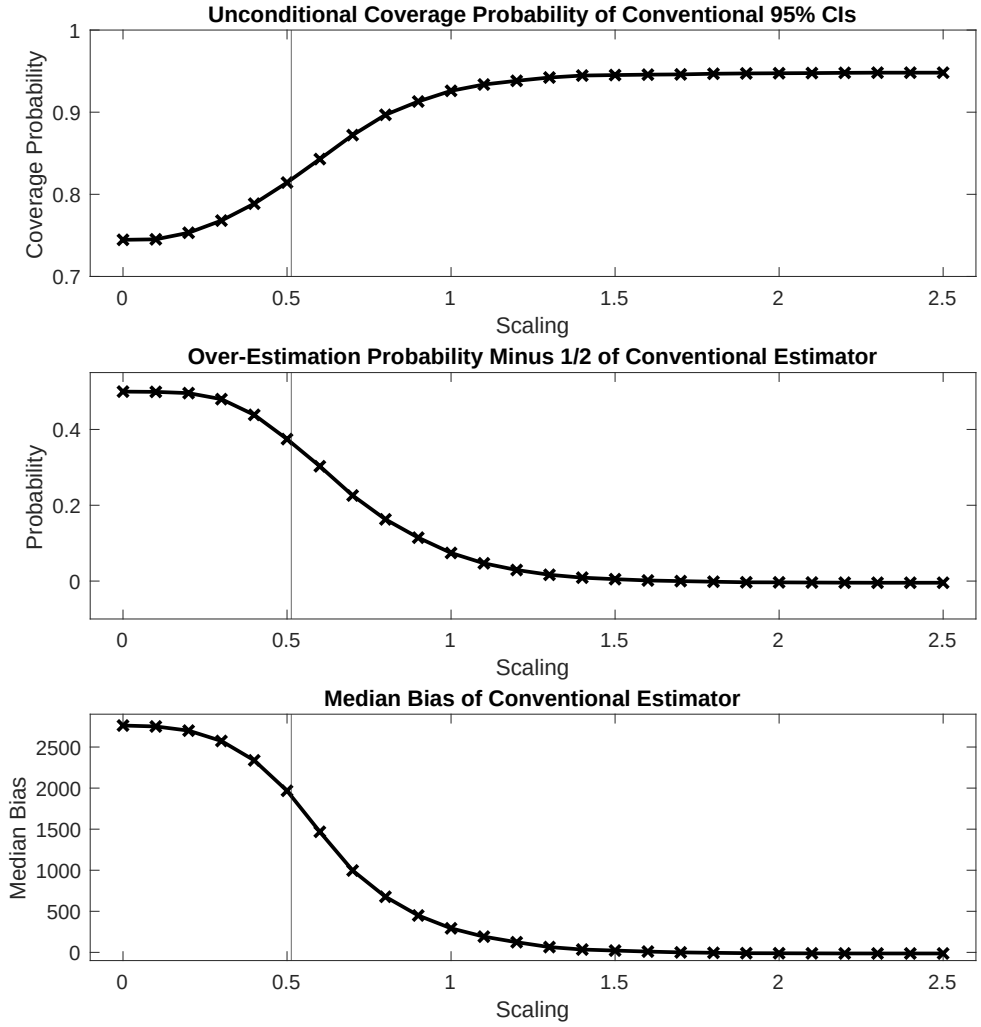

Figure I: Coverage, over-estimation probability, and median bias in dollars for the ATE on cumulative earnings in years three and four in simulations calibration to JOBSTART data, where $X \sim N(s \cdot \hat{\mu}_X, \Sigma)$ for $\hat{\mu}_X$ the JOBSTART point estimates and Σ the diagonal matrix with the squared JOBSTART standard errors on the diagonal. The horizontal axis varies the scaling factor s, and our preferred scaling s^* is marked with a vertical line.

