

Inference on Winners*

Isaiah Andrews[†] Toru Kitagawa[‡] Adam McCloskey[§]

June 27, 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, MIT, iandrews@mit.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

¹The 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

²While 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.

³While 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.

⁴Cave 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 than \$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.

⁵Miller 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, $\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} X(\theta)$.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

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

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

⁶Since 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).

⁷This is distinct from the problem of inference on the effect of the *true* best site, $\mu_X(\theta^*)$ for $\theta^* \in \operatorname{argmax}_{\theta \in \Theta} \mu_X(\theta)$. 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 $\mu_X(\hat{\theta})$.

⁸It 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 $\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} X(\theta)$, 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 ± 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

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

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

⁹Recall 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 $CI_{ET} = [\hat{\mu}_{\alpha/2}, \hat{\mu}_{1-\alpha/2}]$ 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:

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

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

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

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

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

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,

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

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

¹¹In 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 $CI_{ET}^H = \left[\hat{\mu}_{\frac{\alpha-\beta}{2(1-\beta)}}^H, \hat{\mu}_{1-\frac{\alpha-\beta}{2(1-\beta)}}^H \right]$. See Section 5.2 for details on this adjustment. By construction, hybrid intervals are never longer than the level $1-\beta$ projection interval CI_P^β . 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 $\hat{\mu}_{\frac{1}{2}}^H$ and hybrid confidence interval CI_{ET}^H . 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

¹²One 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.

¹³Intuitively, 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 $Pr_\pi\{\mu_X(\hat{\theta})\in CR|\hat{\theta}\}=1-\alpha$ 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 $X=(X(\theta_1),\dots,X(\theta_{|\Theta|}))'$ and $Y=(Y(\theta_1),\dots,Y(\theta_{|\Theta|}))'$ 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

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

¹⁴Indeed, 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 $\tilde{\Sigma}_n$ 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 ,

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

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

¹⁵Manski (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 $\Sigma_X = \text{Var}(X)$ 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.

¹⁶We 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 $X = (\Sigma_X^{*-1} + \Omega^{-1})^{-1}(\Sigma_X^{*-1}X^* + \Omega^{-1}\eta)$ 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 k th-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

¹⁷Note 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

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

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

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

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,

¹⁸Indeed, 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

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

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 $\hat{\theta}$:

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

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

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

¹⁹This 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 $\theta^* = \operatorname{argmax}_{\theta \in \Theta} \mu_X(\theta)$.

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 :

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

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

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

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

To account for the effect of conditioning, let

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

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 $\{X \in \mathcal{X}(\tilde{\theta})\}$, and the conditional distribution of $Y(\tilde{\theta})$ given $\{\hat{\theta} = \tilde{\theta}, Z_{\tilde{\theta}} = z\}$ is a $N(\mu_Y(\tilde{\theta}), \Sigma_Y(\tilde{\theta}))$ distribution truncated to the set

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

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 $\Sigma_{XY}(\tilde{\theta}) = \text{Cov}(X(\tilde{\theta}), Y(\tilde{\theta}))$. 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 $N(\mu_Y(\tilde{\theta}), \Sigma_Y(\tilde{\theta}))$ 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 $CI_{ET} = [\hat{\mu}_{\alpha/2}, \hat{\mu}_{1-\alpha/2}]$ 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 $Pr_\mu \{ \hat{\theta} = \tilde{\theta} \}$ 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 $Pr_{\mu_m} \{ \hat{\theta} = \tilde{\theta} \} \rightarrow 1$ as $m \rightarrow \infty$. Then under μ_m , $CI_{ET} \rightarrow_p CI_N$ and $\hat{\mu}_{\frac{1}{2}} \rightarrow_p Y(\tilde{\theta})$ both conditional on $\hat{\theta} = \tilde{\theta}$ and unconditionally, where for confidence intervals \rightarrow_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 $\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\}$. 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 $1-\alpha$,

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

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} , $FI = \{Y_2(\hat{\theta}) : 0 \in CI_{ET}^{1-2}\}$. 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).

²²An 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 $\theta, \tilde{\theta} \in \Theta$ 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

²³Similar 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 .

²⁴The 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 $F_{TN}^H(Y(\hat{\theta});\mu,\hat{\theta},Z_{\hat{\theta}})=1-\alpha$ 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 $\mu_Y(\hat{\theta})\in CI_P^\beta$:

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

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.

²⁵Zrnica 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 $\hat{\mu}_{\frac{1}{2}}^H$ (measured as the deviation of the overestimation probability from $1/2$) is bounded above by $\beta/2$. On the other hand, since $\hat{\mu}_{\frac{1}{2}}^H \in CI_P^\beta$ we have $|\hat{\mu}_{\frac{1}{2}}^H - Y(\hat{\theta})| \leq c_\beta \sqrt{\Sigma_Y(\tilde{\theta})}$, so the difference between $\hat{\mu}_{\frac{1}{2}}^H$ and the conventional estimator $Y(\hat{\theta})$ is bounded above by half the width of CI_P^β .

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

Proposition 6

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

$$Pr_\mu\left\{\mu_Y(\hat{\theta}) \in CI_{ET}^H \mid \mu_Y(\hat{\theta}) \in CI_P^\beta\right\} = \frac{1-\alpha}{1-\beta} \text{ for all } \mu.$$

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

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

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_P^β . 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 $Pr_\mu\{\hat{\theta} = \tilde{\theta}\} \rightarrow 1$, 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 $\hat{\theta}_n = \operatorname{argmax}_{\theta \in \Theta} \tilde{X}_n(\theta)$. 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

²⁶Romano, 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 $\mu_Y(\theta) \in CI_P^\beta$, 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 ,

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

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 $\tilde{\Sigma}_n$ such that $n \cdot \tilde{\Sigma}_n$ 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, $\tilde{\Sigma}_n$ consists of four copies of a diagonal matrix $\tilde{\Sigma}_{X,n}$, where the diagonal entry of $\tilde{\Sigma}_{X,n}$ 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 $\tilde{\Sigma}_n$ 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 $\tilde{\Sigma}_n$ in place of (X, Y) and Σ in all expressions.

Alternatively, to implement our approach directly, one should compute

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

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

$$\begin{aligned} \mathcal{L}(\hat{\theta}_n, \tilde{Z}_{\hat{\theta},n}) &= \max_{\theta \in \Theta: \tilde{\Sigma}_{XY,n}(\hat{\theta}_n) > \tilde{\Sigma}_{XY,n}(\hat{\theta}_n, \theta)} \frac{\tilde{\Sigma}_{Y,n}(\hat{\theta}_n) \left(\tilde{Z}_{\hat{\theta},n}(\theta) - \tilde{Z}_{\hat{\theta},n}(\hat{\theta}_n) \right)}{\tilde{\Sigma}_{XY,n}(\hat{\theta}_n) - \tilde{\Sigma}_{XY,n}(\hat{\theta}_n, \theta)}, \\ \mathcal{U}(\hat{\theta}_n, \tilde{Z}_{\hat{\theta},n}) &= \min_{\theta \in \Theta: \tilde{\Sigma}_{XY,n}(\hat{\theta}_n) < \tilde{\Sigma}_{XY,n}(\hat{\theta}_n, \theta)} \frac{\tilde{\Sigma}_{Y,n}(\hat{\theta}_n) \left(\tilde{Z}_{\hat{\theta},n}(\theta) - \tilde{Z}_{\hat{\theta},n}(\hat{\theta}_n) \right)}{\tilde{\Sigma}_{XY,n}(\hat{\theta}_n) - \tilde{\Sigma}_{XY,n}(\hat{\theta}_n, \theta)}. \end{aligned}$$

For $F_{TN}(y; \mu_Y(\tilde{\theta}), \tilde{Z}_{\tilde{\theta},n})$ the distribution function for a $N(\mu_Y(\tilde{\theta}), \tilde{\Sigma}_{Y,n}(\tilde{\theta}))$ 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 $CI_{ET,n} = [\hat{\mu}_{\frac{\alpha}{2},n}, \hat{\mu}_{1-\frac{\alpha}{2},n}]$.

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 ξ_1, \dots, ξ_S where for each s , $\xi_s \sim N(0, \tilde{\Sigma}_{Y,n})$. For each S compute the maximum absolute studentized deviation $\max_{\theta \in \Theta} |\xi_s(\theta)| / \sqrt{\tilde{\Sigma}_{Y,n}(\theta)}$, and define \hat{c}_β as the $1 - \beta$ quantile of these maximum absolute deviations across the simulation draws. Define

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

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

$$CI_{ET,n}^H = \left[\hat{\mu}_{\frac{\alpha-\beta}{2(1-\beta)},n}^H, \hat{\mu}_{1-\frac{\alpha-\beta}{2(1-\beta)},n}^H \right].$$

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 $n \rightarrow \infty$ (i) the convergence in distribution (13) holds uniformly over $P \in \mathcal{P}_n$, where the variance Σ may depend on P , (ii) $n \cdot \tilde{\Sigma}$ 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 $n \rightarrow \infty$, 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-recipient 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

²⁷We 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.

²⁸We 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.

²⁹Appendix H reports additional simulation and empirical results, based on an experiment by Karlan and List (2007) studying charitable giving, where 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.

³⁰We 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, $\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} X(\theta)$. 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 $Y(\theta) = \frac{1}{|\theta|} \sum_{t \in \theta} \hat{\mu}_t - \frac{\sum_{t \in \mathcal{T}} c_t \hat{\mu}_t}{\sum_{t \in \mathcal{T}} c_t}$.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, $E[\mu_Y(\hat{\theta})]$, across the 50 CZs considered. Targeted tracts are associated with higher than-average mobility (that is, $E[\mu_Y(\hat{\theta})] > 0$) 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.

³¹Specifically, 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

³³In 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.

³⁴We 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.

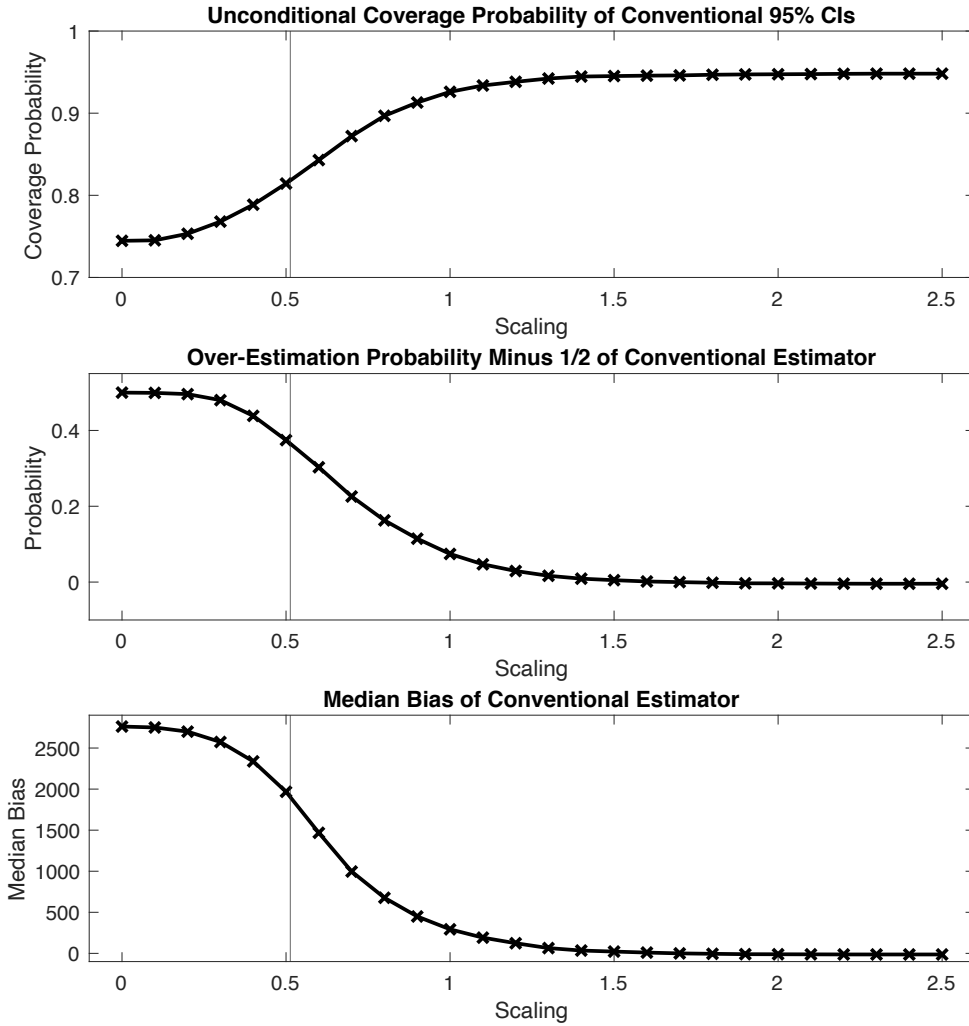


Figure I: Coverage, over-estimation probability, and median bias in dollars for the ATE on cumulative earnings in years three and four in simulations calibration to JOBSTART data, where $X \sim N(s \cdot \hat{\mu}_X, \Sigma)$ for $\hat{\mu}_X$ the JOBSTART point estimates and Σ the diagonal matrix with 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.

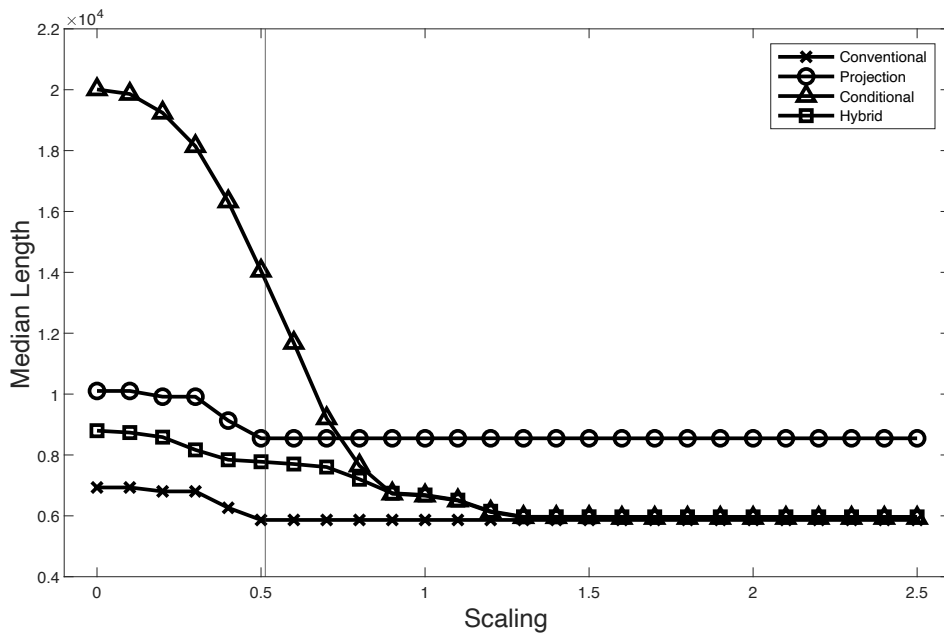


Figure II: Median length for 95% confidence intervals in simulations calibrated to the results of the JOBSTART demonstration, where $X \sim N(s \cdot \hat{\mu}_X, \Sigma)$ for $\hat{\mu}_X$ the JOBSTART point estimates and Σ the diagonal matrix with the squared JOBSTART standard errors on the diagonal.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

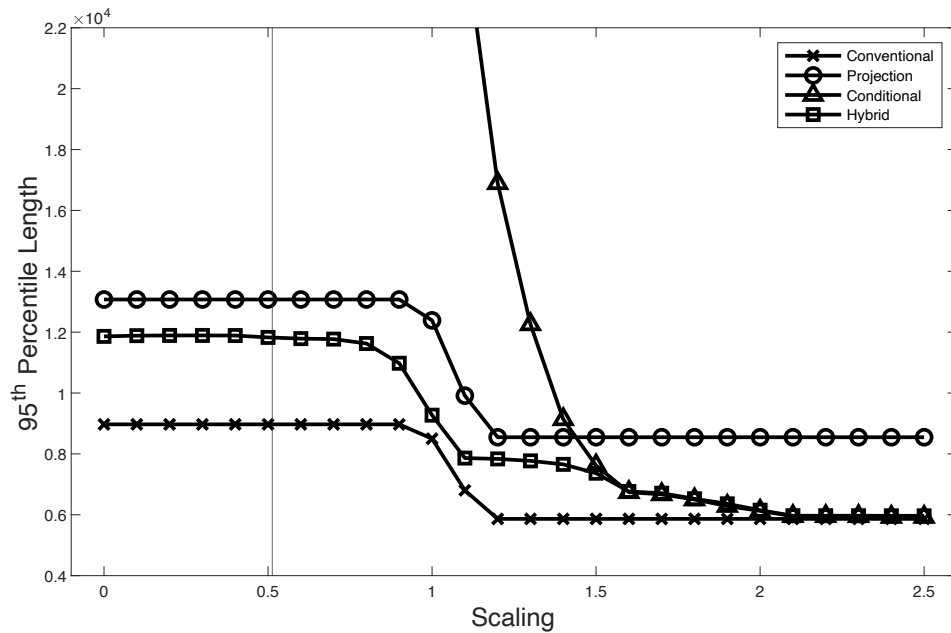


Figure III: 95th percentile length for 95% confidence intervals in simulations calibrated to the results of the JOBSTART demonstration, where $X \sim N(s \cdot \hat{\mu}_X, \Sigma)$ for $\hat{\mu}_X$ the JOBSTART point estimates and Σ the diagonal matrix with the squared JOBSTART standard errors on the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

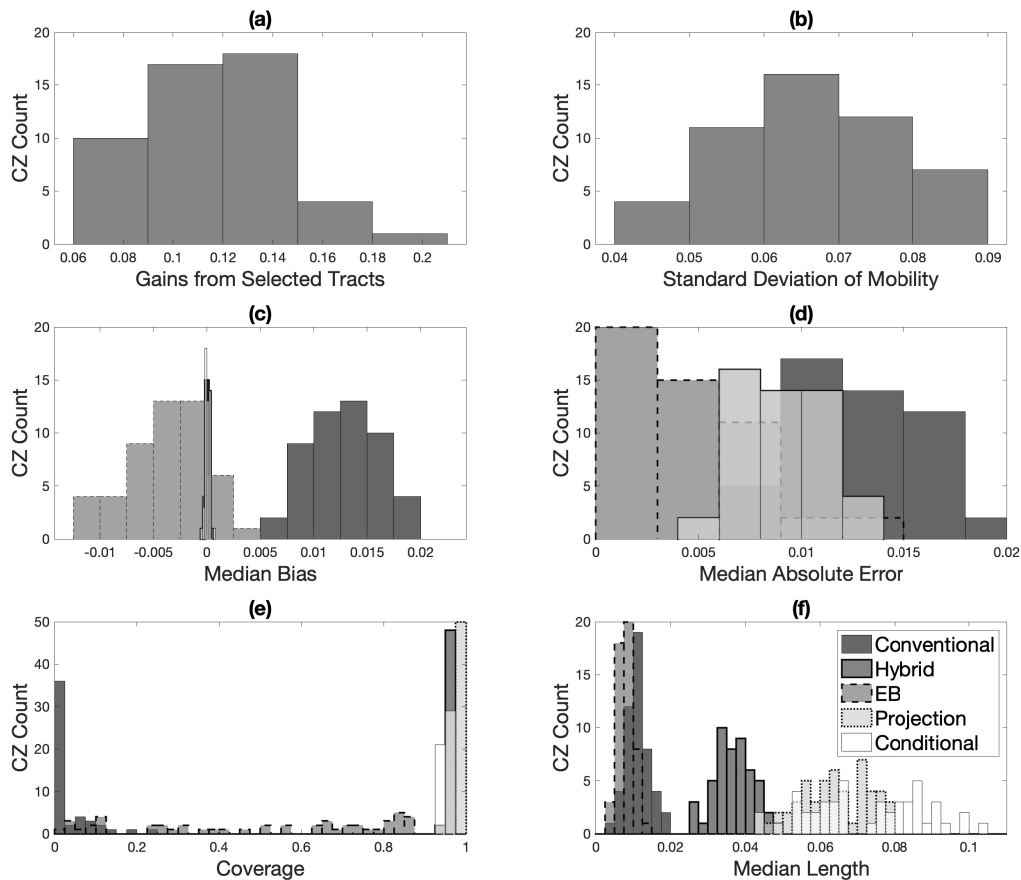


Figure IV: Simulation results from calibration to the Chetty et al. (2020) Opportunity Atlas. Panel (a) shows the distribution of average improvement in economic mobility in selected tracts, relative to the within-CZ weighted average, across the 50 largest CZs. A coefficient of 0.1 implies that the target 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 $\hat{\theta}$.

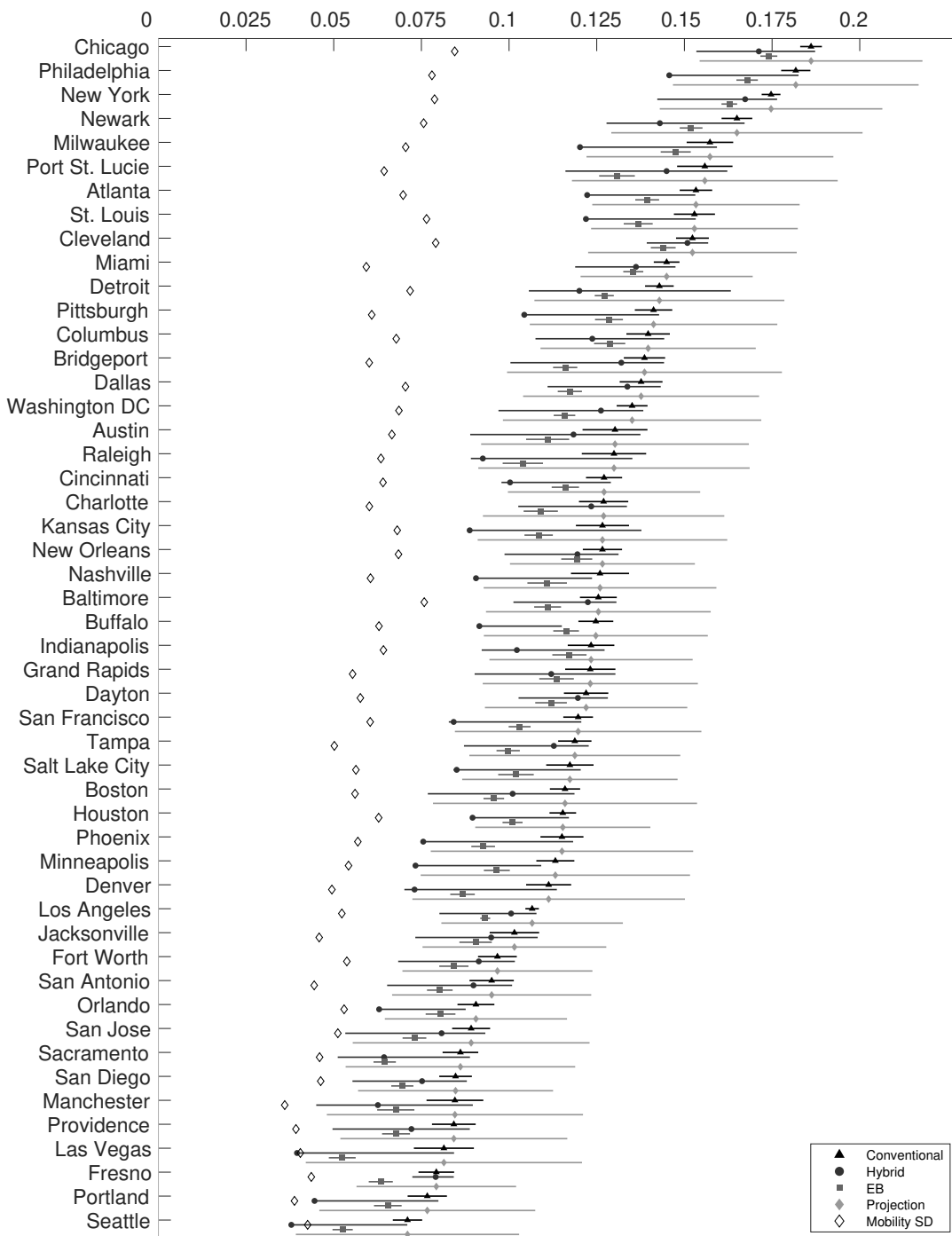


Figure V: Estimates and confidence intervals for average economic mobility for selected census tracts based on the Chetty et al. (2020) Opportunity Atlas, relative to the within-CZ average, weighted by number of voucher recipient households with children. CZs are ordered by the magnitude of the conventional 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.