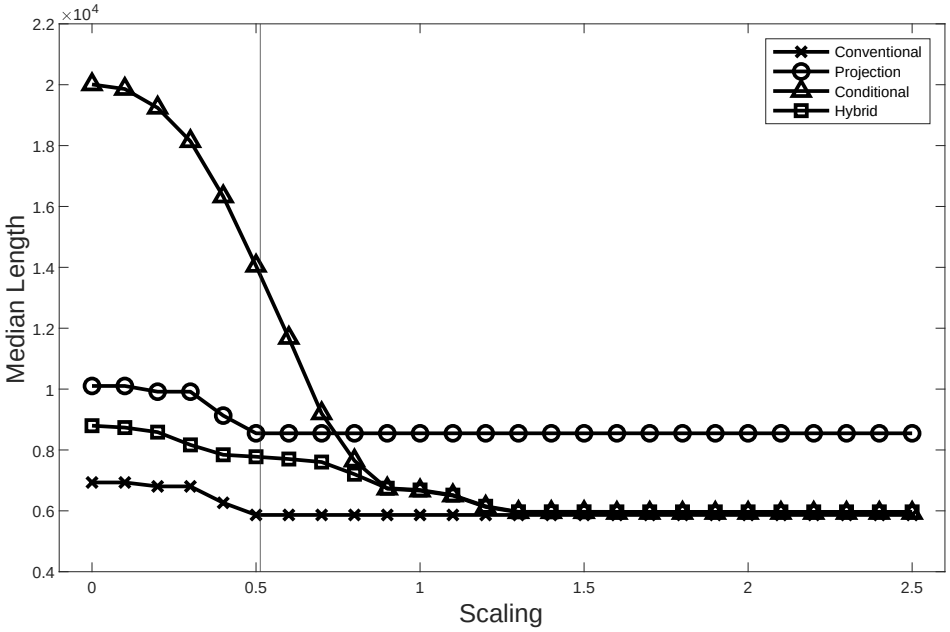

Figure II: Median length for 95% confidence intervals in simulations calibrated to the results of the JOBSTART demonstration, where $X \sim N(s \cdot \hat{\mu}_X, \Sigma)$ for $\hat{\mu}_X$ the JOBSTART point estimates and Σ the diagonal matrix with the squared JOBSTART standard errors on the diagonal. The horizontal axis varies the scaling factor s, and our preferred scaling s^* is marked with a vertical line.

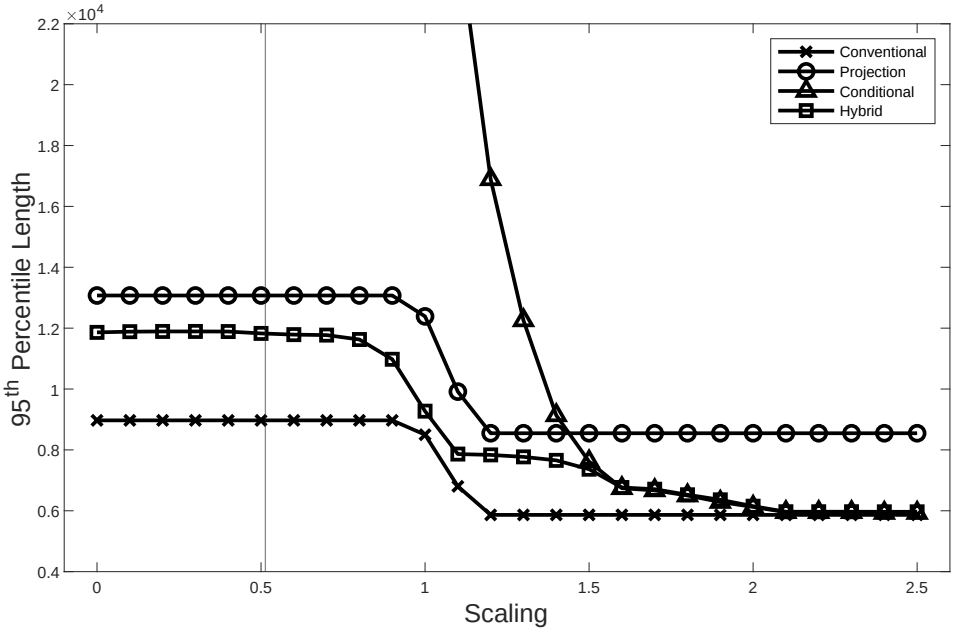

Figure III: 95th percentile length for 95% confidence intervals in simulations calibrated to the results of the JOBSTART demonstration, where $X \sim N(s \cdot \hat{\mu}_X, \Sigma)$ for $\hat{\mu}_X$ the JOBSTART point estimates and Σ the diagonal matrix with the squared JOBSTART standard errors on the diagonal. The horizontal axis varies the scaling factor s, and our preferred scaling s^* is marked with a vertical line.