Online Appendix for
Inference on Winners

Isaiah Andrews

Toru Kitagawa

Adam McCloskey

June 27, 2023

This online appendix contains proofs and additional results for the paper “Inference on Winners.” Appendix A details the assumptions used to impute standard errors for the JOBSTART estimates reported in Cave et al. (1993), and reports additional results for our simulations calibrated to the JOBSTART data. Appendix B shows that our unconditional and conditional coverage requirements arise as necessary conditions for minimax decision rules in two-step decision problems. Appendix C generalizes the conditional inference results discussed in the main text, extending these results to allow additional conditioning variables and unbiased confidence intervals. Appendix D proves our results for the finite-sample normal model. Appendix E provides further details on how our conditional and unconditional inference procedures can be adapted to provide forecast intervals. Appendix F states and proves the uniform asymptotic results referenced in the main text. Appendix G provides additional results and discussion to complement the application in Section 7 of the main text. Finally, Appendix H provides empirical results for an additional empirical example, based on Karlan and List (2007).

Contents

A	Details and Additional Results for JOBSTART Application	3
A.1	Standard Error Imputation	3
A.2	Conditional Coverage Results	8
A.3	Split-Sample Results	8
B	A Decision-Theoretic Model of Inference After Selection	9
C	Conditional Inference	12
C.1	Additional Conditioning Events	12
C.2	Unbiased Confidence Intervals	14
C.3	Behavior When $Pr_{\mu}\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\}$ is Large	16
D	Proofs	16
D.1	Proofs for Results in Section C	16
D.2	Proofs for Results in Main Text	24
E	Forecast Intervals	26
F	Uniform Asymptotic Validity	28
F.1	Uniform Asymptotic Validity	30
F.2	Auxiliary Lemmas	33
F.3	Proofs for Auxiliary Lemmas	35
F.4	Proofs for Uniformity Results	41
G	Additional Materials for Neighborhoods Application	46
G.1	Simulation Design and Target Parameters	46
G.2	Fixed-Length Projection Intervals	47
G.3	Empirical Bayes and Winner’s Curse	48
G.4	Additional Figure for Movers Application	49
G.5	Evaluating the Normality Assumption in Empirical Bayes	49
H	Application: Charitable Giving	54

A Details and Additional Results for JOBSTART Application

This appendix provides additional details and results for the JOBSTART application discussed in Section 2 of the main text.

A.1 Standard Error Imputation

This appendix shows how we obtain the standard errors for the average treatment effects reported in Table I of the main text based upon estimates and standard errors in Tables 5.13 and B.4 of Cave et al. (1993).

The results in Table 5.13 of Cave et al. (1993) are based on what we term “the long regression”. Letting i denote the individual, $t=1$ denote the third year (months 25–36) and $t=2$ the fourth year (months 37–48) after individual assignment, we suppose that the long regression has the form

$$Y_{it} = \delta_0 + \gamma_{s_r} D_i 1_{s_r}\{S_i\} + \sum_{s \in \mathcal{S} \setminus s_r} (\beta_s(1 - D_i) + \gamma_s D_i) 1_s\{S_i\} + X_i' \delta + \varepsilon_{it}, \quad (\text{A.1})$$

where Y_{it} is the outcome of interest (annual earnings in year t), D_i is an indicator of treatment, $1_s\{S_i\}$ is an indicator for site s , X_i is a vector of time-invariant binary characteristics, and ε_{it} is the unobserved residual for individual i in period t . We base our specification of the long regression upon information in and around Table 5.13 of Cave et al. (1993) since the regression equation is not formally stated. We equate the reference site, which we denote by $s_r \in \mathcal{S}$ for \mathcal{S} the set of sites, with SER/Jobs for Progress in Corpus Christi, TX in accordance with Table B.4 of Cave et al. (1993). An (implicit) maintained assumption in our analysis of this example is that all regression coefficients are time-invariant.

Table 5.13 in Cave et al. (1993) reports estimates for the average *cumulative* treatment effect on outcomes over months 25–48 after the treatment. Using the form of the long regression, we can write these cumulative outcomes as

$$Y_{i1} + Y_{i2} = 2\delta_0 + 2\gamma_{s_r} D_i 1_{s_r}\{S_i\} + \sum_{s \in \mathcal{S} \setminus s_r} (2\beta_s(1 - D_i) + 2\gamma_s D_i) 1_s\{S_i\} + 2X_i' \delta + \varepsilon_{i1} + \varepsilon_{i2}. \quad (\text{A.2})$$

Denoting estimates of the parameters of (A.1) in the usual way, Table 5.13 thus reports,

for each binary d and $s \in \mathcal{S}$,

$$\hat{E}[Y_{i1} + Y_{i2} | D_i = d, S_i = s, X_i = \bar{x}_{ds}] = \begin{cases} 2\hat{\delta}_0 + 2\hat{\gamma}_{s_r}d + 2\bar{x}'_{ds_r}\hat{\delta}, & \text{for } s = s_r, \\ 2\hat{\delta}_0 + 2\hat{\beta}_s(1-d) + 2\hat{\gamma}_sd + 2\bar{x}'_{ds}\hat{\delta}, & \text{for } s \neq s_r, \end{cases} \quad (\text{A.3})$$

where \bar{x}_{ds} is a group average.

To give a causal interpretation to the results from the long regression (A.1), we assume that potential outcomes are also linear, with the potential outcome for person i in period t under treatment status d taking the form

$$Y_{it}(d) = \delta_0 + \gamma_{s_r}d1_{s_r}\{S_i\} + \sum_{s \in \mathcal{S} \setminus s_r} (\beta_s(1-d) + \gamma_s d)1_s\{S_i\} + X_i'\delta + \varepsilon_{it}(d),$$

where $E[\varepsilon_{it}(d)] = 0$. Note that this model implies that conditional average treatment effects on earnings given (S_i, X_i) depend on S_i but not on X_i , and that the site-specific average treatment effect on cumulative earnings at site $s \in \mathcal{S}$ is then

$$\begin{aligned} \tau_s &\equiv E[Y_{i1}(1) + Y_{i2}(1) | S_i = s] - E[Y_{i1}(0) + Y_{i2}(0) | S_i = s] \\ &= E[Y_{i1} + Y_{i2} | D_i = 1, S_i = s, X_i = x] - E[Y_{i1} + Y_{i2} | D_i = 0, S_i = s, X_i = x], \\ &= \begin{cases} 2\gamma_{s_r} & \text{for } s = s_r \\ 2\gamma_s - 2\beta_s & \text{for } s \neq s_r, \end{cases} \end{aligned}$$

where the second line follows from randomization of D_i and homogeneity of the conditional average treatment effects in X_i . The estimates for τ_s reported in Table 5.13 of Cave et al. (1993) thus correspond to

$$\hat{\tau}_s = \begin{cases} 2\hat{\gamma}_{s_r} & \text{for } s = s_r \\ 2\hat{\gamma}_s - 2\hat{\beta}_s & \text{for } s \neq s_r, \end{cases} \quad (\text{A.4})$$

which measures the difference between the average outcomes in the treated and control groups at a given site. Cave et al. (1993) report neither confidence intervals nor p-values for these estimates; we must, therefore, impute them.

We base this imputation on the results reported in Table B.4 of Cave et al. (1993).

These results are based on what we call “the short regression”,

$$Y_{i2} = \eta_0 + \psi D_i + \sum_{s \in \mathcal{S} \setminus s_r} \lambda_s 1_s \{S_i\} + X_i' \eta + \nu_{i2}, \quad (\text{A.5})$$

where the definition of the regressors is the same as in (A.1). Note that, in contrast with Table 5.13, the dependent variable of the regression reported in Table B.4 is earnings over months 37–48, rather than cumulative earnings over months 25–48. Moreover, since the functional form of the short regression differs from (A.1), ν_{i2} and ε_{i2} differ.

Table B.4 of Cave et al. (1993) reports both estimates and standard errors for the parameters in (A.5). Our goal is to infer standard errors for $\hat{\tau}_s$ based upon the standard error estimates obtained for the linear regression coefficients of (A.5). In order to do so, we rely on several assumptions.

Assumption 1 (Independence of binary characteristics)

X_i is independent of (D_i, S_i) .

Although useful for simplifying our analysis, Assumption 1 can be rejected using the data reported in Cave et al. (1993). Specifically, the ratio of men to women varies across sites in Table 2.1 of Cave et al. (1993), and this variation is more than we would expect due solely to sampling variability. To accommodate this and other potential failures of our assumptions, we examine the sensitivity of our empirical results in this application to changes in the standard errors.

Assumption 2 (Uniform random assignment at sites)

$\Pr\{D_i = 1 | S_i = s\} = 1/2$.

Imposing Assumption 2 means we do not have to worry about the relative sizes of the treatment and controls groups at each site, or their influence on the variance of any obtained estimates.¹ Although there are slight differences in the ratio of treated to control individuals across sites, these differences are small.

We now rewrite (A.1) using Assumptions 1 and 2. Let $p_s = \mathbb{E}[1_s \{S_i\}]$ for all $s \in \mathcal{S}$, and rewrite (A.1) for $t=2$ as

$$Y_{i2} = \delta_0 + \gamma_{sr} D_i + \sum_{s \in \mathcal{S} \setminus s_r} (\gamma_s - \beta_s - \gamma_{sr}) D_i 1_s \{S_i\} + \sum_{s \in \mathcal{S} \setminus s_r} \beta_s 1_s \{S_i\} + X_i' \delta + \varepsilon_{i2}$$

¹Differences in the number of individuals at each site are already captured by the standard errors in Table B.4.

$$= \eta_0 + \psi D_i + \sum_{s \in \mathcal{S} \setminus s_r} \frac{1}{2} [\gamma_s + \beta_s - \gamma_{s_r}] 1_s \{S_i\} + X_i' \delta + \nu_{i2}, \quad (\text{A.6})$$

where

$$\nu_{i2} = \sum_{s \in \mathcal{S} \setminus s_r} (\gamma_s - \beta_s - \gamma_{s_r}) (D_i - 1/2) (1_s \{S_i\} - p_s) + \varepsilon_{i2}. \quad (\text{A.7})$$

Given that ν_{i2} is mean-zero, and is uncorrelated with any of the regressors in (A.6) under Assumptions 1 and 2, the projection coefficients in (A.6) coincide with the coefficients of (A.5). We can hence associate, for all $s \in \mathcal{S} \setminus s_r$,

$$\lambda_s = \frac{1}{2} [\gamma_s + \beta_s - \gamma_{s_r}].$$

Table B.4 in Cave et al. (1993) reports the standard error of $\tilde{\lambda}_s$ for $s \in \mathcal{S} \setminus s_r$, where $\tilde{\lambda}_s$ is the OLS estimator for the site dummy coefficient using observations collected over months 37–48. We then let $\tilde{\gamma}_s$ and $\tilde{\beta}_s$, $s \in \mathcal{S}$, be the OLS estimators that we would obtain via (A.1) using earning observations from $t=2$ only.

Under Assumption 1, $\tilde{\gamma}_s$ and $\tilde{\beta}_s$ are uncorrelated. In addition, (A.7) implies that conditional on the regressors (D_i, S_i, X_i) , ε_{i2} in the long regression of $t=2$ is necessarily smaller (stochastically) than ν_{i2} in the short regression, in the sense that $E[\nu_{i2}^2 | D_i, S_i, X_i] \geq E[\varepsilon_{i2}^2 | D_i, S_i, X_i]$ almost surely. Accordingly, the asymptotic variance of the OLS estimator for $\tilde{\lambda}_s$ obtained from the short regression is larger than the asymptotic variance of the estimator for λ_s based on the coefficient estimates from the long regression, i.e.,

$$\text{Var}(\tilde{\lambda}_s) \geq \text{Var}\left(\frac{1}{2}(\tilde{\gamma}_s + \tilde{\beta}_s - \tilde{\gamma}_{s_r})\right) \geq \frac{1}{4} \left[\text{Var}(\tilde{\gamma}_s) + \text{Var}(\tilde{\beta}_s) + \text{Var}(\tilde{\gamma}_{s_r}) \right], \quad (\text{A.8})$$

for each s in large samples.

Next, to relate the variances of $\tilde{\gamma}_s$ and $\tilde{\beta}_s$ to the variances of the OLS coefficients in the long regression for cumulative earnings (i.e., to relate these variances to the coefficients in (A.2)), we impose the following assumption.

Assumption 3 (Homoskedasticity of the earning shocks)

Conditional on regressors (D_i, S_i, X_i) , the variance of ε_{i1} is identical to the variance of ε_{i2} , and ε_{i1} is uncorrelated with ε_{i2} .

Under Assumption 3, the variance of the OLS coefficients of (A.1) with $t=2$ is half of the variance of the OLS estimators for the linear regression coefficients of (A.2); given that

the regressor matrix is common to both equations, the variance of the OLS estimators is proportional to the variance of the regression residuals. Thus, $\text{Var}(\tilde{\gamma}_s) = \text{Var}(2\hat{\gamma}_s)/2$ and $\text{Var}(\tilde{\beta}_s) = \text{Var}(2\hat{\beta}_s)/2$. Accordingly, inequality (A.8) can be written as

$$\text{Var}(\tilde{\lambda}_s) \geq \frac{1}{8} \left[\text{Var}(2\hat{\gamma}_s) + \text{Var}(2\hat{\beta}_s) + \text{Var}(2\hat{\gamma}_{s_r}) \right]. \quad (\text{A.9})$$

Note that if Assumption 3 fails and ε_{i1} and ε_{i2} are correlated, this would tend to increase the standard errors for the long regression, and hence the importance of our winner's curse corrections.

Our goal is to obtain the variance of $\hat{\tau}_s$, $s \in \mathcal{S}$, which is

$$\text{Var}(\hat{\tau}_s) = \begin{cases} \text{Var}(2\hat{\gamma}_{s_r}) & \text{for } s = s_r, \\ \text{Var}(2\hat{\gamma}_s) + \text{Var}(2\hat{\beta}_s) & \text{for } s \neq s_r. \end{cases} \quad (\text{A.10})$$

To pin this quantity down, we impose an additional assumption:

Assumption 4 (Variance)

(i.) The inequality in (A.9) holds with equality. (ii.) For all $s \in \mathcal{S} \setminus s_r$ and $t = 1, 2$, $\text{Var}(Y_{it}(0)|S_i = s) = \text{Var}(Y_{it}(1)|S_i = s)$, which implies that $\text{Var}(2\hat{\beta}_s) = \text{Var}(2\hat{\gamma}_s)$. (iii.) $\text{Var}(Y_{i1}(1) + Y_{i2}(1)|S_i = s_r) = \text{Var}(Y_{i1}(1) + Y_{i2}(1)|S_i = s_p)$, where s_p indicates Connelley Skill Learning Center in Pittsburgh, PA.

We justify our choice of Connelley Skill Learning Center as a reference for SER/Jobs for Progress by noting that the sample mean of $Y_{i1}(1) + Y_{i2}(1)$ in Pittsburgh is closest to that of Corpus Christi, TX.

Focusing on $s = s_p$ under Assumptions 1 and 4, and noting that the sample size of Connelley Skill Learning Center is 2/3 that of SER/Jobs for Progress, we obtain

$$\text{Var}(\tilde{\lambda}_{s_p}) = \frac{1}{8} \left(2\text{Var}(2\hat{\gamma}_{s_p}) + \frac{2}{3}\text{Var}(2\hat{\gamma}_{s_p}) \right) = \frac{1}{3}\text{Var}(2\hat{\gamma}_{s_p}),$$

from (A.9). Hence, we can pin down the variance of $\hat{\tau}_{s_p}$ as

$$\text{Var}(\hat{\tau}_{s_p}) = 2\text{Var}(2\hat{\gamma}_{s_p}) = 6\text{Var}(\tilde{\lambda}_{s_p}).$$

Similarly, for $s = s_r$, (A.10) gives

$$\text{Var}(\hat{\tau}_{s_r}) = \text{Var}(2\hat{\gamma}_{s_r}) = \frac{2}{3}\text{Var}(2\hat{\gamma}_{s_p}) = 2\text{Var}(\tilde{\lambda}_{s_p}),$$

whilst for the remaining sites, $s \in \mathcal{S} \setminus \{s_p, s_r\}$, we obtain

$$\text{Var}(\hat{\tau}_s) = \text{Var}(\widehat{2\gamma}_s) + \text{Var}(\widehat{2\beta}_s) = 8\text{Var}(\tilde{\lambda}_s) - \text{Var}(2\hat{\gamma}_{s_r}) = 8\text{Var}(\tilde{\lambda}_s) - 2\text{Var}(\tilde{\lambda}_{s_p}),$$

which allows us to impute standard errors for the remaining sites. Note that our imputed standard errors for all sites are consistent with the reported significance levels in Table 5.13 of Cave et al. (1993).

A.2 Conditional Coverage Results

To complement the simulation results reported in the main text, Figure 7 plots the conditional coverage given $\hat{\theta} = \theta^*$ for θ^* the site yielding the largest effect in the JOBSTART data (i.e. CET), where to illustrate coverage distortions we extend the horizontal axis to include negative scaling factors. As expected the conditional interval has correct conditional coverage, while coverage distortions appear for the hybrid and projection intervals for negative scaling factors. In this case $\hat{\theta} = \theta^*$ with low probability, but conditional on this event $X(\hat{\theta})$ tends to be far away from $\mu_X(\theta^*)$, since for $s < 0$ site θ^* has the *smallest* ATE. Consequently, projection and hybrid confidence intervals under-cover.

A.3 Split-Sample Results

This appendix reports the results from applying split-sample methods in the simulations calibrated to the JOBSTART data. As in the main text, we report results corresponding to the case where we use half of the data to select a target site, and the other half is used for inference.

As discussed in the main text, sample splitting changes the site selected. Hence, a first important question when considering sample splitting is to what extent it reduces the quality of the treatment selected, relative to using the full data for targeting. Figure 8 provides one answer to this question, plotting the average difference in treatment effects between the best site (CET/San Jose in this simulation design) and the selected site, that is $\mu_X(\theta^*) - E[\mu_X(\hat{\theta})]$ for θ^* corresponding to CET/San Jose. As in the main text, the horizontal axis varies the scaling s on the site-specific average treatment effects. As these results make clear, there is a substantial loss from sample splitting in this context, with

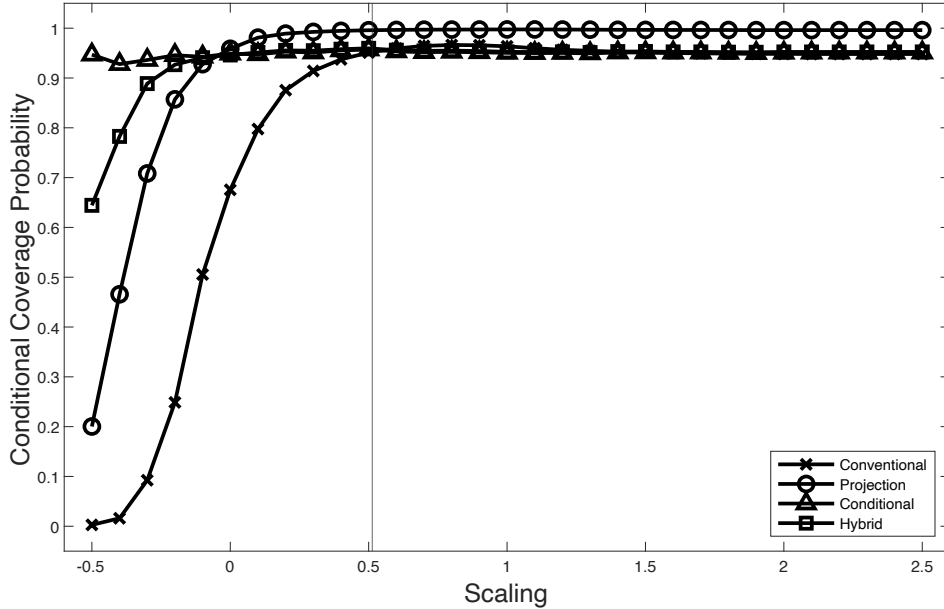


Figure 7: Coverage conditional on $\hat{\theta}=\theta^*$, for θ^* the site with the largest effect in the JOBSTART data (i.e. CET/San Jose), where $X \sim N(s \cdot \hat{\mu}_X, \Sigma)$ for $\hat{\mu}_X$ the JOBSTART point estimates and Σ the diagonal matrix with the squared JOBSTART standard errors on the diagonal. The horizontal axis varies the scaling factor s , and our preferred scaling s^* is marked with a vertical line.

the regret increasing by nearly 40% at our preferred scaling s^* .

B A Decision-Theoretic Model of Inference After Selection

This appendix shows that our unconditional and conditional coverage requirements arise naturally as necessary conditions for minimaxity in a two-step decision-theoretic model.

In particular, consider a decisionmaker who observes

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \Sigma\right),$$

with $\mu = (\mu'_X, \mu'_Y)'$ unknown and Σ known. They must make a two-part decision, first selecting an element $\theta \in \Theta$ and then reporting an interval I intended to cover $\mu_Y(\theta)$. Suppose that the decisionmaker has lexicographic preferences, prioritizing the selection problem first and the inference problem second.

For the first step selection problem, the decisionmaker must select a decision rule

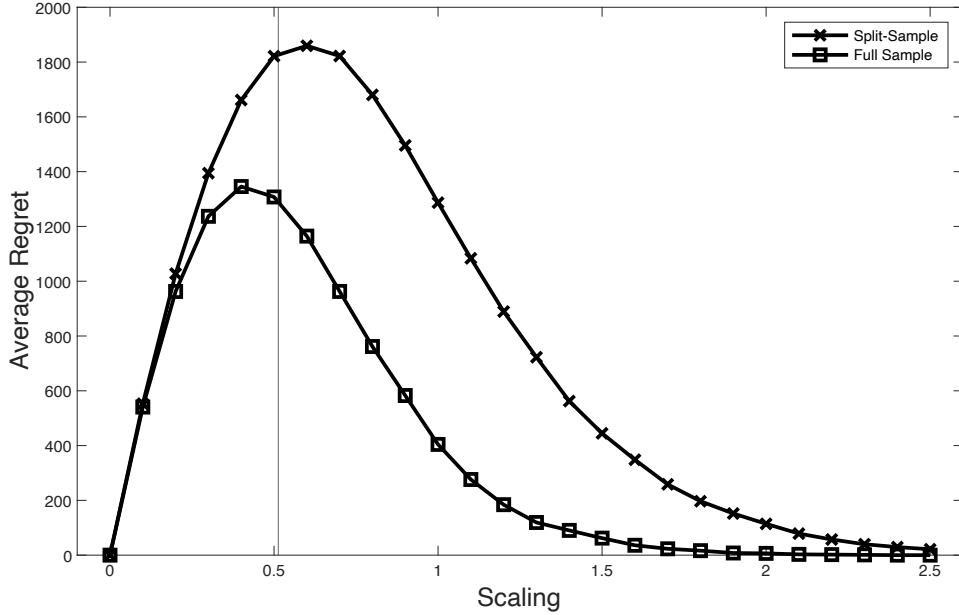


Figure 8: Average regret $\mu_X(\theta^*) - E[\mu_X(\hat{\theta})]$, for selection based on full vs. split sample, where 50% of the data are used for selection.

mapping data realizations to (possibly randomized) selections,

$$\delta_\theta: \mathcal{X} \times \mathcal{Y} \rightarrow \Delta(\Theta)$$

for $\Delta(\Theta)$ the set of probability distributions on Θ . In a slight abuse of notation we use $\delta_\theta(X, Y) \in \Theta$ to denote the realized choice. We assume that the decisionmaker aims to minimize an expected loss that depends on μ ,

$$E_\mu[L_\theta(\delta_\theta(X, Y), \mu)].$$

Since μ is unknown the decisionmaker must aggregate across μ values in some way, for example putting a prior on μ and selecting Bayes decision rules or focusing on the worst case and selecting a minimax rule.