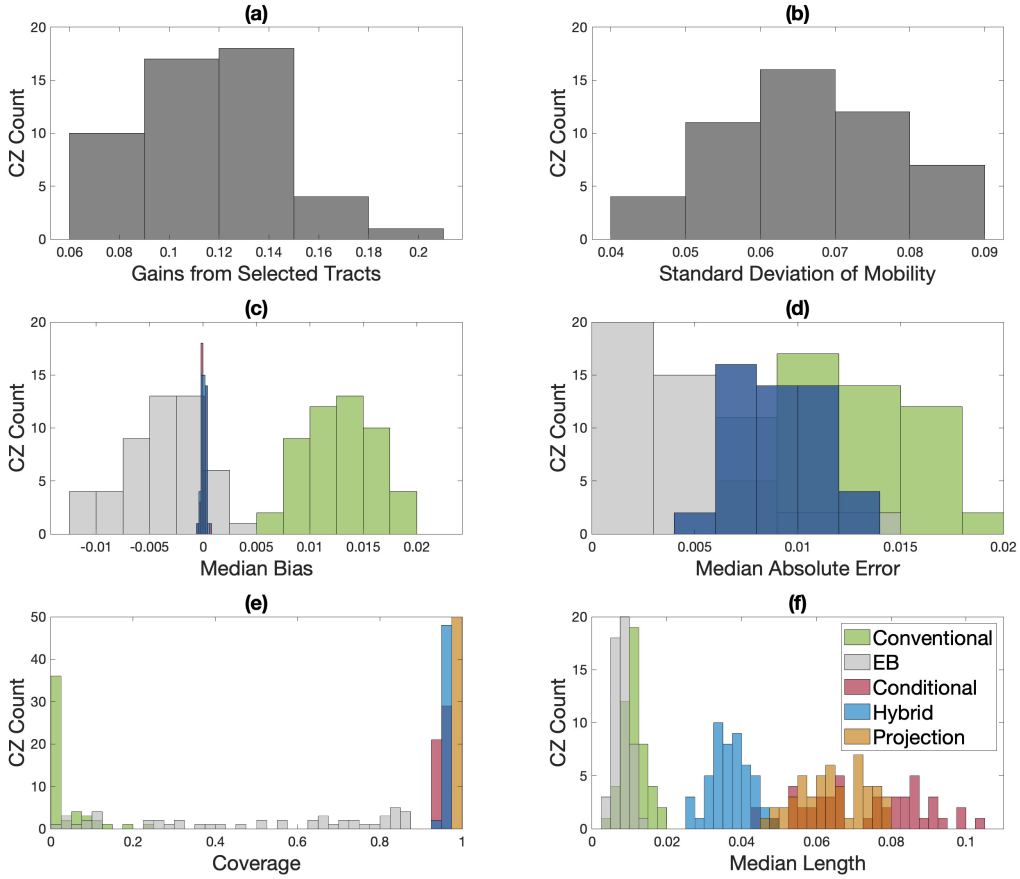

Figure IV: Simulation results from calibration to the Chetty et al. (2020) Opportunity Atlas. Panel (a) shows the distribution of average improvement in economic mobility in selected tracts, relative to the within-CZ weighted average, across the 50 largest CZs. A coefficient of 0.1 implies that the target tracts are associated with a 10 percentile point higher average household income, in adulthood, relative to the weighted average across the CZ. Panel (b) shows the distribution of within-CZ standard deviation of mobility. Panel (c) shows the distributions of median biases of different estimators across the 50 CZs. Panel (d) plots the distributions of median absolute error across the same CZs. Note that the results for the conditional and hybrid fully overlap in this case. Panel (e) shows the distribution of coverage of confidence intervals across the 50 largest CZs, while panel (f) does the same for their median lengths. All quantities reported are unconditional, and so aggregate across values of $\hat{\theta}$.

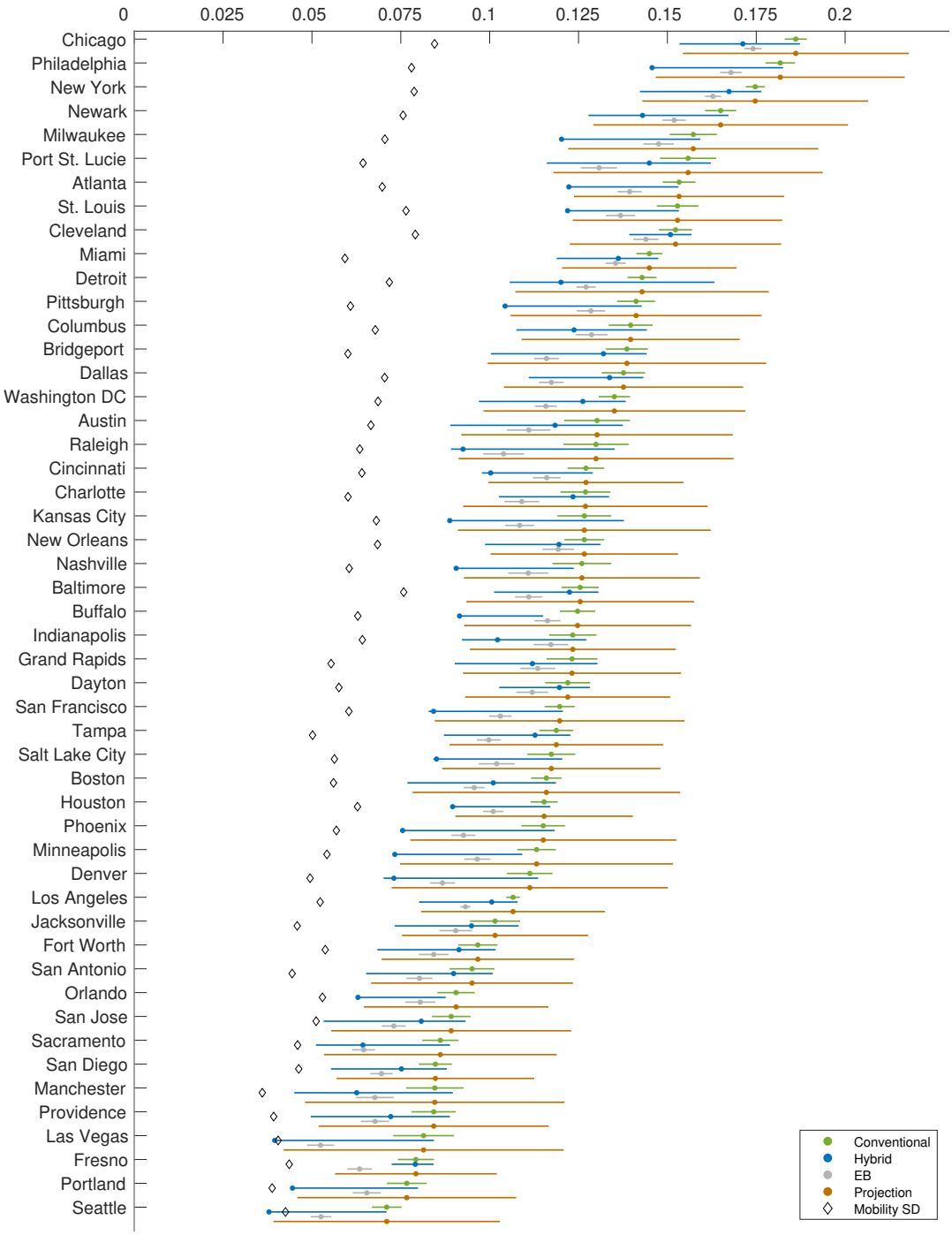

Figure V: Estimates and confidence intervals for average economic mobility for selected census tracts based on the Chetty et al. (2020) Opportunity Atlas, relative to the within-CZ average, weighted by number of voucher recipient households with children. CZs are ordered by the magnitude of the conventional estimate. A coefficient of 0.1 implies that the target tracts are associated with a 10 percentile point higher average household income in adulthood, for children growing up in households at the 25th percentile of the income distribution, relative to the weighted average across the CZ. Diamonds plot the estimated standard deviation of mobility across all tracts in each CZ.