As a concrete example of the first-stage preference, consider a decisionmaker who selects δ_θ to minimize the maximum regret based on μ_X

$$\min_{\delta_\theta} \sup_{\mu} \left\{ \max_{\theta \in \Theta} \mu_X(\theta) - E_\mu[\mu_X(\delta_\theta(X, Y))] \right\}.$$

Results from Lehmann (1966) and Eaton (1967) imply that picking $\delta_\theta(X,Y) = \operatorname{argmax} X(\theta)$ solves this problem when μ is unconstrained and Σ_X is proportional to the identity matrix. More broadly there is a substantial statistical decision theory literature on optimal selection. For our purposes, we will take the solution to the first-step decision problem as given and focus on the second-stage decision problem.

In the second stage the decisionmaker solves an inference problem. Define the loss

$$L_2(\theta, I, \mu) = f(|I|) + 1\{\mu_Y(\theta) \notin I\} = f(|I|) + \sum_{\tilde{\theta} \in \Theta} 1\{\theta = \tilde{\theta}\} 1\{\mu_Y(\tilde{\theta}) \notin I\}$$

where $|I|$ is the Lebesgue measure of the interval I while $f(\cdot)$ is a non-negative and weakly increasing function with $f(\infty) = \alpha^*$. We consider two versions of expected loss in the second-step problem. The first simply averages the loss across (X,Y) realizations. For second-step decision rule $\delta_I: \mathcal{X} \times \mathcal{Y} \rightarrow \Delta(\mathcal{I})$ (for \mathcal{I} the set of intervals on \mathbb{R} and $\Delta(\mathcal{I})$ the set of probability distributions on \mathcal{I}) this yields expected loss

$$E_\mu[f(|\delta_I(X,Y)|) + 1\{\mu_Y(\delta_\theta(X,Y)) \notin \delta_I(X,Y)\}]. \quad (\text{A.11})$$

Alternatively, for settings where we are concerned about selection we introduce a selection dummy S where as in the main text we assume that S is independent of (X,Y) conditional on $\delta_\theta(X,Y)$. Denote the conditional distribution of $S|\delta_\theta(X,Y)$ by $F_{S|\delta_\theta}$. If we care about the second-step loss only in those cases where $S=1$, this yields expected loss

$$E_{\mu, F_{S|\delta_\theta}}[f(|\delta_I(X,Y)|) + 1\{\mu_Y(\delta_\theta(X,Y)) \notin \delta_I(X,Y)\} | S=1]. \quad (\text{A.12})$$

We next show that minimax decision rules in the second stage imply correct coverage, with criterion (A.11) yielding unconditional coverage and criterion (A.12) yielding conditional coverage. For (A.11), note that always choosing $I = \mathbb{R}$ yields a loss of $f(\infty) = \alpha^*$ regardless of the value of μ . Hence, for a minimax decision rule δ_I^* we must have that the worst-case expected loss is weakly less than α^* ,

$$\sup_{\mu} E_\mu[f(|\delta_\theta(X,Y)|) + 1\{\mu_Y(\delta_\theta(X,Y)) \notin \delta_I^*(X,Y)\}] \leq \alpha^*.$$

Since $f(\cdot)$ is non-negative this implies that

$$\sup_{\mu} Pr_{\mu} \{ \mu_Y(\delta_{\theta}(X, Y)) \notin \delta_I^*(X, Y) \} \leq \alpha^*,$$

which is precisely our unconditional coverage requirement. Hence, minimax second-stage rules under criterion (A.11) always have unconditional coverage at least $1 - \alpha^*$.

If we instead consider (A.12), note that the unknown parameters now include both μ and the conditional distribution $F_{S|\delta_{\theta}(X, Y)}$. By the same argument as in the unconditional case the minimax expected loss must be bounded above by α^* ,

$$\sup_{\mu} \sup_{F_{S|\delta_{\theta}}} E_{\mu, F_{S|\delta_{\theta}}} [f(|\delta_{\theta}(X, Y)|) + 1 \{ \mu_Y(\delta_{\theta}(X, Y)) \notin \delta_I^*(X, Y) \} | S = 1] \leq \alpha^*,$$

which implies that

$$\sup_{\mu} \sup_{F_{S|\delta_{\theta}}} Pr_{\mu, F_{S|\delta_{\theta}}} \{ \mu_Y(\delta_{\theta}(X, Y)) \notin \delta_I^*(X, Y) | S = 1 \} \leq \alpha^*.$$

As noted in the main text, however, to ensure that

$$\sup_{F_{S|\delta_{\theta}}} Pr_{\mu, F_{S|\delta_{\theta}}} \{ \mu_Y(\delta_{\theta}(X, Y)) \notin \delta_I^*(X, Y) | S = 1 \} \leq \alpha^*$$

we must have that

$$Pr_{\mu} \{ \mu_Y(\delta_{\theta}(X, Y)) \notin \delta_I^*(X, Y) | \delta_{\theta}(X, Y) \} \leq \alpha^*,$$

so minimax rules in this setting must ensure conditional coverage at least $1 - \alpha^*$.

C Conditional Inference

This section extends the conditional inference results developed in Section 4 of the main text in two directions, first allowing dependence on additional conditioning variables, and then introducing uniformly most accurate unbiased confidence intervals.

C.1 Additional Conditioning Events

Suppose that in addition to conditioning on $\{\hat{\theta} = \tilde{\theta}\}$, we also want to condition on an additional event $\{\hat{\gamma} = \tilde{\gamma}\}$, for $\hat{\gamma} = \gamma(X)$ some function of X . We thus seek estimators that

are quantile-unbiased conditional on $(\hat{\theta}, \hat{\gamma})$,

$$Pr_{\mu} \left\{ \hat{\mu}_{\alpha} \geq \mu_Y(\hat{\theta}) \mid \hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma} \right\} = \alpha \text{ for all } \tilde{\theta} \in \Theta, \tilde{\gamma} \in \Gamma, \text{ and all } \mu, \quad (\text{A.13})$$

and confidence sets with correct conditional coverage

$$Pr_{\mu} \left\{ \mu_Y(\hat{\theta}) \in CI \mid \hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma} \right\} \geq 1 - \alpha \text{ for all } \tilde{\theta} \in \Theta, \tilde{\gamma} \in \Gamma, \text{ and all } \mu. \quad (\text{A.14})$$

One reason we might want to allow such additional conditioning is that we are interested in performance conditional on $S=1$ for an unobserved variable $S \in \{0,1\}$ as discussed in Section 3 of the main text, where we generalize the assumption in the main text and assume that S is conditionally independent of (X, Y) given the pair $(\hat{\theta}, \hat{\gamma})$. If we have no other restrictions on the distribution of S , then in order to guarantee conditional coverage given $S=1$,

$$\inf_{\mu} \inf_{F_{S|\hat{\theta}, \hat{\gamma}}} Pr_{\mu, F_{S|\hat{\theta}, \hat{\gamma}}} \left\{ \mu_Y(\hat{\theta}) \in CS \mid S=1 \right\} \geq 1 - \alpha,$$

it is both necessary and sufficient that we have conditional coverage (A.14).

As in the main text, we re-write the conditioning event in terms of the sample space of X as $\left\{ X : \hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma} \right\} = \mathcal{X}(\tilde{\theta}, \tilde{\gamma})$, and study the conditional distribution of $(X, Y(\tilde{\theta}))$ given $X \in \mathcal{X}(\tilde{\theta}, \tilde{\gamma})$. For $Z_{\tilde{\theta}}$ as defined in (10) of the main text, let

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

Conditional on $\hat{\theta} = \tilde{\theta}$, $\hat{\gamma} = \tilde{\gamma}$, and $Z_{\tilde{\theta}} = z$, $Y(\hat{\theta})$ again follows a one-dimensional normal distribution $N(\mu_Y(\tilde{\theta}), \Sigma_Y(\tilde{\theta}))$ truncated to $\mathcal{Y}(\tilde{\theta}, \tilde{\gamma}, z)$.

To characterize $\mathcal{Y}(\tilde{\theta}, \tilde{\gamma}, z)$, note that for $\mathcal{X}(\tilde{\theta})$ as derived in the main text, we can write $\mathcal{X}(\tilde{\theta}, \tilde{\gamma}) = \mathcal{X}(\tilde{\theta}) \cap \mathcal{X}_{\tilde{\gamma}}(\tilde{\gamma})$ for $\mathcal{X}_{\tilde{\gamma}}(\tilde{\gamma}) = \{X \in \mathcal{X} : \gamma(X) = \tilde{\gamma}\}$. Likewise, for $\mathcal{Y}_{\tilde{\gamma}}(\tilde{\gamma}, z)$ defined analogously to (11) in the main text, $\mathcal{Y}(\tilde{\theta}, \tilde{\gamma}, z) = \mathcal{Y}(\tilde{\theta}, z) \cap \mathcal{Y}_{\tilde{\gamma}}(\tilde{\gamma}, z)$. The form of $\mathcal{X}_{\tilde{\gamma}}(\tilde{\gamma})$ and $\mathcal{Y}_{\tilde{\gamma}}(\tilde{\gamma}, z)$ depends on the conditioning variables $\hat{\gamma}$ considered.

To construct quantile-unbiased estimators, let $F_{TN}(y; \mu_Y(\tilde{\theta}), \tilde{\theta}, \tilde{\gamma}, z)$ denote the distribution function for a $N(\mu_Y(\tilde{\theta}), \Sigma_Y(\tilde{\theta}))$ distribution truncated to $\mathcal{Y}(\tilde{\theta}, \tilde{\gamma}, z)$. This function is strictly decreasing in $\mu_Y(\tilde{\theta})$, so define $\hat{\mu}_{\alpha}$ as the unique solution to

$$F_{TN}(Y(\hat{\theta}); \hat{\mu}_{\alpha}, \tilde{\theta}, \tilde{\gamma}, Z_{\tilde{\theta}}) = 1 - \alpha. \quad (\text{A.15})$$

To establish optimality, we impose one additional assumption.

Assumption 5

If $\Sigma = \text{Cov}((X', Y)')$ has full rank, then the parameter space for μ is $\mathbb{R}^{2|\Theta|}$. Otherwise, there exists some μ^* such that the parameter space for μ is $\left\{ \mu^* + \Sigma^{\frac{1}{2}}v : v \in \mathbb{R}^{2|\Theta|} \right\}$, where $\Sigma^{\frac{1}{2}}$ is the symmetric square root of Σ .

This assumption requires that the parameter space for μ be sufficiently rich. When Σ is degenerate (for example when $X = Y$, as in Section 2 of the main text), this assumption further implies that (X, Y) have the same support for all values of μ . This rules out cases in which a pair of parameter values μ_1, μ_2 can be perfectly distinguished based on the data. Under this assumption, $\hat{\mu}_\alpha$ is an optimal quantile-unbiased estimator.

Proposition 7

Let $\hat{\mu}_\alpha$ solve (A.15). $\hat{\mu}_\alpha$ is conditionally α -quantile-unbiased in the sense of (A.13). If Assumption 5 holds, then $\hat{\mu}_\alpha$ is the uniformly most concentrated α -quantile-unbiased estimator in that for any other conditionally α -quantile-unbiased estimator $\hat{\mu}_\alpha^*$ and any loss function $L(d, \mu_Y(\tilde{\theta}))$ that attains its minimum at $d = \mu_Y(\tilde{\theta})$ and is quasiconvex in d for all $\mu_Y(\tilde{\theta})$,

$$E_\mu \left[L\left(\hat{\mu}_\alpha, \mu_Y(\tilde{\theta})\right) \mid \hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma} \right] \leq E_\mu \left[L\left(\hat{\mu}_\alpha^*, \mu_Y(\tilde{\theta})\right) \mid \hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma} \right]$$

for all μ and all $\tilde{\theta} \in \Theta, \tilde{\gamma} \in \Gamma$.

Proposition 7 shows that $\hat{\mu}_\alpha$ is optimal in the strong sense that it has lower expected loss than any other quantile-unbiased estimator for a large class of loss functions. Hence, $\hat{\mu}_{\frac{1}{2}}$ is an optimal median-unbiased estimator, while $CI_{ET} = [\hat{\mu}_{\frac{\alpha}{2}}, \hat{\mu}_{1-\frac{\alpha}{2}}]$ is an optimal equal-tailed confidence interval.

C.2 Unbiased Confidence Intervals

Rather than considering equal-tailed intervals, we can alternatively consider unbiased confidence intervals. Following Lehmann and Romano (2005), we say that a level $1 - \alpha$ two-sided confidence interval CI is unbiased if its probability of covering any given false parameter value is bounded above by $1 - \alpha$. Likewise, a one sided lower (upper) confidence interval is unbiased if its probability of covering a false parameter value above (below) the true value is bounded above by $1 - \alpha$. Using the duality between tests and confidence intervals, a level $1 - \alpha$ confidence interval CI is unbiased if and only if $\phi(\mu_{Y,0}) = 1 \{ \mu_{Y,0} \notin CI \}$

is an unbiased test for the corresponding family of hypotheses.² The results of Lehmann and Scheffé (1955) applied in our setting imply that optimal unbiased tests conditional on $\{\hat{\theta}=\tilde{\theta}, \hat{\gamma}=\tilde{\gamma}\}$ are the same as optimal unbiased tests conditional on $\{\hat{\theta}=\tilde{\theta}, \hat{\gamma}=\tilde{\gamma}, Z_{\tilde{\theta}}=z_{\tilde{\theta}}\}$. These optimal tests take a simple form.

Define a size α test of the two-sided hypothesis $H_0: \mu_Y(\tilde{\theta}) = \mu_{Y,0}$ as

$$\phi_{TS,\alpha}(\mu_{Y,0}) = 1 \left\{ Y(\tilde{\theta}) \notin [c_l(Z_{\tilde{\theta}}), c_u(Z_{\tilde{\theta}})] \right\}$$

where $c_l(z)$, $c_u(z)$ solve

$$Pr\{\zeta \in [c_l(z), c_u(z)]\} = 1 - \alpha, \quad E[\zeta 1\{\zeta \in [c_l(z), c_u(z)]\}] = (1 - \alpha)E[\zeta]$$

for ζ that follows a truncated normal distribution

$$\zeta \sim \xi | \xi \in \mathcal{Y}(\tilde{\theta}, \tilde{\gamma}, z), \quad \xi \sim N(\mu_{Y,0}, \Sigma_Y(\tilde{\theta})).$$

Likewise, define a size α test of the one-sided hypothesis $H_0: \mu_Y(\tilde{\theta}) \geq \mu_{Y,0}$ as

$$\phi_{OS-, \alpha}(\mu_{Y,0}) = 1 \left\{ F_{TN}(Y(\tilde{\theta}); \mu_{Y,0}, \tilde{\theta}, \tilde{\gamma}, z) \leq \alpha \right\}$$

and a test of $H_0: \mu_Y(\tilde{\theta}) \leq \mu_{Y,0}$ as

$$\phi_{OS+, \alpha}(\mu_{Y,0}) = 1 \left\{ F_{TN}(Y(\tilde{\theta}); \mu_{Y,0}, \tilde{\theta}, \tilde{\gamma}, z) \geq 1 - \alpha \right\}.$$

Proposition 8

If Assumption 5 holds, $\phi_{TS,\alpha}$, $\phi_{OS-, \alpha}$, and $\phi_{OS+, \alpha}$ are uniformly most powerful unbiased size α tests of their respective null hypotheses conditional on $\hat{\theta}=\tilde{\theta}$ and $\hat{\gamma}=\tilde{\gamma}$.

To form uniformly most accurate unbiased confidence intervals we collect the values not rejected by these tests. The two-sided uniformly most accurate unbiased confidence interval is $CI_U = \{\mu_{Y,0} : \phi_{TS,\alpha}(\mu_{Y,0}) = 0\}$. CI_U is unbiased and has conditional coverage $1 - \alpha$ by construction. Likewise, we can form lower and upper one-sided uniformly most accurate unbiased confidence intervals as $CI_{U,-} = \{\mu_{Y,0} : \phi_{OS-, \alpha}(\mu_{Y,0}) = 0\} = (-\infty, \hat{\mu}_{1-\alpha}]$, and $CI_{U,+} = \{\mu_{Y,0} : \phi_{OS+, \alpha}(\mu_{Y,0}) = 0\} = [\hat{\mu}_\alpha, \infty)$, respectively. Hence, we can view CI_{ET} as

²That is, $H_0: \mu_Y(\tilde{\theta}) = \mu_{Y,0}$ for a two-sided confidence interval, $H_0: \mu_Y(\tilde{\theta}) \geq \mu_{Y,0}$ for a lower confidence interval and $H_0: \mu_Y(\tilde{\theta}) \leq \mu_{Y,0}$ for an upper confidence interval.

the intersection of level $1 - \frac{\alpha}{2}$ uniformly most accurate unbiased upper and lower confidence intervals. Unfortunately, no such simplification is generally available for CI_U , though Lemma 5.5.1 of Lehmann and Romano (2005) guarantees that this set is an interval.

C.3 Behavior When $Pr_\mu\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\}$ is Large

In Proposition 3 of the main text, we showed that our median-unbiased estimators and equal-tailed confidence intervals converge to conventional ones when $Pr_\mu\{\hat{\theta}=\tilde{\theta}\} \rightarrow 1$. The same result holds for general conditioning events and unbiased confidence intervals.

Lemma 2

Consider any sequence of values $\mu_{Y,m}$ and $z_{\tilde{\theta},m}$ such that $Pr_{\mu_{Y,m}}\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}|Z_{\tilde{\theta}}=z_{\tilde{\theta},m}\} \rightarrow 1$. Then under $\mu_{Y,m}$, conditional on $\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma},Z_{\tilde{\theta}}=z_{\tilde{\theta},m}\}$ we have $CI_U \rightarrow_p CI_N$, $CI_{ET} \rightarrow_p CI_N$, and $\hat{\mu}_{\frac{1}{2}} \rightarrow_p Y(\tilde{\theta})$.

Proposition 9

Consider any sequence of values μ_m such that $Pr_{\mu_m}\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\} \rightarrow 1$. Then under μ_m , we have $CI_U \rightarrow_p CI_N$, $CI_{ET} \rightarrow_p CI_N$, and $\hat{\mu}_{\frac{1}{2}} \rightarrow_p Y(\tilde{\theta})$ both conditional on $\{\hat{\theta}=\tilde{\theta},\hat{\gamma}=\tilde{\gamma}\}$ and unconditionally.

D Proofs

We first prove the results stated in Section C, and then build on these to prove the results for the finite-sample normal model discussed in the main text.

D.1 Proofs for Results in Section C

Proof of Proposition 7 For ease of reference, let us abbreviate $(Y(\tilde{\theta}),\mu_Y(\tilde{\theta}),Z_{\tilde{\theta}})$ by $(\tilde{Y},\tilde{\mu}_Y,\tilde{Z})$. Let $Y(-\tilde{\theta})$ collect the elements of Y other than $Y(\tilde{\theta})$ and define $\mu_Y(-\theta)$ analogously. Let

$$Y^* = Y(-\tilde{\theta}) - Cov\left(Y(-\tilde{\theta}), \begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right) Var\left(\begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right)^+ \begin{pmatrix} \tilde{Y} \\ X \end{pmatrix},$$

$$\mu_Y^* = \mu_Y(-\tilde{\theta}) - Cov\left(Y(-\tilde{\theta}), \begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right) Var\left(\begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right)^+ \begin{pmatrix} \tilde{\mu}_Y \\ \mu_X \end{pmatrix},$$

and $\tilde{\mu}_Z = \mu_X - (\Sigma_{XY}(\cdot,\tilde{\theta})/\Sigma_Y(\tilde{\theta}))\mu_Y$. Here we use A^+ to denote the Moore-Penrose pseudoinverse of a matrix A . Note that $(\tilde{Z},\tilde{Y},Y^*)$ is a one-to-one transformation of (X,Y) ,

and thus that observing $(\tilde{Z}, \tilde{Y}, Y^*)$ is equivalent to observing (X, Y) . Likewise, $(\tilde{\mu}_Z, \tilde{\mu}_Y, \mu_{Y^*}^*)$ is a one-to-one linear transformation of (μ_X, μ_Y) , and if the set of possible values for the latter contains an open set, that for the former does as well (relative to the appropriate linear subspace).

Note, next, that since $(\tilde{Z}, \tilde{Y}, Y^*)$ is a linear transformation of (X, Y) , $(\tilde{Z}, \tilde{Y}, Y^*)$ is jointly normal (with a potentially degenerate distribution). Note next that the subvectors of $(\tilde{Z}, \tilde{Y}, Y^*)$ are mutually uncorrelated, and thus independent. That \tilde{Z} and \tilde{Y} are uncorrelated is straightforward to verify. To show that Y^* is likewise uncorrelated with the other elements, note that we can write $Cov(Y^*, (\tilde{Y}, X)')$ as

$$Cov\left(Y(-\tilde{\theta}), \begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right) - Cov\left(Y(-\tilde{\theta}), \begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right) Var\left(\begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right)^+ Var\left(\begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right).$$

For $V\Lambda V'$ an eigendecomposition of $Var((\tilde{Y}, X)')$ (so $VV' = I$), note that we can write

$$Var\left(\begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right)^+ Var\left(\begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right) = VDV'$$

for D a diagonal matrix with ones in the entries corresponding to the nonzero entries of Λ and zeros everywhere else. For any column v of V corresponding to a zero entry of D , $v'Var((\tilde{Y}, X)')v = 0$, so the Cauchy-Schwarz inequality implies that

$$Cov\left(Y(-\tilde{\theta}), \begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right)v = 0.$$

Thus,

$$Cov\left(Y(-\tilde{\theta}), \begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right)VDV' = Cov\left(Y(-\tilde{\theta}), \begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right)VV' = Cov\left(Y(-\tilde{\theta}), \begin{pmatrix} \tilde{Y} \\ X \end{pmatrix}\right),$$

so Y^* is uncorrelated with $(\tilde{Y}, X)'$.

Using independence, the joint density of $(\tilde{Z}, \tilde{Y}, Y^*)$ absent truncation is given by

$$f_{N, \tilde{Z}}(\tilde{z}; \tilde{\mu}_Z) f_{N, \tilde{Y}}(\tilde{y}; \tilde{\mu}_Y) f_{N, Y^*}(\tilde{y}^*; \mu_{Y^*}^*)$$

for f_N normal densities with respect to potentially degenerate base measures:

$$f_{N,\tilde{Z}}(\tilde{z};\tilde{\mu}_Z) = \tilde{\det}(2\pi\Sigma_{\tilde{Z}})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\tilde{z}-\tilde{\mu}_Z)'\Sigma_{\tilde{Z}}^+(\tilde{z}-\tilde{\mu}_Z)\right)$$

$$f_{N,\tilde{Y}}(\tilde{y};\tilde{\mu}_Y) = (2\pi\Sigma_{\tilde{Y}})^{-\frac{1}{2}} \exp\left(-\frac{(\tilde{y}-\tilde{\mu}_Y)^2}{2\Sigma_{\tilde{Y}}}\right)$$

$$f_{N,Y^*}(y^*;\mu_Y^*) = \tilde{\det}(2\pi\Sigma_{Y^*})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y^*-\tilde{\mu}_Y^*)'\Sigma_{Y^*}^+(y^*-\mu_Y^*)\right),$$

where $\tilde{\det}(A)$ denotes the pseudodeterminant of a matrix A , $\Sigma_{\tilde{Z}} = \text{Var}(\tilde{Z})$, $\Sigma_{\tilde{Y}} = \Sigma_Y(\tilde{\theta})$, and $\Sigma_{Y^*} = \text{Var}(Y^*)$.

The event $\{X \in \mathcal{X}(\tilde{\theta}, \tilde{\gamma})\}$ depends only on (\tilde{Z}, \tilde{Y}) since it can be expressed as

$$\left\{ \left(\tilde{Z} + \frac{\Sigma_{XY}(\cdot, \tilde{\theta})}{\Sigma_Y(\tilde{\theta})} \tilde{Y} \right) \in \mathcal{X}(\tilde{\theta}, \tilde{\gamma}) \right\},$$

so conditional on this event Y^* remains independent of (\tilde{Z}, \tilde{Y}) . In particular, we can write the joint density conditional on $\{X \in \mathcal{X}(\tilde{\theta}, \tilde{\gamma})\}$ as

$$\frac{1\left\{ \left(\tilde{z} + \Sigma_{XY}(\cdot, \tilde{\theta})\Sigma_Y(\tilde{\theta})^{-1}\tilde{y} \right) \in \mathcal{X}(\tilde{\theta}, \tilde{\gamma}) \right\}}{\text{Pr}_{\tilde{\mu}_Z, \tilde{\mu}_Y} \left\{ X \in \mathcal{X}(\tilde{\theta}, \tilde{\gamma}) \right\}} f_{N,\tilde{Z}}(\tilde{z};\tilde{\mu}_Z) f_{N,\tilde{Y}}(\tilde{y};\tilde{\mu}_Y) f_{N,Y^*}(\tilde{y}^*;\mu_Y^*). \quad (\text{A.16})$$

The density (A.16) has the same structure as (5.5.14) of Pfanzagl (1994), and satisfies properties (5.5.1)-(5.5.3) of Pfanzagl (1994) as well. Part 1 of the proposition then follows immediately from Theorem 5.5.9 of Pfanzagl (1994). Part 2 of the proposition follows by using Theorem 5.5.9 of Pfanzagl (1994) to verify the conditions of Theorem 5.5.15 of Pfanzagl (1994). \square

Proof of Proposition 8 In the proof of Proposition 7, we showed that the joint density of $(\tilde{Z}, \tilde{Y}, Y^*)$ (defined in that proof) has the exponential family structure assumed in equation 4.10 of Lehmann and Romano (2005). Moreover, Assumption 5 implies that the parameter space for (μ_X, μ_Y) is convex and is not contained in any proper linear subspace. Thus, the parameter space for $(\tilde{\mu}_Z, \tilde{\mu}_Y, \mu_Y^*)$ inherits the same property, and satisfies the conditions of Theorem 4.4.1 of Lehmann and Romano (2005). The result follows immediately. \square

Proof of Lemma 2 Recall that conditional on $Z_{\tilde{\theta}} = z_{\tilde{\theta}}$, $\hat{\theta} = \tilde{\theta}$ and $\hat{\gamma} = \tilde{\gamma}$ if and only if $Y(\tilde{\theta}) \in \mathcal{Y}(\tilde{\theta}, \tilde{\gamma}, z_{\tilde{\theta}})$. Hence, the assumption of the lemma implies that

$$Pr_{\mu_{Y,m}} \left\{ Y(\tilde{\theta}) \in \mathcal{Y}(\tilde{\theta}, \tilde{\gamma}, Z_{\tilde{\theta}}) | Z_{\tilde{\theta}} = z_{\tilde{\theta},m} \right\} \rightarrow 1.$$

Note, next, that both the conventional and conditional confidence intervals are equivariant under shifts, in the sense that the conditional confidence interval for $\mu_Y(\tilde{\theta})$ based on observing $Y(\tilde{\theta})$ conditional on $Y(\tilde{\theta}) \in \mathcal{Y}(\tilde{\theta}, \tilde{\gamma}, Z_{\tilde{\theta}})$ is equal to the conditional confidence interval for $\mu_Y(\tilde{\theta})$ based on observing $Y(\tilde{\theta}) - \mu_Y^*(\tilde{\theta})$ conditional on $Y(\tilde{\theta}) - \mu_Y^*(\tilde{\theta}) \in \mathcal{Y}(\tilde{\theta}, \tilde{\gamma}, Z_{\tilde{\theta}}) - \mu_Y^*(\tilde{\theta})$ for any constant $\mu_Y^*(\tilde{\theta})$. Hence, rather than considering a sequence of values $\mu_{Y,m}$, we can fix some μ_Y^* and note that $Pr_{\mu_Y^*} \left\{ Y(\tilde{\theta}) \in \mathcal{Y}_m^* | Z_{\tilde{\theta}} = z_{\tilde{\theta},m} \right\} \rightarrow 1$, where $\mathcal{Y}_m^* = \mathcal{Y}(\tilde{\theta}, \tilde{\gamma}, Z_{\tilde{\theta}}) - \mu_{Y,m}(\tilde{\theta}) + \mu_Y^*(\tilde{\theta})$. Confidence intervals for $\mu_{Y,m}(\tilde{\theta})$ in the original problem are equal to those for $\mu_Y^*(\tilde{\theta})$ in the new problem, shifted by $\mu_{Y,m}(\tilde{\theta}) - \mu_Y^*(\tilde{\theta})$. Hence, to prove the result it suffices to prove the equivalence of conditional and conventional confidence intervals in the problem with μ_Y fixed (and likewise for estimators).

To prove the result, we make use of the following lemma, which is proved below. First, we must introduce the following notation. Let $(c_{l,ET}(\mu_{Y,0}, \mathcal{Y}), c_{u,ET}(\mu_{Y,0}, \mathcal{Y}))$ denote the critical values for an equal-tailed test of $H_0 : \mu_Y(\tilde{\theta}) = \mu_{Y,0}$ for $Y(\tilde{\theta}) \sim N(\mu_Y(\tilde{\theta}), \Sigma_Y(\tilde{\theta}))$ conditional on $Y(\tilde{\theta}) \in \mathcal{Y}$. That is, $(c_{l,ET}(\mu_{Y,0}, \mathcal{Y}), c_{u,ET}(\mu_{Y,0}, \mathcal{Y}))$ solve $F_{TN}(c_{l,ET}(\mu_{Y,0}, \mathcal{Y}); \mu_{Y,0}, \mathcal{Y}) = \frac{\alpha}{2}$ and $F_{TN}(c_{u,ET}(\mu_{Y,0}, \mathcal{Y}); \mu_{Y,0}, \mathcal{Y}) = 1 - \frac{\alpha}{2}$, where $F_{TN}(\cdot; \mu_{Y,0}, \mathcal{Y})$ is the distribution function for the normal distribution $N(\mu_{Y,0}, \Sigma_Y(\tilde{\theta}))$ truncated to \mathcal{Y} . Similarly, let $(c_{l,U}(\mu_{Y,0}, \mathcal{Y}), c_{u,U}(\mu_{Y,0}, \mathcal{Y}))$ denote the critical values for the corresponding unbiased test. That is, $(c_{l,U}(\mu_{Y,0}, \mathcal{Y}), c_{u,U}(\mu_{Y,0}, \mathcal{Y}))$ solve $Pr\{\zeta \in [c_{l,U}(\mu_{Y,0}, \mathcal{Y}), c_{u,U}(\mu_{Y,0}, \mathcal{Y})]\} = 1 - \alpha$ and $E[\zeta 1\{\zeta \in [c_{l,U}(\mu_{Y,0}, \mathcal{Y}), c_{u,U}(\mu_{Y,0}, \mathcal{Y})]\}] = (1 - \alpha) E[\zeta]$ for $\zeta \sim \xi | \xi \in \mathcal{Y}$ where $\xi \sim N(\mu_{Y,0}, \Sigma_Y(\tilde{\theta}))$.

Lemma 3

Suppose that we observe $Y(\tilde{\theta}) \sim N(\mu_Y(\tilde{\theta}), \Sigma_Y(\tilde{\theta}))$ conditional on $Y(\tilde{\theta})$ falling in a set \mathcal{Y} . If we hold $(\Sigma_Y(\tilde{\theta}), \mu_{Y,0})$ fixed and consider a sequence of sets \mathcal{Y}_m such that $Pr\{Y(\tilde{\theta}) \in \mathcal{Y}_m\} \rightarrow 1$, we have that for

$$\phi_{ET}(\mu_{Y,0}) = 1 \left\{ Y(\tilde{\theta}) \notin [c_{l,ET}(\mu_{Y,0}, \mathcal{Y}_m), c_{u,ET}(\mu_{Y,0}, \mathcal{Y}_m)] \right\} \quad (\text{A.17})$$

and

$$\phi_U(\mu_{Y,0}) = 1 \left\{ Y(\tilde{\theta}) \notin [c_{l,U}(\mu_{Y,0}, \mathcal{Y}_m), c_{u,U}(\mu_{Y,0}, \mathcal{Y}_m)] \right\}, \quad (\text{A.18})$$

$$(c_{l,ET}(\mu_{Y,0}, \mathcal{Y}_m), c_{u,ET}(\mu_{Y,0}, \mathcal{Y}_m)) \rightarrow \left(\mu_{Y,0} - c_{\frac{\alpha}{2}, N} \sqrt{\Sigma_Y(\tilde{\theta})}, \mu_{Y,0} + c_{\frac{\alpha}{2}, N} \sqrt{\Sigma_Y(\tilde{\theta})} \right)$$

and

$$(c_{l,U}(\mu_{Y,0}, \mathcal{Y}_m), c_{u,U}(\mu_{Y,0}, \mathcal{Y}_m)) \rightarrow \left(\mu_{Y,0} - c_{\frac{\alpha}{2}, N} \sqrt{\Sigma_Y(\tilde{\theta})}, \mu_{Y,0} + c_{\frac{\alpha}{2}, N} \sqrt{\Sigma_Y(\tilde{\theta})} \right).$$

To complete the proof, first note that CI_{ET} and CI_U are formed by inverting (families of) equal-tailed and unbiased tests, respectively. Let CI_m denote a generic conditional confidence interval formed by inverting a family of tests

$$\phi_m(\mu_{Y,0}) = 1 \left\{ Y(\tilde{\theta}) \notin [c_l(\mu_{Y,0}, \mathcal{Y}_m^*), c_u(\mu_{Y,0}, \mathcal{Y}_m^*)] \right\}.$$

Hence, we want to show that

$$CI_m \rightarrow_p \left[Y(\tilde{\theta}) - c_{\frac{\alpha}{2}, N} \sqrt{\Sigma_Y(\tilde{\theta})}, Y(\tilde{\theta}) + c_{\frac{\alpha}{2}, N} \sqrt{\Sigma_Y(\tilde{\theta})} \right], \quad (\text{A.19})$$

as $m \rightarrow \infty$, for CI_m formed by inverting either (A.17) or (A.18).

We note that CI_m is a finite interval for all m , which holds trivially for the equal-tailed confidence interval CI_{ET} , and holds for CI_U by Lemma 5.5.1 of Lehmann and Romano (2005). For each value $\mu_{Y,0}$ our Lemma 3 implies that

$$\phi_m(\mu_{Y,0}) \rightarrow_p 1 \left\{ Y(\tilde{\theta}) \notin \left[\mu_{Y,0} - c_{\frac{\alpha}{2}, N} \sqrt{\Sigma_Y(\tilde{\theta})}, \mu_{Y,0} + c_{\frac{\alpha}{2}, N} \sqrt{\Sigma_Y(\tilde{\theta})} \right] \right\}$$

for ϕ_m equal to either (A.17) or (A.18). This convergence in probability holds jointly for all finite collections of values $\mu_{Y,0}$, however, which implies (A.19). The same argument works for the median unbiased estimator $\hat{\mu}_{\frac{1}{2}}$, which can also be viewed as the upper endpoint of a one-sided 50% confidence interval. \square

Proof of Proposition 9 We prove this result for the unconditional case, noting that since $Pr_{\mu_m} \left\{ \hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma} \right\} \rightarrow 1$, the result conditional on $\left\{ \hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma} \right\}$ follows immediately.

Note that $Pr_{\mu_m} \left\{ \hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma} \right\} \rightarrow 1$ implies $Pr_{\mu_{Y,m}} \left\{ \hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma} | Z_{\tilde{\theta}} \right\} \rightarrow_p 1$. Hence, for $g(\mu_Y, z) = Pr_{\mu_Y} \left\{ \hat{\theta} = \tilde{\theta}, \hat{\gamma} = \tilde{\gamma} | Z_{\tilde{\theta}} = z \right\}$, we see that $g(\mu_{Y,m}, Z_{\tilde{\theta}}) \rightarrow_p 1$. Note, next,

that for d the Euclidian distance between the endpoints, if we define $h_\varepsilon(\mu_Y, z) = Pr_{\mu_Y}\{d(CI_U, CI_N) > \varepsilon | Z_{\tilde{\theta}} = z\}$, Lemma 2 implies that for any sequence $(\mu_{Y,m}, z_m)$ such that $g(\mu_{Y,m}, z_m) \rightarrow 1$, $h_\varepsilon(\mu_{Y,m}, z_m) \rightarrow 0$. Hence, if we define $\mathcal{G}(\delta) = \{(\mu_Y, z) : g(\mu_Y, z) > 1 - \delta\}$ and $\mathcal{H}(\varepsilon) = \{(\mu_Y, z) : h_\varepsilon(\mu_Y, z) < \varepsilon\}$, for all $\varepsilon > 0$ there exists $\delta(\varepsilon) > 0$ such that $\mathcal{G}(\delta(\varepsilon)) \subseteq \mathcal{H}(\varepsilon)$.

Hence, since our argument above implies that for all $\delta > 0$, $Pr_{\mu_m}\{(\mu_{Y,m}, Z_{\tilde{\theta}}) \in \mathcal{G}(\delta)\} \rightarrow 1$, we see that for all $\varepsilon > 0$, $Pr_{\mu_m}\{(\mu_{Y,m}, Z_{\tilde{\theta}}) \in \mathcal{H}(\varepsilon)\} \rightarrow 1$ as well, which suffices to prove the desired claim for confidence intervals. The same argument likewise implies the result for our median unbiased estimator. \square

Proof of Lemma 3 Note that we can assume without loss of generality that $\mu_{Y,0} = 0$ and $\Sigma_Y(\tilde{\theta}) = 1$ since we can define $Y^*(\tilde{\theta}) = (Y(\tilde{\theta}) - \mu_{Y,0}) / \sqrt{\Sigma_Y(\tilde{\theta})}$ and consider the problem of testing that the mean of $Y^*(\tilde{\theta})$ is zero (transforming the set \mathcal{Y}_m accordingly). After deriving critical values (c_l^*, c_u^*) in this transformed problem, we can recover critical values for our original problem as $(c_l, c_u) = \sqrt{\Sigma_Y(\tilde{\theta})}(c_l^*, c_u^*) + \mu_{Y,0}$. Hence, for the remainder of the proof we assume that $\mu_{Y,0} = 0$ and $\Sigma_Y(\tilde{\theta}) = 1$.

Equal-Tailed Test We consider first the equal-tailed test. Note that this test rejects if and only if $Y(\tilde{\theta}) \notin [c_{l,ET}(\mathcal{Y}), c_{u,ET}(\mathcal{Y})]$, where we suppress the dependence of the critical values on $\mu_{Y,0} = 0$ for simplicity, and $(c_{l,ET}(\mathcal{Y}), c_{u,ET}(\mathcal{Y}))$ solve $F_{TN}(c_{l,ET}(\mathcal{Y}), \mathcal{Y}) = \frac{\alpha}{2}$ and $F_{TN}(c_{u,ET}(\mathcal{Y}), \mathcal{Y}) = 1 - \frac{\alpha}{2}$, for $F_{TN}(\cdot, \mathcal{Y})$ the distribution function of a standard normal random variable truncated to \mathcal{Y} . Recall that we can write the density corresponding to $F_{TN}(y, \mathcal{Y})$ as $\frac{1_{\{y \in \mathcal{Y}\}}}{Pr\{\xi \in \mathcal{Y}\}} f_N(y)$ where f_N is the standard normal density and $Pr\{\xi \in \mathcal{Y}\}$ is the probability that $\xi \in \mathcal{Y}$ for $\xi \sim N(0, 1)$. Hence, we can write $F_{TN}(y, \mathcal{Y}) = \frac{\int_{-\infty}^y 1_{\{\tilde{y} \in \mathcal{Y}\}} f_N(\tilde{y}) d\tilde{y}}{Pr\{\xi \in \mathcal{Y}\}}$.

Note next that for all y we can write $F_{TN}(y, \mathcal{Y}_m) = a_m(y) + F_N(y)$, where F_N is the standard normal distribution function and $a_m(y) = \frac{\int_{-\infty}^y 1_{\{\tilde{y} \in \mathcal{Y}_m\}} f_N(\tilde{y}) d\tilde{y}}{Pr\{\xi \in \mathcal{Y}_m\}} - F_N(y)$. Recall, however, that $Pr\{\xi \in \mathcal{Y}_m\} \rightarrow 1$ and

$$\begin{aligned} & \left| \int_{-\infty}^y 1_{\{\tilde{y} \in \mathcal{Y}_m\}} f_N(\tilde{y}) d\tilde{y} - F_N(y) \right| = \left| \int_{-\infty}^y [1_{\{\tilde{y} \in \mathcal{Y}_m\}} - 1] f_N(\tilde{y}) d\tilde{y} \right| \\ & = \int_{-\infty}^y 1_{\{\tilde{y} \notin \mathcal{Y}_m\}} f_N(\tilde{y}) d\tilde{y} \leq Pr\{\xi \notin \mathcal{Y}_m\} \rightarrow 0 \end{aligned}$$

for all y , so $a_m(y) \rightarrow 0$ for all y . Theorem 2.11 in van der Vaart (1998) then implies that $a_m(y) \rightarrow 0$ uniformly in y as well.

Note next that $F_{TN}(c_{l,ET}(\mathcal{Y}_m), \mathcal{Y}_m) = a_m(c_{l,ET}(\mathcal{Y}_m)) + F_N(c_{l,ET}(\mathcal{Y}_m)) = \frac{\alpha}{2}$ implies $c_{l,ET}(\mathcal{Y}_m) = F_N^{-1}(\frac{\alpha}{2} - a_m(c_{l,ET}(\mathcal{Y}_m)))$, and thus that $c_{l,ET}(\mathcal{Y}_m) \rightarrow F_N^{-1}(\frac{\alpha}{2})$. Using the same

argument, we can show that $c_{u,ET}(\mathcal{Y}_m) \rightarrow F_N^{-1}(1 - \frac{\alpha}{2})$, as desired.

Unbiased Test We next consider the unbiased test. Recall that critical values $c_{l,U}(\mathcal{Y}), c_{u,U}(\mathcal{Y})$ for the unbiased test solve $Pr\{\zeta \in [c_{l,U}(\mathcal{Y}), c_{u,U}(\mathcal{Y})]\} = 1 - \alpha$ and $E[\zeta 1\{\zeta \in [c_{l,U}(\mathcal{Y}), c_{u,U}(\mathcal{Y})]\}] = (1 - \alpha)E[\zeta]$ for $\zeta \sim \xi | \xi \in \mathcal{Y}$ where $\xi \sim N(0,1)$.

Note that for ζ_m the truncated normal random variable corresponding to \mathcal{Y}_m , we can write $Pr\{\zeta_m \in [c_l, c_u]\} = a_m(c_l, c_u) + (F_N(c_u) - F_N(c_l))$ with

$$a_m(c_l, c_u) = (F_N(c_l) - Pr\{\zeta_m \leq c_l\}) - (F_N(c_u) - Pr\{\zeta_m \leq c_u\}).$$

As in the argument for equal-tailed tests above, we see that both $F_N(c_u) - Pr\{\zeta_m \leq c_u\}$ and $F_N(c_l) - Pr\{\zeta_m \leq c_l\}$ converge to zero pointwise, and thus uniformly in c_u and c_l by Theorem 2.11 in van der Vaart (1998). Hence, $a_m(c_l, c_u) \rightarrow 0$ uniformly in (c_l, c_u) .

Note, next, that we can write $E[\zeta_m 1\{\zeta_m \in [c_l, c_u]\}] = [E[\xi 1\{\xi \in [c_l, c_u]\}]] + b_m(c_l, c_u)$ for

$$b_m(c_l, c_u) = E[\zeta_m 1\{\zeta_m \in [c_l, c_u]\}] - [E[\xi 1\{\xi \in [c_l, c_u]\}]] = \int_{c_l}^{c_u} \left(\frac{1\{y \in \mathcal{Y}_m\}}{Pr\{\xi \in \mathcal{Y}_m\}} - 1 \right) y f_N(y) dy.$$

Note, however, that $\int_{c_l}^{c_u} (1\{y \in \mathcal{Y}_m\} - 1) y f_N(y) dy \leq E[|\xi| 1\{\xi \notin \mathcal{Y}_m\}]$. Hence, since

$$\begin{aligned} & \left| \int_{c_l}^{c_u} \left(\frac{1\{y \in \mathcal{Y}_m\}}{Pr\{\xi \in \mathcal{Y}_m\}} - 1 \right) y f_N(y) dy \right| \\ & \leq \left| \int_{c_l}^{c_u} (1\{y \in \mathcal{Y}_m\} - 1) y f_N(y) dy \right| + \left| \int_{c_l}^{c_u} \left(\frac{1\{y \in \mathcal{Y}_m\}}{Pr\{\xi \in \mathcal{Y}_m\}} - 1\{y \in \mathcal{Y}_m\} \right) y f_N(y) dy \right| \\ & \leq E[|\xi| 1\{\xi \notin \mathcal{Y}_m\}] + \left| \left(\frac{1}{Pr\{\xi \in \mathcal{Y}_m\}} - 1 \right) \int_{c_l}^{c_u} 1\{y \in \mathcal{Y}_m\} |y| f_N(y) dy \right| \\ & \leq \sqrt{P(\xi \notin \mathcal{Y}_m)} + \left| \left(\frac{1}{Pr\{\xi \in \mathcal{Y}_m\}} - 1 \right) \right| E[|\xi|] \end{aligned}$$

by the Cauchy-Schwartz Inequality, where the right hand side tends to zero and doesn't depend on (c_l, c_u) , $b_m(c_l, c_u)$ converges to zero uniformly in (c_l, c_u) .

Next, let us define $(c_{l,m}, c_{u,m})$ as the solutions to $Pr\{\zeta_m \in [c_{l,m}, c_{u,m}]\} = 1 - \alpha$ and $E[\zeta_m 1\{\zeta_m \in [c_{l,m}, c_{u,m}]\}] = (1 - \alpha)E[\zeta_m]$. From our results above, we can re-write the problem solved by $(c_{l,m}, c_{u,m})$ as $F_N(c_u) - F_N(c_l) = 1 - \alpha - a_m(c_l, c_u)$, $E[\xi 1\{\xi \in [c_l, c_u]\}] = (1 - \alpha)E[\zeta_m] - b_m(c_l, c_u)$. Letting $\bar{a}_m = \sup_{c_l, c_u} |a_m(c_l, c_u)|$, and $\bar{b}_m = \sup_{c_l, c_u} |b_m(c_l, c_u)|$ we thus see that $(c_{l,m}, c_{u,m})$ solves $F_N(c_u) - F_N(c_l) = 1 - \alpha - a_m^*$ and $E[\xi 1\{\xi \in [c_l, c_u]\}] = (1 - \alpha)E[\zeta_m] - b_m^*$ for some $a_m^* \in [-\bar{a}_m, \bar{a}_m]$, $b_m^* \in [-\bar{b}_m, \bar{b}_m]$. We will next show that for

any sequence of values (a_m^*, b_m^*) such that $a_m^* \in [-\bar{a}_m, \bar{a}_m]$ and $b_m^* \in [-\bar{b}_m, \bar{b}_m]$ for all m , the implied solutions $c_{l,m}(a_m^*, b_m^*)$, $c_{u,m}(a_m^*, b_m^*)$ converge to $F_N^{-1}(\frac{\alpha}{2})$ and $F_N^{-1}(1 - \frac{\alpha}{2})$. This follows from the next lemma, which is proved below.

Lemma 4

Suppose that $c_{l,m}$ and $c_{u,m}$ solve $Pr\{\xi \in [c_l, c_u]\} = 1 - \alpha + a_m$ and $E[\xi 1\{\xi \in [c_l, c_u]\}] = d_m$ for $a_m, d_m \rightarrow 0$. Then $(c_{l,m}, c_{u,m}) \rightarrow (-c_{\frac{\alpha}{2}, N}, c_{\frac{\alpha}{2}, N})$.

Using this lemma, since $E[\zeta_m] \rightarrow 0$ as $m \rightarrow \infty$ we see that for any sequence of values $(a_m^*, b_m^*) \rightarrow 0$, $(c_{l,m}(a_m^*, b_m^*), c_{u,m}(a_m^*, b_m^*)) \rightarrow (-c_{\frac{\alpha}{2}, N}, c_{\frac{\alpha}{2}, N})$. However, since $\bar{a}_m, \bar{b}_m \rightarrow 0$ we know that the values a_m^* and b_m^* corresponding to the true $c_{l,m}, c_{u,m}$ must converge to zero. Hence $(c_{l,m}, c_{u,m}) \rightarrow (-c_{\frac{\alpha}{2}, N}, c_{\frac{\alpha}{2}, N})$ as we wanted to show. \square

Proof of Lemma 4 Note that the critical values solve

$$f(a_m, d_m, c) = \begin{pmatrix} F_N(c_u) - F_N(c_l) - (1 - \alpha) - a_m \\ \int_{c_l}^{c_u} y f_N(y) dy - d_m \end{pmatrix} = 0.$$

We can simplify this expression, since $\frac{\partial}{\partial y} f_N(y) = -y f_N(y)$, so $\int_{c_l}^{c_u} y f_N(y) dy = f_N(c_l) - f_N(c_u)$.

We thus must solve the system of equations $g(c) - v_m = 0$, for

$$g(c) = \begin{pmatrix} F_N(c_u) - F_N(c_l) \\ f_N(c_l) - f_N(c_u) \end{pmatrix}, \quad v_m = \begin{pmatrix} a_m + (1 - \alpha) \\ d_m \end{pmatrix}.$$

Note that for $v_m = (1 - \alpha, 0)'$ this system is solved by $c = (-c_{\frac{\alpha}{2}, N}, c_{\frac{\alpha}{2}, N})$. Further,

$$\frac{\partial}{\partial c} g(c) = \begin{pmatrix} -f_N(c_l) & f_N(c_u) \\ -c_l f_N(c_l) & c_u f_N(c_u) \end{pmatrix},$$

which evaluated at $c = (-c_{\frac{\alpha}{2}, N}, c_{\frac{\alpha}{2}, N})$ is equal to

$$\begin{pmatrix} -f_N(c_{\frac{\alpha}{2}, N}) & f_N(c_{\frac{\alpha}{2}, N}) \\ c_{\frac{\alpha}{2}, N} f_N(c_{\frac{\alpha}{2}, N}) & c_{\frac{\alpha}{2}, N} f_N(c_{\frac{\alpha}{2}, N}) \end{pmatrix}$$

and has full rank for all $\alpha \in (0, 1)$. Thus, by the implicit function theorem there exists an open neighborhood V of $v_\infty = (1 - \alpha, 0)$ such that $g(c) - v = 0$ has a unique solution $c(v)$ for $v \in V$ and $c(v)$ is continuously differentiable. Hence, if we consider any sequence of

values $v_m \rightarrow (1-\alpha, 0)$, we see that $c(v_m) \rightarrow \begin{pmatrix} -c_{\frac{\alpha}{2}, N} \\ c_{\frac{\alpha}{2}, N} \end{pmatrix}$, again as we wanted to show. \square

D.2 Proofs for Results in Main Text

Proof of Proposition 1 Let us assume without loss of generality that $\tilde{\theta} = \theta_1$. Note that the conditioning event $\{\max_{\theta \in \Theta} X(\theta) = X(\theta_1)\}$ is equivalent to $\{MX \geq 0\}$, where

$$M \equiv \begin{pmatrix} 1 & -1 & 0 & 0 & \dots & 0 \\ 1 & 0 & -1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \dots & -1 \end{pmatrix}$$

is a $(|\Theta|-1) \times |\Theta|$ matrix and the inequality is taken element-wise. Let $A = \begin{bmatrix} -M & 0_{(|\Theta|-1) \times |\Theta|} \end{bmatrix}$, where $0_{(|\Theta|-1) \times |\Theta|}$ denotes the $(|\Theta|-1) \times |\Theta|$ matrix of zeros. Let $W = (X', Y)'$ and note that we can re-write the event of interest as $\{W : AW \leq 0\}$ and that we are interested in inference on $\eta' \mu$ for η the $2|\Theta| \times 1$ vector with one in the $(|\Theta|+1)$ st entry and zeros everywhere else. Define

$$\tilde{Z}_{\tilde{\theta}}^* = W - cY(\tilde{\theta}),$$

for $c = Cov(W, Y(\tilde{\theta})) / \Sigma_Y(\tilde{\theta})$, noting that the definition of $Z_{\tilde{\theta}}$ in (10) of the main text corresponds to extracting the elements of $\tilde{Z}_{\tilde{\theta}}^*$ corresponding to X . By Lemma 5.1 of Lee et al. (2016),

$$\{W : AW \leq 0\} = \left\{ W : \mathcal{L}(\tilde{\theta}, \tilde{Z}_{\tilde{\theta}}^*) \leq Y(\tilde{\theta}) \leq \mathcal{U}(\tilde{\theta}, \tilde{Z}_{\tilde{\theta}}^*), \mathcal{V}(\tilde{\theta}, \tilde{Z}_{\tilde{\theta}}^*) \geq 0 \right\},$$

where for $(v)_j$ the j th element of a vector v ,

$$\mathcal{L}(\tilde{\theta}, z) = \max_{j: (Ac)_j < 0} \frac{-(Az)_j}{(Ac)_j}$$

$$\mathcal{U}(\tilde{\theta}, z) = \min_{j: (Ac)_j > 0} \frac{-(Az)_j}{(Ac)_j}$$

$$\mathcal{V}(\tilde{\theta}, z) = \min_{j: (Ac)_j = 0} -(Az)_j.$$

Note, however, that

$$\left(A\tilde{Z}_{\tilde{\theta}}^*\right)_j = Z_{\tilde{\theta}}(\theta_j) - Z_{\tilde{\theta}}(\theta_1)$$

and

$$(Ac)_j = -\frac{\Sigma_{XY}(\theta_1, \theta_1) - \Sigma_{XY}(\theta_1, \theta_j)}{\Sigma_Y(\theta_1)}.$$

Hence, we can re-write

$$\frac{-(A\tilde{Z}_{\tilde{\theta}}^*)_j}{(Ac)_j} = \frac{\Sigma_Y(\theta_1)(Z_{\tilde{\theta}}(\theta_j) - Z_{\tilde{\theta}}(\theta_1))}{\Sigma_{XY}(\theta_1, \theta_1) - \Sigma_{XY}(\theta_1, \theta_j)},$$

$$\begin{aligned}\mathcal{L}(\tilde{\theta}, \tilde{Z}_{\tilde{\theta}}^*) &= \max_{j: \Sigma_{XY}(\theta_1, \theta_1) > \Sigma_{XY}(\theta_1, \theta_j)} \frac{\Sigma_Y(\theta_1)(Z_{\tilde{\theta}}(\theta_j) - Z_{\tilde{\theta}}(\theta_1))}{\Sigma_{XY}(\theta_1, \theta_1) - \Sigma_{XY}(\theta_1, \theta_j)}, \\ \mathcal{U}(\tilde{\theta}, \tilde{Z}_{\tilde{\theta}}^*) &= \min_{j: \Sigma_{XY}(\theta_1, \theta_1) < \Sigma_{XY}(\theta_1, \theta_j)} \frac{\Sigma_Y(\theta_1)(Z_{\tilde{\theta}}(\theta_j) - Z_{\tilde{\theta}}(\theta_1))}{\Sigma_{XY}(\theta_1, \theta_1) - \Sigma_{XY}(\theta_1, \theta_j)},\end{aligned}$$

and

$$\mathcal{V}(\tilde{\theta}, \tilde{Z}_{\tilde{\theta}}^*) = \min_{j: \Sigma_{XY}(\theta_1, \theta_1) = \Sigma_{XY}(\theta_1, \theta_j)} -(Z_{\tilde{\theta}}(\theta_j) - Z_{\tilde{\theta}}(\theta_1)).$$

Note, however, that these are functions of $Z_{\tilde{\theta}}$, as expected. The result follows. \square

Proof of Proposition 2 Follows as a special case of Proposition 7. \square

Proof of Proposition 3 Follows as a special case of Proposition 9. \square

Proof of Proposition 4 We prove the result for coverage, while the result for median unbiasedness is analogous. Provided $\hat{\theta}$ is unique with probability one, we can write

$$Pr_{\mu}\left\{\mu(\hat{\theta}) \in CI\right\} = \sum_{\tilde{\theta} \in \Theta} Pr_{\mu}\left\{\hat{\theta} = \tilde{\theta}\right\} Pr_{\mu}\left\{\mu(\tilde{\theta}) \in CI \mid \hat{\theta} = \tilde{\theta}\right\}.$$

Since $\sum_{\tilde{\theta} \in \Theta} Pr_{\mu}\left\{\hat{\theta} = \tilde{\theta}\right\} = 1$, the result of the proposition follows immediately. \square

Proof of Lemma 1 The assumption of the lemma implies that $X(\tilde{\theta}) - X(\theta)$ has a non-degenerate normal distribution for all μ . Since Θ is finite, almost-sure uniqueness of $\hat{\theta}$ follows immediately. \square

Proof of Proposition 5 We first establish uniqueness of $\hat{\mu}_{\alpha}^H$. To do so, it suffices to show that $F_{TN}^H(Y(\tilde{\theta}); \mu_Y(\tilde{\theta}), \tilde{\theta}, Z_{\tilde{\theta}})$ is strictly decreasing in $\mu_Y(\tilde{\theta})$. Note first that this holds for the truncated normal assuming truncation that does not depend on $\mu_Y(\tilde{\theta})$ by Lemma A.1 of

Lee et al. (2016). When we instead consider $F_{TN}^H(Y(\tilde{\theta}); \mu_Y(\tilde{\theta}), \tilde{\theta}, Z_{\tilde{\theta}})$, we further truncate to

$$Y(\tilde{\theta}) \in \left[\mu_Y(\tilde{\theta}) - c_\beta \sqrt{\Sigma_Y(\tilde{\theta})}, \mu_Y(\tilde{\theta}) + c_\beta \sqrt{\Sigma_Y(\tilde{\theta})} \right].$$

Since this interval shifts upwards as we increase $\mu_Y(\tilde{\theta})$, $F_{TN}^H(Y(\tilde{\theta}); \mu_Y(\tilde{\theta}), \tilde{\theta}, Z_{\tilde{\theta}})$ is a fortiori decreasing in $\mu_Y(\tilde{\theta})$. Uniqueness of $\hat{\mu}_\alpha^H$ for $\alpha \in (0, 1)$ follows. Note, next, that $F_{TN}^H(Y(\tilde{\theta}); \mu_Y(\tilde{\theta}), \tilde{\theta}, Z_{\tilde{\theta}}) \in \{0, 1\}$ for $\mu_Y(\tilde{\theta}) \notin CI_P^\beta$ from which we immediately see that $\hat{\mu}_\alpha^H \in CI_P^\beta$.

Finally, note that for $\mu_Y(\tilde{\theta})$ the true value, $F_{TN}^H(Y(\hat{\theta}); \mu_Y(\tilde{\theta}), \tilde{\theta}, Z_{\tilde{\theta}}) \sim U[0, 1]$ conditional on $\{\hat{\theta} = \tilde{\theta}, Z_{\hat{\theta}} = z_{\tilde{\theta}}, \mu_Y(\tilde{\theta}) \in CI_P^\beta\}$. Since $F_{TN}^H(Y(\hat{\theta}); \mu_Y(\tilde{\theta}), \tilde{\theta}, Z_{\tilde{\theta}})$ is decreasing in $\mu_Y(\tilde{\theta})$,

$$\begin{aligned} & Pr_\mu \left\{ \hat{\mu}_\alpha^H \geq \mu_Y(\tilde{\theta}) \mid \hat{\theta} = \tilde{\theta}, Z_{\hat{\theta}} = z_{\tilde{\theta}}, \mu_Y(\tilde{\theta}) \in CI_P^\beta \right\} \\ &= Pr_\mu \left\{ F_{TN}^H(Y(\hat{\theta}); \mu_Y(\tilde{\theta}), \tilde{\theta}, \tilde{\gamma}, Z_{\tilde{\theta}}) \geq 1 - \alpha \mid \hat{\theta} = \tilde{\theta}, Z_{\hat{\theta}} = z_{\tilde{\theta}}, \mu_Y(\tilde{\theta}) \in CI_P^\beta \right\} = \alpha, \end{aligned}$$

and thus $\hat{\mu}_\alpha^H$ is α -quantile-unbiased conditional on $\{\hat{\theta} = \tilde{\theta}, Z_{\hat{\theta}} = z_{\tilde{\theta}}, \mu_Y(\tilde{\theta}) \in CI_P^\beta\}$. We can drop the conditioning on $Z_{\tilde{\theta}}$ by the law of iterated expectations, and α -quantile unbiasedness conditional on $\mu_Y(\tilde{\theta}) \in CI_P^\beta$ follows by the same argument as in the proof of Proposition 4.

Proof of Proposition 6 The first part of the proposition follows immediately from Proposition 5. For the second part of the proposition, note that

$$\begin{aligned} & Pr_\mu \left\{ \mu_Y(\hat{\theta}) \in CI_{ET}^H \right\} = Pr_\mu \left\{ \mu_Y(\hat{\theta}) \in CI_P^\beta \right\} \times \\ & \sum_{\tilde{\theta} \in \Theta} Pr_\mu \left\{ \hat{\theta} = \tilde{\theta} \mid \mu_Y(\hat{\theta}) \in CI_P^\beta \right\} Pr_\mu \left\{ \mu_Y(\tilde{\theta}) \in CI_{ET}^H \mid \hat{\theta} = \tilde{\theta}, \mu_Y(\tilde{\theta}) \in CI_P^\beta \right\} \\ &= Pr_\mu \left\{ \mu_Y(\hat{\theta}) \in CI_P^\beta \right\} \frac{1 - \alpha}{1 - \beta} \geq (1 - \beta) \frac{1 - \alpha}{1 - \beta} = 1 - \alpha, \end{aligned}$$

where the second equality follows from the first part of the proposition. The upper bound follows by the same argument and the fact that $Pr_\mu \left\{ \mu_Y(\hat{\theta}) \in CI_P^\beta \right\} \leq 1$. \square

E Forecast Intervals

This appendix provides additional details on the forecast intervals discussed in the main text. Let $Z_{1-2, \tilde{\theta}} = X - \frac{\Sigma_{XY}(\cdot, \tilde{\theta})}{\Sigma_{Y_{1-2}}(\tilde{\theta})} Y_{1-2}(\tilde{\theta})$ denote the analog of $Z_{\tilde{\theta}}$ in the main text which uses Y_{1-2} in place of Y . The same argument as in Section 4 of the main text implies

that conditional on $\hat{\theta}=\tilde{\theta}$ and $Z_{1-2,\tilde{\theta}}=z$, for $\Sigma_{Y_{1-2}}=\Sigma_Y+\Sigma_{Y_2}$, $Y_{1-2}(\tilde{\theta})$ is distributed as a $N(0,\Sigma_{Y_{1-2}}(\tilde{\theta}))$ variable truncated to $\mathcal{Y}_{1-2}(\tilde{\theta},z)=[\mathcal{L}_{1-2}(\tilde{\theta},z),\mathcal{U}_{1-2}(\tilde{\theta},z)]$, where

$$\mathcal{L}_{1-2}(\tilde{\theta},z)=\max_{\theta\in\Theta:\Sigma_{XY}(\tilde{\theta})>\Sigma_{XY}(\tilde{\theta},\theta)}\frac{\Sigma_{Y_{1-2}}(\tilde{\theta})(z(\theta)-z(\tilde{\theta}))}{\Sigma_{XY}(\tilde{\theta})-\Sigma_{XY}(\tilde{\theta},\theta)},$$

$$\mathcal{U}_{1-2}(\tilde{\theta},z)=\min_{\theta\in\Theta:\Sigma_{XY}(\tilde{\theta})<\Sigma_{XY}(\tilde{\theta},\theta)}\frac{\Sigma_{Y_{1-2}}(\tilde{\theta})(z(\theta)-z(\tilde{\theta}))}{\Sigma_{XY}(\tilde{\theta})-\Sigma_{XY}(\tilde{\theta},\theta)}.$$

The interval CI_{ET}^{1-2} fails to cover zero only when $Y_{1-2}(\hat{\theta})$ lies in the tails of this conditional distribution. Letting $q_\alpha(\tilde{\theta},z)$ denote the α quantile of the truncated normal distribution, the resulting *conditional* forecast interval for $Y_2(\hat{\theta})$ is thus

$$FI=\{y_2:Y(\hat{\theta})-q_{1-\alpha/2}(\hat{\theta},Z_{1-2,\hat{\theta}}(y_2))\leq y_2\leq Y(\hat{\theta})-q_{\alpha/2}(\hat{\theta},Z_{1-2,\hat{\theta}}(y_2))\},$$

where with a slight abuse of notation $Z_{1-2,\tilde{\theta}}(y_2)=X-\frac{\Sigma_{XY}(\cdot,\tilde{\theta})}{\Sigma_{Y_{1-2}}(\tilde{\theta})}(Y(\tilde{\theta})-y_2)$. To see why this forecast interval has correct coverage conditional on any realization of $\hat{\theta}$, note that

$$\begin{aligned} &Pr_\mu\left\{Y_2(\hat{\theta})\in FI|\hat{\theta}=\tilde{\theta},Z_{1-2,\hat{\theta}}=z\right\} \\ &=Pr_\mu\left\{Y(\hat{\theta})-q_{1-\alpha/2}(\hat{\theta},Z_{1-2,\hat{\theta}}(Y_2(\hat{\theta})))\leq Y_2(\hat{\theta})\leq Y(\hat{\theta})-q_{\alpha/2}(\hat{\theta},Z_{1-2,\hat{\theta}}(Y_2(\hat{\theta})))|\hat{\theta}=\tilde{\theta},Z_{1-2,\hat{\theta}}=z\right\} \\ &=Pr_\mu\left\{q_{\alpha/2}(\tilde{\theta},z)\leq Y_{1-2}(\tilde{\theta})\leq q_{1-\alpha/2}(\tilde{\theta},z)|\hat{\theta}=\tilde{\theta},Z_{1-2,\hat{\theta}}=z\right\}=1-\alpha \end{aligned}$$

since $Z_{1-2,\tilde{\theta}}(Y_2(\tilde{\theta}))=Z_{1-2,\tilde{\theta}}$. We can construct conditional upper and lower one-sided forecast intervals analogously.

We can similarly construct *unconditional* forecast intervals for $Y_2(\hat{\theta})$ based on our hybrid approach. Let $q_\alpha^H(\tilde{\theta},\beta,z)$ denote the α -quantile of $Y_{1-2}(\tilde{\theta})$ truncated to $\mathcal{Y}^H(\tilde{\theta},\beta,z)=[\mathcal{L}_{1-2}^H(\tilde{\theta},\beta,z),\mathcal{U}_{1-2}^H(\tilde{\theta},\beta,z)]$, where

$$\mathcal{L}_{1-2}^H(\tilde{\theta},\beta,z)=\max\left\{-c_\beta^{1-2}\sqrt{\Sigma_{Y_{1-2}}(\tilde{\theta})},\mathcal{L}_{1-2}(\tilde{\theta},z)\right\}$$

$$\mathcal{U}_{1-2}^H(\tilde{\theta},\beta,z)=\min\left\{c_\beta^{1-2}\sqrt{\Sigma_{Y_{1-2}}(\tilde{\theta})},\mathcal{U}_{1-2}(\tilde{\theta},z)\right\}$$

for c_β^{1-2} the $1-\beta$ quantile of $\max_\theta|Y_{1-2}(\theta)|/\sqrt{\Sigma_{Y_{1-2}}(\theta)}$ (i.e. the projection critical value).

The unconditional forecast interval is then

$$FI^H = \left\{ y_2 : Y(\hat{\theta}) - q_{1-\frac{\alpha-\beta}{2(1-\beta)}}^H(\hat{\theta}, \beta, Z_{1-2, \hat{\theta}}(y_2)) \leq y_2 \leq Y(\hat{\theta}) - q_{\frac{\alpha-\beta}{2(1-\beta)}}^H(\hat{\theta}, \beta, Z_{1-2, \hat{\theta}}(y_2)) \right\}.$$

To see why this forecast interval has correct unconditional coverage, note that

$$\begin{aligned} & Pr_{\mu} \left\{ Y_2(\hat{\theta}) \in FI^H \right\} \\ &= Pr_{\mu} \left\{ Y_2(\hat{\theta}) \in FI^H \mid -c_{\beta}^{1-2} \sqrt{\Sigma_{Y_{1-2}}(\hat{\theta})} \leq Y_{1-2}(\hat{\theta}) \leq c_{\beta}^{1-2} \sqrt{\Sigma_{Y_{1-2}}(\hat{\theta})} \right\} \\ &\quad \times Pr_{\mu} \left\{ -c_{\beta}^{1-2} \sqrt{\Sigma_{Y_{1-2}}(\hat{\theta})} \leq Y_{1-2}(\hat{\theta}) \leq c_{\beta}^{1-2} \sqrt{\Sigma_{Y_{1-2}}(\hat{\theta})} \right\} \\ &= Pr_{\mu} \left\{ Y(\hat{\theta}) - q_{1-\frac{\alpha-\beta}{2(1-\beta)}}^H(\hat{\theta}, \beta, Z_{1-2, \hat{\theta}}(Y_2(\hat{\theta}))) \leq Y_2(\hat{\theta}) \leq Y(\hat{\theta}) - q_{\frac{\alpha-\beta}{2(1-\beta)}}^H(\hat{\theta}, \beta, Z_{1-2, \hat{\theta}}(Y_2(\hat{\theta}))) \right. \\ &\quad \left. \mid -c_{\beta}^{1-2} \sqrt{\Sigma_{Y_{1-2}}(\hat{\theta})} \leq Y_{1-2}(\hat{\theta}) \leq c_{\beta}^{1-2} \sqrt{\Sigma_{Y_{1-2}}(\hat{\theta})} \right\} \\ &\quad \times Pr_{\mu} \left\{ -c_{\beta}^{1-2} \sqrt{\Sigma_{Y_{1-2}}(\hat{\theta})} \leq Y_{1-2}(\hat{\theta}) \leq c_{\beta}^{1-2} \sqrt{\Sigma_{Y_{1-2}}(\hat{\theta})} \right\} \\ &\geq Pr_{\mu} \left\{ q_{\frac{\alpha-\beta}{2(1-\beta)}}^H(\hat{\theta}, \beta, Z_{1-2, \hat{\theta}}) \leq Y_{1-2}(\hat{\theta}) \leq q_{1-\frac{\alpha-\beta}{2(1-\beta)}}^H(\hat{\theta}, \beta, Z_{1-2, \hat{\theta}}) \right. \\ &\quad \left. \mid -c_{\beta}^{1-2} \sqrt{\Sigma_{Y_{1-2}}(\hat{\theta})} \leq Y_{1-2}(\hat{\theta}) \leq c_{\beta}^{1-2} \sqrt{\Sigma_{Y_{1-2}}(\hat{\theta})} \right\} \times (1-\beta) \\ &= \left(1 - \frac{\alpha-\beta}{1-\beta} \right) (1-\beta) = 1-\alpha, \end{aligned}$$

where the inequality follows from the fact that

$$Pr_{\mu} \left\{ -c_{\beta}^{1-2} \sqrt{\Sigma_{Y_{1-2}}(\hat{\theta})} \leq Y_{1-2}(\hat{\theta}) \leq c_{\beta}^{1-2} \sqrt{\Sigma_{Y_{1-2}}(\hat{\theta})} \right\} \geq 1-\beta.$$

F Uniform Asymptotic Validity

This section establishes uniform asymptotic validity for plug-in versions of the procedures discussed in the main text. One could use arguments along the same lines as those below to derive results for additional conditioning variables $\hat{\gamma}_n$, but since such arguments would be case-specific, we do not pursue such an extension here.

Feasible finite-sample estimators and confidence intervals are denoted as their coun-

terparts in Sections 4–5 of the main text, with the addition of an n subscript. We suppose that the sample of size n is drawn from some (unknown) distribution $P \in \mathcal{P}_n$. To simplify repetitive notation we work with scaled estimates (X_n, Y_n) which correspond to $\sqrt{n} \cdot (\tilde{X}_n, \tilde{Y}_n)$ for $(\tilde{X}_n, \tilde{Y}_n)$ as discussed in Section 6 of the main text. Similarly, we work with a variance estimator $\hat{\Sigma}_n = n \cdot \tilde{\Sigma}_n$ for $\tilde{\Sigma}_n$ as discussed in the main text.

We first impose that (X_n, Y_n) are uniformly asymptotically normal under $P \in \mathcal{P}_n$, where the centering vectors $(\mu_{X,n}, \mu_{Y,n})$ and the limiting variance Σ may depend on P .

Assumption 6

For the class of Lipschitz functions that are bounded in absolute value by one and have Lipschitz constant bounded by one, BL_1 , there exist sequences of functions $\mu_{X,n}(P)$ and $\mu_{Y,n}(P)$ and a function $\Sigma(P)$ such that for $\xi_P \sim N(0, \Sigma(P))$,

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} \sup_{f \in BL_1} \left| E_P \left[f \begin{pmatrix} X_n - \mu_{X,n}(P) \\ Y_n - \mu_{Y,n}(P) \end{pmatrix} \right] - E[f(\xi_P)] \right| = 0.$$

Uniform convergence in bounded Lipschitz metric is one formalization for uniform convergence in distribution. When X_n and Y_n are scaled sample averages based on independent data, as in Section 2 of the main text, Assumption 6 will follow from moment bounds, while for dependent data it will follow from moment and dependence bounds.

We next assume that the asymptotic variance is uniformly consistently estimable.

Assumption 7

The estimator $\hat{\Sigma}_n$ is uniformly consistent in the sense that for all $\varepsilon > 0$

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} Pr_P \left\{ \left\| \hat{\Sigma}_n - \Sigma(P) \right\| > \varepsilon \right\} = 0.$$

Provided we use a variance estimator appropriate to the setting (e.g. the sample variance for iid data, long-run variance estimators for time series, and so on) Assumption 7 will follow from the same sorts of sufficient conditions as for Assumption 6.

Finally, we restrict the asymptotic variance.

Assumption 8

There exists a finite $\bar{\lambda} > 0$ such that

$$1/\bar{\lambda} \leq \Sigma_X(\theta; P), \Sigma_Y(\theta; P) \leq \bar{\lambda}, \text{ for all } \theta \in \Theta \text{ and all } P \in \mathcal{P}_n,$$

$$1/\bar{\lambda} \leq \sqrt{\Sigma_X(\theta; P)\Sigma_X(\tilde{\theta}; P) - \Sigma_X(\theta, \tilde{\theta}; P)} \text{ for all } \theta, \tilde{\theta} \in \Theta \text{ with } \theta \neq \tilde{\theta} \text{ and all } P \in \mathcal{P}_n.$$

The upper bounds on $\Sigma_X(\theta; P)$ and $\Sigma_Y(\theta; P)$ ensure that the random variables ξ_P in Assumption 6 are stochastically bounded, while the lower bounds ensure that each entry (X_n, Y_n) has a nonzero asymptotic variance. The assumption of nonzero variance rules out the case where one element of X_n is a non-random threshold (as discussed in Section 3 of main text), but our asymptotic results can be extended to cover this case at the cost of additional notation. The second condition ensures that no two elements of X_n are perfectly (positively) correlated asymptotically, and hence, by Lemma 1, guarantees that $\hat{\theta}_n$ is unique with probability tending to one. Note that this condition is weaker than a standard assumption bounding the eigenvalues of $\Sigma_X(P)$ away from zero.

High-Dimensional Settings Our asymptotic analysis considers settings where $|\Theta|$, and hence the dimension of X_n and Y_n , are fixed as $n \rightarrow \infty$. One might also be interested in settings where $|\Theta|$ grows with n , but this will raise complications for both the normal approximation and estimation of the asymptotic variance. Such an extension is interesting, but beyond the scope of this paper.

Variance Estimation Practically, even for fixed $|\Theta|$ one might still worry about the difficulty of estimating Σ in finite samples, since this matrix has $|\Theta|(|\Theta| + 1)/2$ entries. Fortunately, in many cases Σ has additional structure which renders variance estimation more tractable than in the fully general case. Suppose, for instance, that we want to conduct inference on the best-performing treatment from a randomized trial, as in Section 2 of the main text and Section H below. In this case, provided trial participants are drawn independently, elements of $X_n(\theta)$ corresponding to distinct treatments are uncorrelated and Σ is diagonal. In other cases, such as Section 7 of the main text, $|\Theta|$ may be large, but the elements of X_n are formed by taking combinations of a much lower-dimensional set of random variables. In this case, Σ_X can be written as a known linear transformation of a much lower-dimensional variance matrix.

F.1 Uniform Asymptotic Validity

In the finite-sample normal model, we study both conditional and unconditional properties of our methods. We would like to do the same in our asymptotic analysis, but may have $Pr\{\hat{\theta}_n = \tilde{\theta}\} \rightarrow 0$ for some $\tilde{\theta}$, in which case conditioning on $\hat{\theta}_n = \tilde{\theta}$ is problematic. To address this, we multiply conditional statements by the probability of the conditioning event.

Asymptotic uniformity results for conditional inference procedures were established by Tibshirani et al. (2018) and Andrews, Kitagawa, and McCloskey (2021) for settings where the target parameter is chosen in other ways. Their results, however, limit attention to

classes of data generating processes with asymptotically bounded means $(\mu_{X,n}, \mu_{Y,n})$. This rules out e.g. the conventional pointwise asymptotic case that fixes P and takes $n \rightarrow \infty$. We do not require such boundedness. Moreover, the results of Tibshirani et al. (2018) do not cover quantile-unbiased estimation, and also do not cover hybrid procedures, which are new to the literature.³

Our proofs are based on subsequencing arguments as in D. Andrews, Cheng, and Guggenberger (2020), though due to the differences in our setting (our interest in conditional inference, and the fact that our target is random from an unconditional perspective) we cannot directly apply their results. We first establish the asymptotic validity of our quantile-unbiased estimators.

Proposition 10

Under Assumptions 6-8, for $\hat{\mu}_{\alpha,n}$ the α -quantile unbiased estimator,

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} \left| Pr_P \left\{ \hat{\mu}_{\alpha,n} \geq \mu_{Y,n}(\hat{\theta}_n; P) \mid \hat{\theta}_n = \tilde{\theta} \right\} - \alpha \right| Pr_P \left\{ \hat{\theta}_n = \tilde{\theta} \right\} = 0, \quad (\text{A.20})$$

for all $\tilde{\theta} \in \Theta$, and

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} \left| Pr_P \left\{ \hat{\mu}_{\alpha,n} \geq \mu_{Y,n}(\hat{\theta}_n; P) \right\} - \alpha \right| = 0. \quad (\text{A.21})$$

This immediately implies asymptotic validity of equal-tailed confidence intervals.

Corollary 1

Under Assumptions 6-8, for $CI_{ET,n}$ the level $1-\alpha$ equal-tailed confidence interval

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} \left| Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_{ET,n} \mid \hat{\theta}_n = \tilde{\theta} \right\} - (1-\alpha) \right| Pr_P \left\{ \hat{\theta}_n = \tilde{\theta} \right\} = 0,$$

for all $\tilde{\theta} \in \Theta$, and

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} \left| Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_{ET,n} \right\} - (1-\alpha) \right| = 0.$$

We can likewise establish uniform asymptotic validity of projection confidence intervals.

Proposition 11

³In a follow-up paper, Andrews, Kitagawa, and McCloskey (2021), we apply the conditional and hybrid approaches developed here to settings where $\hat{\theta} = \operatorname{argmax} \|X(\theta)\|$.

Under Assumptions 6-8, for $CI_{P,n}$ the level $1-\alpha$ projection confidence interval,

$$\liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_n} Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_{P,n} \right\} \geq 1-\alpha. \quad (\text{A.22})$$

To state results for hybrid estimators and confidence intervals, let $C_n^H(\tilde{\theta}; P) = 1 \left\{ \hat{\theta}_n = \tilde{\theta}, \mu_{Y,n}(\hat{\theta}_n; P) \in CI_{P,n}^\beta \right\}$ be an indicator for the hybrid conditioning event that $\hat{\theta}_n$ is equal to $\tilde{\theta}$ and the parameter of interest $\mu_Y(\tilde{\theta})$ falls in the level β projection confidence interval $CI_{P,n}^\beta$. We can establish quantile unbiasedness of hybrid estimators given this event, along with bounded unconditional bias.

Proposition 12

Under Assumptions 6-8, for $\hat{\mu}_{\alpha,n}^H$ the α -quantile unbiased hybrid estimator based on $CI_{P,n}^\beta$,

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} \left| Pr_P \left\{ \hat{\mu}_{\alpha,n}^H \geq \mu_{Y,n}(\hat{\theta}_n; P) \mid C_n^H(\tilde{\theta}; P) = 1 \right\} - \alpha \right| E_P \left\{ C_n^H(\tilde{\theta}; P) \right\} = 0, \quad (\text{A.23})$$

for all $\tilde{\theta} \in \Theta$, and

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} \left| Pr_P \left\{ \hat{\mu}_{\alpha,n}^H \geq \mu_{Y,n}(\hat{\theta}_n; P) \right\} - \alpha \right| \leq \max\{\alpha, 1-\alpha\} \beta. \quad (\text{A.24})$$

Validity of hybrid estimators again implies validity of hybrid confidence intervals.

Corollary 2

Under Assumptions 6-8, for $CI_{ET,n}^H$ the level $1-\alpha$ equal-tailed hybrid confidence interval based on $CI_{P,n}^\beta$,

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} \left| Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_{ET,n}^H \mid C_n^H(\tilde{\theta}; P) = 1 \right\} - \frac{1-\alpha}{1-\beta} \right| E_P \left\{ C_n^H(\tilde{\theta}; P) \right\} = 0, \quad (\text{A.25})$$

for all $\tilde{\theta} \in \Theta$,

$$\liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_n} Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_{ET,n}^H \right\} \geq 1-\alpha, \quad (\text{A.26})$$

and

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_{ET,n}^H \right\} \leq \frac{1-\alpha}{1-\beta} \leq 1-\alpha+\beta. \quad (\text{A.27})$$

Hence, our procedures are uniformly asymptotically valid, unlike conventional inference.⁴

⁴The bootstrap also fails to deliver uniform validity, as it implicitly tries to estimate the difference be-

F.2 Auxiliary Lemmas

This section collects lemmas that we will use to prove our uniformity results.

Lemma 5

Under Assumption 8, for any sequence of confidence intervals CI_n , any sequence of sets $\mathcal{C}_n(P)$ indexed by P , $C_n(P) = 1 \left\{ \left(X_n, Y_n, \widehat{\Sigma}_n \right) \in \mathcal{C}_n(P) \right\}$, and any constant α , to show that

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} \left| Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_n \mid C_n(P) = 1 \right\} - \alpha \right| Pr_P \{ C_n(P) = 1 \} = 0$$

it suffices to show that for all subsequences $\{n_s\} \subseteq \{n\}$, $\{P_{n_s}\} \in \mathcal{P}^\infty = \times_{n=1}^\infty \mathcal{P}_n$ with:

1. $\Sigma(P_{n_s}) \rightarrow \Sigma^* \in \mathcal{S}$ for

$$\mathcal{S} = \left\{ \Sigma : 1/\bar{\lambda} \leq (\Sigma_X(\theta), \Sigma_Y(\theta)) \leq \bar{\lambda}, 1/\bar{\lambda} \leq \sqrt{\Sigma_X(\theta; P)\Sigma_X(\tilde{\theta}; P) - \Sigma_X(\theta, \tilde{\theta}; P)} \right\}, \quad (\text{A.28})$$

2. $Pr_{P_{n_s}} \{ C_{n_s}(P_{n_s}) = 1 \} \rightarrow p^* \in (0, 1]$, and

3. $\mu_{X, n_s}(P_{n_s}) - \max_{\theta} \mu_{X, n_s}(\theta; P_{n_s}) \rightarrow \mu_X^* \in \mathcal{M}_X^*$ for

$$\mathcal{M}_X^* = \left\{ \mu_X \in [-\infty, 0]^{|\Theta|} : \max_{\theta} \mu_X(\theta) = 0 \right\},$$

we have

$$\lim_{s \rightarrow \infty} Pr_{P_{n_s}} \left\{ \mu_{Y, n_s}(\hat{\theta}_{n_s}; P_{n_s}) \in CI_{n_s} \mid C_{n_s}(P_{n_s}) = 1 \right\} = \alpha. \quad (\text{A.29})$$

Lemma 6

For collections of sets $\mathcal{C}_{n,1}(P), \dots, \mathcal{C}_{n,J}(P)$, and $C_{n,j}(P) = 1 \left\{ \left(X_n, Y_n, \widehat{\Sigma}_n \right) \in \mathcal{C}_{n,j}(P) \right\}$, if $\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} Pr_P \{ C_{n,j}(P) = 1, C_{n,j'}(P) = 1 \} = 0$ for all $j \neq j'$ and

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} \left| Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_n \mid C_{n,j}(P) = 1 \right\} - (1-\alpha) \right| Pr_P \{ C_{n,j}(P) = 1 \} = 0$$

for all j , then

$$\liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_n} Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_n \right\} \geq (1-\alpha) \cdot \liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_n} \sum_j Pr_P \{ C_{n,j}(P) = 1 \},$$

tween the “winning” policy and the others, which cannot be done with sufficient precision. We are unaware of results for subsampling, m-out-of-n bootstrap, or other resampling-based approaches for this setting.

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_n \right\} \leq 1 - \alpha \cdot \liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_n} \sum_j Pr_P \{ C_{n,j}(P) = 1 \}.$$

To state the next lemma, define

$$\mathcal{L}(\tilde{\theta}, Z, \Sigma) = \max_{\theta \in \Theta: \Sigma_{XY}(\tilde{\theta}) > \Sigma_{XY}(\tilde{\theta}, \theta)} \frac{\Sigma_Y(\tilde{\theta})(Z(\theta) - Z(\tilde{\theta}))}{\Sigma_{XY}(\tilde{\theta}) - \Sigma_{XY}(\tilde{\theta}, \theta)} \quad (\text{A.30})$$

$$\mathcal{U}(\tilde{\theta}, Z, \Sigma) = \min_{\theta \in \Theta: \Sigma_{XY}(\tilde{\theta}) < \Sigma_{XY}(\tilde{\theta}, \theta)} \frac{\Sigma_Y(\tilde{\theta})(Z(\theta) - Z(\tilde{\theta}))}{\Sigma_{XY}(\tilde{\theta}) - \Sigma_{XY}(\tilde{\theta}, \theta)}, \quad (\text{A.31})$$

where we define a maximum over the empty set as $-\infty$ and a minimum over the empty set as $+\infty$. For

$$\begin{pmatrix} X_n^* \\ Y_n^* \end{pmatrix} = \begin{pmatrix} X_n - \max_{\theta} \mu_{X,n}(\theta; P) \\ Y_n - \mu_{Y,n}(P) \end{pmatrix},$$

we next show that using $(X_n^*, Y_n^*, \hat{\Sigma}_n)$ in our calculations yields the same bounds \mathcal{L} and \mathcal{U} as using $(X_n, Y_n, \hat{\Sigma}_n)$, up to additive shifts.

Lemma 7

For $\mathcal{L}(\tilde{\theta}, Z, \Sigma)$ and $\mathcal{U}(\tilde{\theta}, Z, \Sigma)$ as defined in (A.30) and (A.31), and

$$Z_{\tilde{\theta},n} = X_n - \frac{\hat{\Sigma}_{XY,n}(\cdot, \tilde{\theta})}{\hat{\Sigma}_{Y,n}(\tilde{\theta})} Y_n(\tilde{\theta}), \quad Z_{\tilde{\theta},n}^* = X_n^* - \frac{\hat{\Sigma}_{XY,n}(\cdot, \tilde{\theta})}{\hat{\Sigma}_{Y,n}(\tilde{\theta})} Y_n^*(\tilde{\theta}),$$

we have

$$\mathcal{L}(\tilde{\theta}, Z_{\tilde{\theta},n}^*, \hat{\Sigma}_n) = \mathcal{L}(\tilde{\theta}, Z_{\tilde{\theta},n}, \hat{\Sigma}_n) - \mu_{Y,n}(\tilde{\theta}; P), \quad \mathcal{U}(\tilde{\theta}, Z_{\tilde{\theta},n}^*, \hat{\Sigma}_n) = \mathcal{U}(\tilde{\theta}, Z_{\tilde{\theta},n}, \hat{\Sigma}_n) - \mu_{Y,n}(\tilde{\theta}; P).$$

For brevity, going forward we use the shorthand notation

$$\left(\mathcal{L}(\tilde{\theta}, Z_{\tilde{\theta},n}, \hat{\Sigma}_n), \mathcal{U}(\tilde{\theta}, Z_{\tilde{\theta},n}, \hat{\Sigma}_n), \mathcal{L}(\tilde{\theta}, Z_{\tilde{\theta},n}^*, \hat{\Sigma}_n), \mathcal{U}(\tilde{\theta}, Z_{\tilde{\theta},n}^*, \hat{\Sigma}_n) \right) = (\mathcal{L}_n, \mathcal{U}_n, \mathcal{L}_n^*, \mathcal{U}_n^*).$$

Lemma 8

Under Assumptions 6 and 7, for any $\{n_s\}$ and $\{P_{n_s}\}$ satisfying conditions (1)-(3) of Lemma 5 and any $\tilde{\theta}$ with $\mu_X^*(\tilde{\theta}) > -\infty$, $(Y_{n_s}^*, \mathcal{L}_{n_s}^*, \mathcal{U}_{n_s}^*, \hat{\Sigma}_{n_s}, \hat{\theta}_{n_s}) \rightarrow_d (Y^*, \mathcal{L}^*, \mathcal{U}^*, \Sigma^*, \hat{\theta})$, where the

objects on the right hand side are calculated based on (Y^*, X^*, Σ^*) for $(X^*, Y^*)' \sim N(\mu^*, \Sigma^*)$ with $\mu^* = (\mu_X^*, 0)'$.

Lemma 9

For F_N again the standard normal distribution function, the function

$$F_{TN}(Y(\theta); \mu, \Sigma_Y(\theta), \mathcal{L}, \mathcal{U}) = \frac{F_N\left(\frac{Y(\theta) \wedge \mathcal{U} - \mu}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{\mathcal{L} - \mu}{\sqrt{\Sigma_Y(\theta)}}\right)}{F_N\left(\frac{\mathcal{U} - \mu}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{\mathcal{L} - \mu}{\sqrt{\Sigma_Y(\theta)}}\right)} 1\{Y(\theta) \geq \mathcal{L}\} \quad (\text{A.32})$$

is continuous in $(Y(\theta), \mu, \Sigma_Y(\theta), \mathcal{L}, \mathcal{U})$ on the set

$$\{(Y(\theta), \mu, \Sigma_Y(\theta)) \in \mathbb{R}^3, \mathcal{L} \in \mathbb{R} \cup \{-\infty\}, \mathcal{U} \in \mathbb{R} \cup \{\infty\} : \Sigma_Y(\theta) > 0, \mathcal{L} < Y(\theta) < \mathcal{U}\}.$$

F.3 Proofs for Auxiliary Lemmas

Proof of Lemma 5 To prove that

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} \left| Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_n | C_n(P) = 1 \right\} - \alpha \right| Pr_P \{C_n(P) = 1\} = 0$$

it suffices to show that

$$\liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_n} \left(Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_n | C_n(P) = 1 \right\} - \alpha \right) Pr_P \{C_n(P) = 1\} \geq 0 \quad (\text{A.33})$$

and

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} \left(Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_n | C_n(P) = 1 \right\} - \alpha \right) Pr_P \{C_n(P) = 1\} \leq 0. \quad (\text{A.34})$$

We prove that to show (A.33), it suffices to show that for all $\{n_s\}$, $\{P_{n_s}\}$ satisfying conditions (1)-(3) of the lemma,

$$\liminf_{s \rightarrow \infty} Pr_{P_{n_s}} \left\{ \mu_{Y,n_s}(\hat{\theta}_{n_s}; P_{n_s}) \in CI_{n_s} | C_{n_s}(P_{n_s}) = 1 \right\} \geq \alpha. \quad (\text{A.35})$$

An argument along the same lines implies that to prove (A.34) it suffices to show that

$$\limsup_{s \rightarrow \infty} Pr_{P_{n_s}} \left\{ \mu_{Y,n_s}(\hat{\theta}_{n_s}; P_{n_s}) \in CI_{n_s} | C_{n_s}(P_{n_s}) = 1 \right\} \leq \alpha. \quad (\text{A.36})$$

Note, however, that (A.35) and (A.36) together are equivalent to (A.29).

Towards contradiction, suppose that (A.33) fails, so

$$\liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_n} \left(Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_n | C_n(P) = 1 \right\} - \alpha \right) Pr_P \{ C_n(P) = 1 \} < -\varepsilon,$$

for some $\varepsilon > 0$ but that (A.35) holds for all sequences satisfying conditions (1)-(3) of the lemma. Then there exists an increasing sequence of sample sizes n_q and some sequence $\{P_{n_q}\}$ with $P_{n_q} \in \mathcal{P}_{n_q}$ for all q such that

$$\limsup_{q \rightarrow \infty} \left(Pr_{P_{n_q}} \left\{ \mu_{Y,n_q}(\hat{\theta}_{n_q}; P_{n_q}) \in CI_{n_q} | C_{n_q}(P_{n_q}) = 1 \right\} - \alpha \right) Pr_{P_{n_q}} \{ C_{n_q}(P_{n_q}) = 1 \} < -\varepsilon. \quad (\text{A.37})$$

We want to show that there exists a further subsequence $\{n_s\} \subseteq \{n_q\}$ satisfying (1)-(3) in the statement of the lemma, and so establish a contradiction.

Note that since the set \mathcal{S} defined in (A.28) is compact (e.g. in the Frobenius norm), and Assumption 8 implies that $\Sigma(P_{n_q}) \in \mathcal{S}$ for all q , there exists a further subsequence $\{n_r\} \subseteq \{n_q\}$ such that

$$\lim_{r \rightarrow \infty} \Sigma(P_{n_r}) \rightarrow \Sigma^*$$

for some $\Sigma^* \in \mathcal{S}$.

Note, next, that $Pr_{P_{n_r}} \{ C_{n_r}(P_{n_r}) = 1 \} \in [0, 1]$ for all r , and so converges along a subsequence $\{n_t\} \subseteq \{n_r\}$. However, (A.37) implies that $Pr_{P_{n_r}} \{ C_{n_r}(P_{n_r}) = 1 \} \geq \frac{\varepsilon}{\alpha}$ for all r , and thus that $Pr_{P_{n_t}} \{ C_{n_t}(P_{n_t}) = 1 \} \rightarrow p^* \in [\frac{\varepsilon}{\alpha}, 1]$.

Finally, let us define $\mu_{X,n}^*(P) = \mu_{X,n}(P) - \max_{\theta} \mu_{X,n}(\theta; P)$, and note that $\mu_{X,n}^*(P) \leq 0$ by construction. Since $\mu_{X,n}^*(P)$ is finite-dimensional and $\max_{\theta} \mu_{X,n}^*(P; \theta) = 0$, there exists some $\theta \in \Theta$ such that $\mu_{X,n}^*(P; \theta)$ is equal to zero infinitely often. Let $\{n_u\} \subseteq \{n_t\}$ extract the corresponding sequence of sample sizes. The set $[-\infty, 0]^{| \Theta |}$ is compact under the metric $d(\mu_X, \tilde{\mu}_X) = \|F_N(\mu_X) - F_N(\tilde{\mu}_X)\|$ for $F_N(\cdot)$ the standard normal cdf applied elementwise, and $\|\cdot\|$ the Euclidean norm. Hence, there exists a further subsequence $\{n_s\} \subseteq \{n_u\}$ along which $\mu_{X,n_s}^*(P_{n_s})$ converges to a limit in this metric. Note, however, that this means that $\mu_{X,n_s}^*(P_{n_s})$ converges to a limit $\mu^* \in \mathcal{M}^*$ in the usual metric.

Hence, we have shown that there exists a subsequence $\{n_s\} \subseteq \{n_q\}$ that satisfies (1)-(3). By supposition, (A.35) must hold along this subsequence. Thus,

$$\liminf_{n \rightarrow \infty} \left(Pr_{P_{n_s}} \left\{ \mu_{Y,n_s}(\hat{\theta}_{n_s}; P_{n_s}) \in CI_{n_s} | C_{n_s}(P_{n_s}) = 1 \right\} - \alpha \right) Pr_{P_{n_s}} \{ C_{n_s}(P_{n_s}) = 1 \} \geq 0,$$

which contradicts (A.37). Hence, we have established a contradiction and so proved that (A.35) for all subsequences satisfying conditions (1)-(3) of the lemma implies (A.33).

An argument along the same lines shows that (A.36) along all subsequences satisfying conditions (1)-(3) of the lemma implies (A.34). \square

Proof of Lemma 6 Define $C_{n,J+1}(P) = 1\{C_{n,j}(P) = 0 \text{ for all } j \in \{1, \dots, J\}\}$. Note that

$$\begin{aligned} & Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_n \right\} \\ &= \sum_{j=1}^{J+1} Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_n | C_{n,j}(P) = 1 \right\} Pr_P \{C_{n,j}(P) = 1\} + o(1) \end{aligned}$$

where the $o(1)$ term is negligible uniformly over $P \in \mathcal{P}_n$ as $n \rightarrow \infty$. Hence,

$$\begin{aligned} & Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_n \right\} - (1-\alpha) \\ &= \sum_{j=1}^{J+1} \left(Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_n | C_{n,j}(P) = 1 \right\} - (1-\alpha) \right) Pr_P \{C_{n,j}(P) = 1\} + o(1) \end{aligned}$$

and

$$\begin{aligned} & \liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_n} Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_n \right\} - (1-\alpha) \\ &= \liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_n} \sum_{j=1}^{J+1} \left(Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_n | C_{n,j}(P) = 1 \right\} - (1-\alpha) \right) Pr_P \{C_{n,j}(P) = 1\} \\ &= \liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_n} \left(Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_n | C_{n,J+1}(P) = 1 \right\} - (1-\alpha) \right) Pr_P \{C_{n,J+1}(P) = 1\} \\ & \quad \geq -(1-\alpha) \limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} Pr_P \{C_{n,J+1}(P) = 1\} \\ &= -(1-\alpha) \left(1 - \liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_n} \sum_{j=1}^J Pr_P \{C_{n,j}(P) = 1\} \right) \end{aligned}$$

which immediately implies that

$$\liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_n} Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_n \right\} \geq (1-\alpha) \liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_n} \sum_{j=1}^J Pr_P \{C_{n,j}(P) = 1\}.$$

Likewise,

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_n \right\} - (1-\alpha) \\ &= \limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} \sum_{j=1}^{J+1} \left(Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_n | C_{n,j}(P) = 1 \right\} - (1-\alpha) \right) Pr_P \{C_{n,j}(P) = 1\} \end{aligned}$$

$$\begin{aligned}
&= \limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} \left(Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_n | C_{n,J+1}(P) = 1 \right\} - (1-\alpha) \right) Pr_P \{ C_{n,J+1}(P) = 1 \} \\
&\leq \alpha \cdot \limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} Pr_P \{ C_{n,J+1}(P) = 1 \} = \alpha \left(1 - \liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_n} \sum_{j=1}^J Pr_P \{ C_{n,j}(P) = 1 \} \right).
\end{aligned}$$

This immediately implies that

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_n \right\} \leq 1 - \alpha \cdot \liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_n} \sum_{j=1}^J Pr_P \{ C_{n,j}(P) = 1 \},$$

as we wanted to show. \square

Proof of Lemma 7 Note that

$$Z_{\tilde{\theta},n}^* = Z_{\tilde{\theta},n} - \max_{\theta} \mu_{X,n}(\theta; P) + \widehat{\Sigma}_{XY,n}(\cdot, \tilde{\theta}) \frac{\mu_{Y,n}(\tilde{\theta}; P)}{\widehat{\Sigma}_{Y,n}(\tilde{\theta})},$$

so

$$Z_{\tilde{\theta},n}^*(\theta) - Z_{\tilde{\theta},n}^*(\tilde{\theta}) = Z_{\tilde{\theta},n}(\theta) - Z_{\tilde{\theta},n}(\tilde{\theta}) + \left(\widehat{\Sigma}_{XY,n}(\theta, \tilde{\theta}) - \widehat{\Sigma}_{XY,n}(\tilde{\theta}, \tilde{\theta}) \right) \frac{\mu_{Y,n}(\tilde{\theta}; P)}{\widehat{\Sigma}_{Y,n}(\tilde{\theta})}.$$

The result follows immediately. \square

Proof of Lemma 8 By Assumption 6

$$\begin{pmatrix} X_{n_s} - \mu_{X,n_s}(P_{n_s}) \\ Y_{n_s} - \mu_{Y,n_s}(P_{n_s}) \end{pmatrix} \rightarrow_d N(0, \Sigma^*).$$

Hence, by Slutsky's lemma

$$\begin{pmatrix} X_{n_s}^* \\ Y_{n_s}^* \end{pmatrix} = \begin{pmatrix} X_{n_s} - \max_{\theta} \mu_{X,n_s}(\theta; P_{n_s}) \\ Y_{n_s} - \mu_{Y,n_s}(P_{n_s}) \end{pmatrix} \rightarrow_d \begin{pmatrix} X^* \\ Y^* \end{pmatrix} \sim N(\mu^*, \Sigma^*).$$

We begin by considering one $\theta \in \Theta \setminus \{\tilde{\theta}\}$ at a time. Since $\widehat{\Sigma}_{n_s} \rightarrow_p \Sigma^*$ by Assumption

7, if $\Sigma_{XY}^*(\tilde{\theta}) - \Sigma_{XY}^*(\tilde{\theta}, \theta) \neq 0$ then

$$\frac{\widehat{\Sigma}_{Y, n_s}(\tilde{\theta}) \left(Z_{\tilde{\theta}, n_s}^*(\theta) - Z_{\tilde{\theta}, n_s}^*(\tilde{\theta}) \right)}{\widehat{\Sigma}_{XY, n_s}(\tilde{\theta}) - \widehat{\Sigma}_{XY, n_s}(\tilde{\theta}, \theta)} \rightarrow_d \frac{\Sigma_Y^*(\tilde{\theta}) \left(Z_{\tilde{\theta}}^*(\theta) - Z_{\tilde{\theta}}^*(\tilde{\theta}) \right)}{\Sigma_{XY}^*(\tilde{\theta}) - \Sigma_{XY}^*(\tilde{\theta}, \theta)},$$

where the terms on the right hand side are based on (X^*, Y^*, Σ^*) . The limit is finite if $\mu_X^*(\theta) > -\infty$, while otherwise $\mu_X^*(\theta) = -\infty$ and

$$\frac{\Sigma_Y^*(\tilde{\theta}) \left(Z_{\tilde{\theta}}^*(\theta) - Z_{\tilde{\theta}}^*(\tilde{\theta}) \right)}{\Sigma_{XY}^*(\tilde{\theta}) - \Sigma_{XY}^*(\tilde{\theta}, \theta)} = \begin{cases} -\infty & \text{if } \Sigma_{XY}^*(\tilde{\theta}) - \Sigma_{XY}^*(\tilde{\theta}, \theta) > 0 \\ +\infty & \text{if } \Sigma_{XY}^*(\tilde{\theta}) - \Sigma_{XY}^*(\tilde{\theta}, \theta) < 0 \end{cases}.$$

If instead $\Sigma_{XY}^*(\tilde{\theta}) - \Sigma_{XY}^*(\tilde{\theta}, \theta) = 0$, then since $\Sigma_X^*(\tilde{\theta}, \theta) < \sqrt{\Sigma_X^*(\tilde{\theta}) \Sigma_X^*(\theta)}$,

$$Z_{\tilde{\theta}}^*(\theta) - Z_{\tilde{\theta}}^*(\tilde{\theta}) = X^*(\theta) - X^*(\tilde{\theta})$$

is normally distributed with non-zero variance. Hence, in this case

$$\left| \frac{\widehat{\Sigma}_{Y, n_s}(\tilde{\theta}) \left(Z_{n_s, \tilde{\theta}}^*(\theta) - Z_{n_s, \tilde{\theta}}^*(\tilde{\theta}) \right)}{\widehat{\Sigma}_{XY, n_s}(\tilde{\theta}) - \widehat{\Sigma}_{XY, n_s}(\tilde{\theta}, \theta)} \right| \rightarrow \infty. \quad (\text{A.38})$$

Let us define

$$\Theta^*(\tilde{\theta}) = \left\{ \theta \in \Theta \setminus \tilde{\theta} : \Sigma_{XY}^*(\tilde{\theta}) - \Sigma_{XY}^*(\tilde{\theta}, \theta) \neq 0 \right\}.$$

The argument above implies that

$$\begin{aligned} & \max_{\theta \in \Theta^*(\tilde{\theta}) : \widehat{\Sigma}_{XY, n_s}(\tilde{\theta}) > \widehat{\Sigma}_{XY, n_s}(\tilde{\theta}, \theta)} \frac{\widehat{\Sigma}_{Y, n_s}(\tilde{\theta}) \left(Z_{\tilde{\theta}, n_s}^*(\theta) - Z_{\tilde{\theta}, n_s}^*(\tilde{\theta}) \right)}{\widehat{\Sigma}_{XY, n_s}(\tilde{\theta}) - \widehat{\Sigma}_{XY, n_s}(\tilde{\theta}, \theta)} \\ & \rightarrow_d \mathcal{L}^* = \max_{\theta \in \Theta : \Sigma_{XY}^*(\tilde{\theta}) > \Sigma_{XY}^*(\tilde{\theta}, \theta)} \frac{\Sigma_Y^*(\tilde{\theta}) \left(Z_{\tilde{\theta}}^*(\theta) - Z_{\tilde{\theta}}^*(\tilde{\theta}) \right)}{\Sigma_{XY}^*(\tilde{\theta}) - \Sigma_{XY}^*(\tilde{\theta}, \theta)}, \end{aligned} \quad (\text{A.39})$$

and

$$\min_{\theta \in \Theta^*(\tilde{\theta}) : \widehat{\Sigma}_{XY, n_s}(\tilde{\theta}) < \widehat{\Sigma}_{XY, n_s}(\tilde{\theta}, \theta)} \frac{\widehat{\Sigma}_{Y, n_s}(\tilde{\theta}) \left(Z_{\tilde{\theta}, n_s}^*(\theta) - Z_{\tilde{\theta}, n_s}^*(\tilde{\theta}) \right)}{\widehat{\Sigma}_{XY, n_s}(\tilde{\theta}) - \widehat{\Sigma}_{XY, n_s}(\tilde{\theta}, \theta)}$$

$$\rightarrow_d \mathcal{U}^* = \min_{\theta \in \Theta: \Sigma_{XY}^*(\tilde{\theta}) < \Sigma_{XY}^*(\tilde{\theta}, \theta)} \frac{\Sigma_Y^*(\tilde{\theta}) \left(Z_{\tilde{\theta}}^*(\theta) - Z_{\tilde{\theta}}^*(\tilde{\theta}) \right)}{\Sigma_{XY}^*(\tilde{\theta}) - \Sigma_{XY}^*(\tilde{\theta}, \theta)}. \quad (\text{A.40})$$

Since

$$\begin{aligned} & \max_{\theta \in \Theta: \widehat{\Sigma}_{XY, n_s}(\tilde{\theta}) > \widehat{\Sigma}_{XY, n_s}(\tilde{\theta}, \theta)} \frac{\widehat{\Sigma}_{Y, n_s}(\tilde{\theta}) \left(Z_{\tilde{\theta}, n_s}^*(\theta) - Z_{\tilde{\theta}, n_s}^*(\tilde{\theta}) \right)}{\widehat{\Sigma}_{XY, n_s}(\tilde{\theta}) - \widehat{\Sigma}_{XY, n_s}(\tilde{\theta}, \theta)} \leq Y_{n_s}^*(\tilde{\theta}) \\ & \leq \min_{\theta \in \Theta: \widehat{\Sigma}_{XY, n_s}(\tilde{\theta}) < \widehat{\Sigma}_{XY, n_s}(\tilde{\theta}, \theta)} \frac{\widehat{\Sigma}_{Y, n_s}(\tilde{\theta}) \left(Z_{\tilde{\theta}, n_s}^*(\theta) - Z_{\tilde{\theta}, n_s}^*(\tilde{\theta}) \right)}{\widehat{\Sigma}_{XY, n_s}(\tilde{\theta}) - \widehat{\Sigma}_{XY, n_s}(\tilde{\theta}, \theta)} \end{aligned}$$

with probability one for all n_s and $Y_{n_s} \xrightarrow{d} Y^*$, (A.38) implies

$$\frac{\widehat{\Sigma}_{Y, n_s}(\tilde{\theta}) \left(Z_{n_s, \tilde{\theta}}^*(\theta) - Z_{n_s, \tilde{\theta}}^*(\tilde{\theta}) \right)}{\widehat{\Sigma}_{XY, n_s}(\tilde{\theta}) - \widehat{\Sigma}_{XY, n_s}(\tilde{\theta}, \theta)} \rightarrow -\infty$$

when $\Sigma_{XY}^*(\tilde{\theta}) = \Sigma_{XY}^*(\tilde{\theta}, \theta)$ for all $\theta, \tilde{\theta} \in \Theta$ such that $\widehat{\Sigma}_{XY, n_s}(\tilde{\theta}) > \widehat{\Sigma}_{XY, n_s}(\tilde{\theta}, \theta)$. Similarly,

$$\frac{\widehat{\Sigma}_{Y, n_s}(\tilde{\theta}) \left(Z_{n_s, \tilde{\theta}}^*(\theta) - Z_{n_s, \tilde{\theta}}^*(\tilde{\theta}) \right)}{\widehat{\Sigma}_{XY, n_s}(\tilde{\theta}) - \widehat{\Sigma}_{XY, n_s}(\tilde{\theta}, \theta)} \rightarrow \infty$$

when $\Sigma_{XY}^*(\tilde{\theta}) = \Sigma_{XY}^*(\tilde{\theta}, \theta)$ for all $\theta, \tilde{\theta} \in \Theta$ such that $\widehat{\Sigma}_{XY, n_s}(\tilde{\theta}) < \widehat{\Sigma}_{XY, n_s}(\tilde{\theta}, \theta)$. Thus, the same convergence results as (A.39)–(A.40) continue to hold when we minimize and maximize over Θ rather than $\Theta^*(\tilde{\theta})$. Hence, $(\mathcal{L}_{n_s}^*, \mathcal{U}_{n_s}^*) \rightarrow_d (\mathcal{L}^*, \mathcal{U}^*)$. Moreover, $\hat{\theta}_{n_s}$ is almost everywhere continuous in $X_{n_s}^*$, so $(Y_{n_s}^*, \widehat{\Sigma}_{n_s}, \hat{\theta}_{n_s}) \rightarrow_d (Y^*, \Sigma^*, \hat{\theta})$ by the continuous mapping theorem, and this convergence holds jointly with that for $(\mathcal{L}_{n_s}^*, \mathcal{U}_{n_s}^*)$. Hence, we have established the desired convergence. \square

Proof of Lemma 9 Continuity for $\Sigma_Y(\theta) > 0, \mathcal{L} < Y(\theta) < \mathcal{U}$ with all elements finite is immediate from the functional form. Moreover, for fixed $(Y(\theta), \mu, \Sigma_Y(\theta)) \in \mathbb{R}^3$ with $\Sigma_Y(\theta) > 0$ and $\mathcal{L} < Y(\theta) < \mathcal{U}$,

$$\lim_{\mathcal{U} \rightarrow \infty} \frac{F_N \left(\frac{Y(\theta) \wedge \mathcal{U} - \mu}{\sqrt{\Sigma_Y(\theta)}} \right) - F_N \left(\frac{\mathcal{L} - \mu}{\sqrt{\Sigma_Y(\theta)}} \right)}{F_N \left(\frac{\mathcal{U} - \mu}{\sqrt{\Sigma_Y(\theta)}} \right) - F_N \left(\frac{\mathcal{L} - \mu}{\sqrt{\Sigma_Y(\theta)}} \right)} \mathbf{1}\{Y(\theta) \geq \mathcal{L}\} = \frac{F_N \left(\frac{Y(\theta) - \mu}{\sqrt{\Sigma_Y(\theta)}} \right) - F_N \left(\frac{\mathcal{L} - \mu}{\sqrt{\Sigma_Y(\theta)}} \right)}{F_N \left(\frac{\infty}{\sqrt{\Sigma_Y(\theta)}} \right) - F_N \left(\frac{\mathcal{L} - \mu}{\sqrt{\Sigma_Y(\theta)}} \right)}$$

$$\lim_{\mathcal{L} \rightarrow -\infty} \frac{F_N\left(\frac{Y(\theta) \wedge \mathcal{U} - \mu}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{\mathcal{L} - \mu}{\sqrt{\Sigma_Y(\theta)}}\right)}{F_N\left(\frac{\mathcal{U} - \mu}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{\mathcal{L} - \mu}{\sqrt{\Sigma_Y(\theta)}}\right)} \mathbf{1}\{Y(\theta) \geq \mathcal{L}\} = \frac{F_N\left(\frac{Y(\theta) - \mu}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{-\infty}{\sqrt{\Sigma_Y(\theta)}}\right)}{F_N\left(\frac{\mathcal{U} - \mu}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{-\infty}{\sqrt{\Sigma_Y(\theta)}}\right)}$$

and

$$\lim_{(\mathcal{L}, \mathcal{U}) \rightarrow (-\infty, \infty)} \frac{F_N\left(\frac{Y(\theta) \wedge \mathcal{U} - \mu}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{\mathcal{L} - \mu}{\sqrt{\Sigma_Y(\theta)}}\right)}{F_N\left(\frac{\mathcal{U} - \mu}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{\mathcal{L} - \mu}{\sqrt{\Sigma_Y(\theta)}}\right)} \mathbf{1}\{Y(\theta) \geq \mathcal{L}\} = \frac{F_N\left(\frac{Y(\theta) - \mu}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{-\infty}{\sqrt{\Sigma_Y(\theta)}}\right)}{F_N\left(\frac{\infty}{\sqrt{\Sigma_Y(\theta)}}\right) - F_N\left(\frac{-\infty}{\sqrt{\Sigma_Y(\theta)}}\right)}.$$

Hence, we obtain the desired result. \square

F.4 Proofs for Uniformity Results

Proof of Proposition 10 Note that

$$\hat{\mu}_{\alpha, n} \geq \mu_{Y, n}(\hat{\theta}_n; P) \iff \mu_{Y, n}(\hat{\theta}_n; P) \in CI_{U, -, n}$$

for $CI_{U, -, n} = (-\infty, \hat{\mu}_{\alpha, n}]$. Hence, by Lemma 5, to prove that (A.20) holds it suffices to show that for all $\{n_s\}$ and $\{P_{n_s}\}$ such that conditions (1)-(3) of the lemma hold with $C_n(P) = \mathbf{1}\{\hat{\theta}_n = \tilde{\theta}\}$, we have

$$\lim_{s \rightarrow \infty} Pr_{P_{n_s}} \left\{ \hat{\mu}_{Y, n_s}(\hat{\theta}_{n_s}; P_{n_s}) \in CI_{U, -, n_s} | \hat{\theta}_{n_s} = \tilde{\theta} \right\} = \alpha. \quad (\text{A.41})$$

To this end, recall that for $F_{TN}(Y(\theta); \mu, \Sigma_Y(\theta), \mathcal{L}, \mathcal{U})$ as defined in (A.32), the estimator $\hat{\mu}_{\alpha, n}$ solves $F_{TN}(Y_n(\hat{\theta}_n); \mu, \hat{\Sigma}_{Y, n}(\hat{\theta}_n), \mathcal{L}(\hat{\theta}_n, Z_{\hat{\theta}_n, n}, \hat{\Sigma}_n), \mathcal{U}(\hat{\theta}_n, Z_{\hat{\theta}_n, n}, \hat{\Sigma}_n)) = 1 - \alpha$. This cdf is strictly decreasing in μ as argued in the proof of Proposition 5, and is increasing in $Y_n(\hat{\theta})$. Hence, $\hat{\mu}_{\alpha, n} \geq \mu_{Y, n}(\hat{\theta}_n; P)$ if and only if

$$F_{TN}(Y_n(\hat{\theta}_n); \mu_{Y, n}(\hat{\theta}_n; P), \hat{\Sigma}_{Y, n}(\hat{\theta}_n), \mathcal{L}(\hat{\theta}_n, Z_{\hat{\theta}_n, n}, \hat{\Sigma}_n), \mathcal{U}(\hat{\theta}_n, Z_{\hat{\theta}_n, n}, \hat{\Sigma}_n)) \geq 1 - \alpha.$$

Note, next, that by Lemma 7 and the form of the function F_{TN} ,

$$\begin{aligned} & F_{TN}(Y_n(\hat{\theta}_n); \mu_{Y, n}(\hat{\theta}_n; P), \hat{\Sigma}_{Y, n}(\hat{\theta}_n), \mathcal{L}(\hat{\theta}_n, Z_{\hat{\theta}_n, n}, \hat{\Sigma}_n), \mathcal{U}(\hat{\theta}_n, Z_{\hat{\theta}_n, n}, \hat{\Sigma}_n)) \\ &= F_{TN}(Y_n^*(\hat{\theta}_n); 0, \hat{\Sigma}_{Y, n}(\hat{\theta}_n), \mathcal{L}(\hat{\theta}_n, Z_{\hat{\theta}_n, n}^*, \hat{\Sigma}_n), \mathcal{U}(\hat{\theta}_n, Z_{\hat{\theta}_n, n}^*, \hat{\Sigma}_n)), \end{aligned}$$

so $\hat{\mu}_{\alpha,n} \geq \mu_{Y,n}(\hat{\theta}_n; P)$ if and only if

$$F_{TN}\left(Y_n^*(\hat{\theta}_n); 0, \widehat{\Sigma}_{Y,n}(\hat{\theta}_n), \mathcal{L}(\hat{\theta}_n, Z_{\hat{\theta}_n,n}^*, \widehat{\Sigma}_n), \mathcal{U}(\hat{\theta}_n, Z_{\hat{\theta}_n,n}^*, \widehat{\Sigma}_n)\right) \geq 1 - \alpha.$$

Lemma 8 shows that $(Y_n^*(\hat{\theta}_{n_s}), \widehat{\Sigma}_{Y,n_s}(\hat{\theta}_{n_s}), \mathcal{L}_{n_s}^*, \mathcal{U}_{n_s}^*, \hat{\theta}_{n_s})$ converges in distribution as $s \rightarrow \infty$, so since F_{TN} is continuous by Lemma 9 while $\operatorname{argmax}_{\theta} X^*(\theta)$ is almost surely unique and continuous for X^* as in Lemma 8, the continuous mapping theorem implies that

$$\begin{aligned} & \left(F_{TN}\left(Y_{n_s}^*(\hat{\theta}_{n_s}); 0, \widehat{\Sigma}_{Y,n_s}(\hat{\theta}_{n_s}), \mathcal{L}_{n_s}^*, \mathcal{U}_{n_s}^*\right), 1\{\hat{\theta}_{n_s} = \tilde{\theta}\}\right) \\ & \rightarrow_d \left(F_{TN}\left(Y^*(\hat{\theta}); 0, \Sigma_Y^*(\hat{\theta}), \mathcal{L}^*, \mathcal{U}^*\right), 1\{\hat{\theta} = \tilde{\theta}\}\right). \end{aligned}$$

Since we can write

$$\begin{aligned} & Pr_{P_{n_s}} \left\{ F_{TN}\left(Y_{n_s}^*(\hat{\theta}_{n_s}); 0, \widehat{\Sigma}_{Y,n_s}(\hat{\theta}_{n_s}), \mathcal{L}_{n_s}^*, \mathcal{U}_{n_s}^*\right) \geq 1 - \alpha \mid \hat{\theta}_{n_s} = \tilde{\theta} \right\} \\ & = \frac{E_{P_{n_s}} \left[1\left\{ F_{TN}\left(Y_{n_s}^*(\hat{\theta}_{n_s}); 0, \widehat{\Sigma}_{Y,n_s}(\hat{\theta}_{n_s}), \mathcal{L}_{n_s}^*, \mathcal{U}_{n_s}^*\right) \geq 1 - \alpha \right\} 1\{\hat{\theta}_{n_s} = \tilde{\theta}\} \right]}{E_{P_{n_s}} \left[1\{\hat{\theta}_{n_s} = \tilde{\theta}\} \right]}, \end{aligned}$$

and by construction (see also Proposition 7 in the main text),

$$F_{TN}\left(Y^*(\hat{\theta}); 0, \Sigma_Y^*(\hat{\theta}), \mathcal{L}^*, \mathcal{U}^*, \hat{\theta}\right) \mid \hat{\theta} = \tilde{\theta} \sim U[0,1],$$

and $Pr\{\hat{\theta} = \tilde{\theta}\} = p^* > 0$, we thus have that

$$\begin{aligned} & Pr_{P_{n_s}} \left\{ F_{TN}\left(Y_{n_s}^*(\hat{\theta}_{n_s}); 0, \widehat{\Sigma}_{Y,n_s}(\hat{\theta}_{n_s}), \mathcal{L}_{n_s}^*, \mathcal{U}_{n_s}^*\right) \geq 1 - \alpha \mid \hat{\theta}_{n_s} = \tilde{\theta} \right\} \\ & \rightarrow Pr \left\{ F_{TN}\left(Y^*(\hat{\theta}); 0, \Sigma_Y^*(\hat{\theta}), \mathcal{L}^*, \mathcal{U}^*\right) \geq 1 - \alpha \mid \hat{\theta} = \tilde{\theta} \right\} = \alpha, \end{aligned}$$

which verifies (A.41).

Since this argument holds for all $\tilde{\theta} \in \Theta$, and Assumptions 6 and 8 imply that for all $\theta, \tilde{\theta} \in \Theta$ with $\theta \neq \tilde{\theta}$, $\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} Pr_P \left\{ X_n(\theta) = X_n(\tilde{\theta}) \right\} = 0$, Lemma 6 implies (A.21). \square

Proof of Corollary 1 By construction, $CI_{ET,n} = [\hat{\mu}_{\alpha/2,n}, \hat{\mu}_{1-\alpha/2,n}]$, and $\hat{\mu}_{1-\alpha/2,n} > \hat{\mu}_{\alpha/2,n}$ for all $\alpha < 1$. Hence,

$$Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_{ET,n} \mid \hat{\theta}_n = \tilde{\theta} \right\}$$

$$= Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \leq \hat{\mu}_{1-\alpha/2,n} | \hat{\theta}_n = \tilde{\theta} \right\} - Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \leq \hat{\mu}_{\alpha/2,n} | \hat{\theta}_n = \tilde{\theta} \right\},$$

so the result is immediate from Proposition 10 and Lemma 6. \square

Proof of Proposition 11 By the same argument as in the proof of Lemma 5, to show that (A.22) holds it suffices to show that for all $\{n_s\}$, $\{P_{n_s}\}$ satisfying conditions (1)-(3) of Lemma 5, $\liminf_{n \rightarrow \infty} Pr_{P_{n_s}} \left\{ \mu_{Y,n_s}(\hat{\theta}_{n_s}; P_{n_s}) \in CI_{P,n_s} \right\} \geq 1 - \alpha$.

To this end, note that

$$\mu_{Y,n_s}(\hat{\theta}_{n_s}; P_{n_s}) \in CI_{P,n_s} \text{ if and only if } Y_{n_s}^*(\hat{\theta}_{n_s}) \in \left[-c_\alpha(\hat{\Sigma}_{Y,n_s}) \sqrt{\hat{\Sigma}_Y(\hat{\theta}_{n_s})}, c_\alpha(\hat{\Sigma}_{Y,n_s}) \sqrt{\hat{\Sigma}_Y(\hat{\theta}_{n_s})} \right]$$

for $c_\alpha(\Sigma_Y)$ the $1 - \alpha$ quantile of $\max_\theta |\xi(\theta)| / \sqrt{\Sigma_Y(\theta)}$ where $\xi \sim N(0, \Sigma_Y)$. Next, note that $c_\alpha(\Sigma_Y)$ is continuous in Σ on \mathcal{S} as defined in (A.28). Hence, for all θ , $c_\alpha(\Sigma_Y) \sqrt{\Sigma_Y(\theta)}$ is continuous as well. Assumptions 6 and 7 imply that $(Y_{n_s}^*, \hat{\Sigma}_{n_s}, \hat{\theta}_{n_s}) \rightarrow_d (Y^*, \Sigma^*, \hat{\theta})$, which by the continuous mapping theorem implies

$$\left(Y_{n_s}^*(\hat{\theta}_{n_s}), c_\alpha(\hat{\Sigma}_{Y,n_s}) \sqrt{\hat{\Sigma}_Y(\hat{\theta}_{n_s})} \right) \rightarrow_d \left(Y^*(\hat{\theta}), c_\alpha(\Sigma_Y^*) \sqrt{\Sigma_Y^*(\hat{\theta})} \right).$$

Hence, since $Pr \left\{ |Y^*(\hat{\theta})| - c_\alpha(\Sigma_Y^*) \sqrt{\Sigma_Y^*(\hat{\theta})} = 0 \right\} = 0$,

$$Pr_{P_{n_s}} \left\{ \mu_{Y,n_s}(\hat{\theta}_{n_s}; P_{n_s}) \in CI_{P,n_s} \right\} \rightarrow Pr \left\{ Y^*(\hat{\theta}) \in \left[-c_\alpha(\Sigma_Y^*) \sqrt{\Sigma_Y^*(\hat{\theta})}, c_\alpha(\Sigma_Y^*) \sqrt{\Sigma_Y^*(\hat{\theta})} \right] \right\} \quad (\text{A.42})$$

where the right hand side is at least $1 - \alpha$ by construction. \square

Proof of Proposition 12 Note that $\hat{\mu}_{\alpha,n}^H \geq \mu_{Y,n}(\hat{\theta}_n; P)$ if and only if $\mu_{Y,n}(\hat{\theta}_n; P) \in CI_{U,-,n}^H$ for $CI_{U,-,n}^H = (-\infty, \hat{\mu}_{\alpha,n}^H]$. Hence, by Lemma 5, to prove that (A.23) holds it suffices to show that for all $\{n_s\}$ and $\{P_{n_s}\}$ such that conditions (1)-(3) of the lemma hold with $C_n(P) = 1 \left\{ \hat{\theta}_n = \tilde{\theta}, \mu_{Y,n}(\hat{\theta}_n; P_n) \in CI_{P,n}^\beta \right\}$, we have

$$\lim_{s \rightarrow \infty} Pr_{P_{n_s}} \left\{ \hat{\mu}_{Y,n_s}(\hat{\theta}_{n_s}; P_{n_s}) \in CI_{U,-,n_s}^H | \hat{\theta}_{n_s} = \tilde{\theta}, \mu_{Y,n_s}(\hat{\theta}_{n_s}; P_{n_s}) \in CI_{P,n_s}^\beta \right\} = \alpha.$$

Recall that for $F_{TN}(Y(\theta); \mu, \Sigma_Y(\theta), \mathcal{L}, \mathcal{U})$ defined as in (A.32), $\hat{\mu}_{\alpha,n}^H$ solves

$$F_{TN}(Y_n(\hat{\theta}_n); \mu, \hat{\Sigma}_{Y,n}(\hat{\theta}_n), \mathcal{L}_n^H(\mu, \hat{\theta}_n), \mathcal{U}_n^H(\mu, \hat{\theta}_n)) = 1 - \alpha,$$

for

$$\begin{aligned}\mathcal{L}_n^H(\mu, \hat{\theta}_n) &= \max \left\{ \mathcal{L}(\hat{\theta}_n, Z_{\hat{\theta}_n, n}, \hat{\Sigma}_n), \mu - c_\alpha(\hat{\Sigma}_{Y, n}) \sqrt{\hat{\Sigma}_Y(\hat{\theta}_n)} \right\}, \\ \mathcal{U}_n^H(\mu, \hat{\theta}_n) &= \min \left\{ \mathcal{U}(\hat{\theta}_n, Z_{\hat{\theta}_n, n}, \hat{\Sigma}_n), \mu + c_\alpha(\hat{\Sigma}_{Y, n}) \sqrt{\hat{\Sigma}_Y(\hat{\theta}_n)} \right\}.\end{aligned}$$

The proof of Proposition 5 shows that $F_{TN}(Y_n(\hat{\theta}_n); \mu, \hat{\Sigma}_{Y, n}(\hat{\theta}_n), \mathcal{L}_n^H(\mu, \hat{\theta}_n), \mathcal{U}_n^H(\mu, \hat{\theta}_n))$ is strictly decreasing in μ , so for a given value $\mu_{Y, 0}$,

$$\hat{\mu}_{\alpha, n}^H \geq \mu_{Y, 0} \iff F_{TN}(Y_n(\hat{\theta}_n); \mu_{Y, 0}, \hat{\Sigma}_{Y, n}(\hat{\theta}_n), \mathcal{L}_n^H(\mu_{Y, 0}, \hat{\theta}_n), \mathcal{U}_n^H(\mu_{Y, 0}, \hat{\theta}_n)) \geq 1 - \alpha.$$

As in the proof of Proposition 10

$$\begin{aligned}F_{TN}(Y_n(\hat{\theta}_n); \mu_{Y, n}(\hat{\theta}_n; P_n), \hat{\Sigma}_{Y, n}(\hat{\theta}_n), \mathcal{L}_n^H(\mu_{Y, n}(\hat{\theta}_n; P_n), \hat{\theta}_n), \mathcal{U}_n^H(\mu_{Y, n}(\hat{\theta}_n; P_n), \hat{\theta}_n)) \\ = F_{TN}(Y_n^*(\hat{\theta}_n); 0, \hat{\Sigma}_{Y, n}(\hat{\theta}_n), \mathcal{L}_n^{H*}(\hat{\theta}_n), \mathcal{U}_n^{H*}(\hat{\theta}_n)),\end{aligned}$$

where

$$\begin{aligned}\mathcal{L}_n^{H*}(\hat{\theta}_n) &= \max \left\{ \mathcal{L}(\hat{\theta}_n, Z_{\hat{\theta}_n, n}^*, \hat{\Sigma}_n), -c_\alpha(\hat{\Sigma}_{Y, n}) \sqrt{\hat{\Sigma}_Y(\hat{\theta}_n)} \right\}, \\ \mathcal{U}_n^{H*}(\hat{\theta}_n) &= \min \left\{ \mathcal{U}(\hat{\theta}_n, Z_{\hat{\theta}_n, n}^*, \hat{\Sigma}_n), c_\alpha(\hat{\Sigma}_{Y, n}) \sqrt{\hat{\Sigma}_Y(\hat{\theta}_n)} \right\}\end{aligned}$$

so $\hat{\mu}_{\alpha, n}^H \geq \mu_{Y, n}(\hat{\theta}_n; P)$ if and only if $F_{TN}(Y_n^*(\hat{\theta}_n); 0, \hat{\Sigma}_{Y, n}(\hat{\theta}_n), \mathcal{L}_n^{H*}(\hat{\theta}_n), \mathcal{U}_n^{H*}(\hat{\theta}_n)) \geq 1 - \alpha$.

Lemma 8 implies that $(Y_{n_s}^*, \hat{\Sigma}_{Y, n_s}, \mathcal{L}_{n_s}^{H*}(\hat{\theta}), \mathcal{U}_{n_s}^{H*}(\hat{\theta}), \hat{\theta}_{n_s}) \rightarrow_d (Y^*, \Sigma_Y^*, \mathcal{L}^{H*}(\tilde{\theta}), \mathcal{U}^{H*}(\tilde{\theta}), \tilde{\theta})$, where $\mathcal{L}^{H*}(\tilde{\theta})$ and $\mathcal{U}^{H*}(\tilde{\theta})$ are equal to $\mathcal{L}_n^{H*}(\tilde{\theta})$ and $\mathcal{U}_n^{H*}(\tilde{\theta})$ after replacing $(X_n, Y_n, \hat{\Sigma}_n)$ with (X, Y, Σ^*) . Then by the continuous mapping theorem and (A.42),

$$\begin{aligned}& \left(F_{TN}(Y_{n_s}^*(\hat{\theta}_{n_s}); 0, \hat{\Sigma}_{Y, n_s}(\hat{\theta}_{n_s}), \mathcal{L}_{n_s}^{H*}(\tilde{\theta}), \mathcal{U}_{n_s}^{H*}(\tilde{\theta})), 1 \left\{ \hat{\theta}_{n_s} = \tilde{\theta}, \mu_{Y, n_s}(\hat{\theta}_{n_s}; P_{n_s}) \in CI_{P, n_s}^\beta \right\} \right) \\ \rightarrow_d & \left(F_{TN}(Y^*(\tilde{\theta}); 0, \Sigma_Y^*(\tilde{\theta}), \mathcal{L}^{H*}(\tilde{\theta}), \mathcal{U}^{H*}(\tilde{\theta})), 1 \left\{ \hat{\theta} = \tilde{\theta}, Y^*(\hat{\theta}) \in \left[-c_\alpha(\Sigma_Y^*) \sqrt{\Sigma_Y^*(\hat{\theta})}, c_\alpha(\Sigma_Y^*) \sqrt{\Sigma_Y^*(\hat{\theta})} \right] \right\} \right).\end{aligned}$$

Hence, by the same argument as in the proof of Proposition 10,

$$\lim_{s \rightarrow \infty} Pr_{P_{n_s}} \left\{ \mu_{Y, n_s}(\hat{\theta}_{n_s}; P_{n_s}) \in CI_{U, -, n_s}^H | \hat{\theta}_{n_s} = \tilde{\theta}, \mu_{Y, n_s}(\hat{\theta}_{n_s}; P_{n_s}) \in CI_{P, n_s}^\beta \right\} = \alpha,$$

as we aimed to show.

To prove (A.24), note that for $\widetilde{CI}_{U,+}^H = (\hat{\mu}_{\alpha,n}^H, \infty)$,

$$\hat{\mu}_{\alpha,n}^H \geq \mu_{Y,n}(\hat{\theta}_n; P) \iff \mu_{Y,n}(\hat{\theta}_n; P) \notin \widetilde{CI}_{U,+}^H$$

and thus that the argument above proves that

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} \left| Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in \widetilde{CI}_{U,+}^H | C_n^H(\tilde{\theta}; P) \right\} - (1-\alpha) \right| Pr_P \left\{ C_n^H(\tilde{\theta}; P) \right\} = 0$$

for $C_n^H(\tilde{\theta}; P)$ as in the statement of the proposition. Since

$$\sum_{\tilde{\theta} \in \Theta} Pr_P \left\{ \hat{\theta}_{n_s} = \tilde{\theta}, \mu_{Y,n_s}(\hat{\theta}_{n_s}; P_{n_s}) \in CI_{P,n_s}^\beta \right\} = Pr_P \left\{ \mu_{Y,n_s}(\hat{\theta}_{n_s}; P_{n_s}) \in CI_{P,n_s}^\beta \right\} + o(1), \quad (\text{A.43})$$

and Proposition 11 shows that

$$\liminf_{s \rightarrow \infty} \inf_{P \in \mathcal{P}_{n_s}} Pr_P \left\{ \mu_{Y,n_s}(\hat{\theta}_{n_s}; P_{n_s}) \in CI_{P,n_s}^\beta \right\} \geq 1 - \beta,$$

Lemma 6 together with (A.23) implies that

$$\liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_n} Pr_P \left\{ \hat{\mu}_{\alpha,n}^H < \mu_{Y,n}(\hat{\theta}_n; P) \right\} \geq (1-\alpha)(1-\beta) = (1-\alpha) - \beta(1-\alpha)$$

and

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} Pr_P \left\{ \hat{\mu}_{\alpha,n}^H < \mu_{Y,n}(\hat{\theta}_n; P) \right\} \leq 1 - \alpha(1-\beta) = (1-\alpha) + \beta\alpha$$

from which the second result of the proposition follows immediately. \square

Proof of Corollary 2 Note that by construction $CI_{ET,n}^H = \left[\hat{\mu}_{\frac{\alpha-\beta}{2(1-\beta)},n}^H, \hat{\mu}_{1-\frac{\alpha-\beta}{2(1-\beta)},n}^H \right]$, where $\hat{\mu}_{\frac{\alpha-\beta}{2(1-\beta)},n}^H < \hat{\mu}_{1-\frac{\alpha-\beta}{2(1-\beta)},n}^H$ provided $\frac{\alpha-\beta}{1-\beta} < 1$. Hence,

$$\begin{aligned} & Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_{ET,n}^H | C_n^H(\tilde{\theta}; P) \right\} \\ &= Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \leq \hat{\mu}_{1-\frac{\alpha-\beta}{2(1-\beta)},n}^H | C_n^H(\tilde{\theta}; P) \right\} - Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) < \hat{\mu}_{\frac{\alpha-\beta}{2(1-\beta)},n}^H | C_n^H(\tilde{\theta}; P) \right\}, \end{aligned}$$

so Proposition 12 immediately implies (A.25).

Equation (A.43) in the proof of Proposition 12 together with Lemma 6 implies that

$$\liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_n} Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_{ET,n}^H \right\} \geq \frac{1-\alpha}{1-\beta}(1-\beta) = 1-\alpha$$

so (A.26) holds. We could likewise get an upper bound on coverage using Lemma 6, but obtain a sharper bound by proving the result directly. Specifically, note that

$$\mu_{Y,n}(\hat{\theta}_n; P_n) \in CI_{ET,n}^H \Rightarrow \mu_{Y,n}(\hat{\theta}_n; P_n) \in CI_{P,n}^\beta.$$

Hence,

$$\begin{aligned} & Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_{ET,n}^H \right\} \\ = & Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_{ET,n}^H \mid \hat{\mu}_{Y,n}(\hat{\theta}_n; P_n) \in CI_{P,n}^\beta \right\} Pr \left\{ \mu_{Y,n}(\hat{\theta}_n; P_n) \in CI_{P,n}^\beta \right\}. \end{aligned}$$

By the first part of the proposition, this implies that

$$\begin{aligned} \limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} Pr_P \left\{ \mu_{Y,n}(\hat{\theta}_n; P) \in CI_{ET,n}^H \right\} &\leq \frac{1-\alpha}{1-\beta} \limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_n} Pr \left\{ \mu_{Y,n}(\hat{\theta}_n; P_n) \in CI_{P,n}^\beta \right\} \\ &\leq \frac{1-\alpha}{1-\beta}, \end{aligned}$$

so (A.27) holds as well. \square

G Additional Materials for Neighborhoods Application

This appendix provides additional details on the neighborhoods application discussed in Section 7 of the main text.

G.1 Simulation Design and Target Parameters

Our simulations take the census tract-level, un-shrunk estimates from the Opportunity Atlas as the true parameter values. The true parameter value in tract t , μ_t , thus corresponds to tract-level average household income rank in adulthood for children growing up in households at the 25th percentile of the income distribution.⁵ We simulate estimates for tract t by drawing $\hat{\mu}_t \sim N(\mu_t, \sigma_t^2)$, for μ_t the Opportunity Atlas estimate and σ_t the Opportunity Atlas standard error. We treat the draws $\hat{\mu}_t$ as independent across tracts. Hence, in each commuting zone, for \mathcal{T} the set of tracts in that CZ we generate jointly normal sets of estimates and corresponding standard errors, $\{(\hat{\mu}_t, \sigma_t) : t \in \mathcal{T}\}$. Our analysis also drops the single tract in the data where $\sigma_t > 1$, (i.e. a standard error in excess of 100 percentile points) which was located in the Seattle CZ. No other tract in the CZs we consider has a standard error larger than 0.2.

As discussed in the main text, in each commuting zone we define Θ as the set of selections containing one third of the tracts in \mathcal{T} (rounded down), $\Theta = \{\theta \subset \mathcal{T} : |\theta| = \lfloor |\mathcal{T}|/3 \rfloor\}$. For $X(\theta) =$

⁵Specifically, we focus on `kfr_pooled_pooled_p25` from in the file `tract_outcomes_simple.csv`, available at https://opportunityinsights.org/wp-content/uploads/2018/10/tract_outcomes_simple.csv, downloaded on May 10, 2022.

$\frac{1}{|\Theta|} \sum_{t \in \theta} \hat{\mu}_t$ the average estimate over tracts in θ , $\hat{\theta} = \operatorname{argmax}_{\theta} X(\theta)$ thus selects the third of tracts with the largest estimates. For c_t the number of voucher households with children reported to be living in tract t ,⁶ we define $Y(\theta)$ as a difference between $X(\theta)$ and a c_t weighted average of $\hat{\mu}_t$ over all tracts in the commuting zone $Y(\theta) = X(\theta) - \frac{\sum_t c_t \hat{\mu}_t}{\sum_t c_t}$. Our target parameter $\mu_Y(\theta) = E[Y(\theta)] = \frac{1}{|\Theta|} \sum_{t \in \theta} \mu_t - \frac{\sum_t c_t \mu_t}{\sum_t c_t}$ thus measures the difference between the weighted average outcome in the tracts where voucher households with children currently live and the average across targeted tracts.

We examine the performance of several estimators and confidence intervals for $\mu_Y(\hat{\theta})$ in each of our 50 commuting zones. To do so, for each commuting zone and each draw $\hat{\mu} = (\hat{\mu}_1, \dots, \hat{\mu}_{|\mathcal{T}|})$ we compute the corresponding estimators $\hat{\mu}_Y$ and confidence intervals CI . For each procedure considered (detailed in the next section) we report the median bias, $Med(\hat{\mu}_Y - \mu_Y(\hat{\theta}))$, the median absolute error, $Med(|\hat{\mu}_Y - \mu_Y(\hat{\theta})|)$, the coverage probability $Pr\{\mu_Y(\hat{\theta}) \in CI\}$, and the median confidence interval length, $Med(|CI|)$. Specifically, for each commuting zone and each simulation draw $s \in \{1, \dots, S\}$ for $S = 10,000$ we draw $\hat{\mu}^s$ as described above. We then form $X_s(\theta) = \frac{1}{|\Theta|} \sum_{t \in \theta} \hat{\mu}_t^s$ and set $\hat{\theta}_s = \operatorname{argmax}_{\theta} X_s(\theta)$. For $\hat{\mu}_{Y,s}$ and CI_s the resulting point estimate and confidence interval in simulation draw s , we then approximate the average coverage in that commuting zone by the sample average $\frac{1}{S} \sum_s 1\{\mu_Y(\hat{\theta}_s) \in CI_s\}$, while $Med(\hat{\mu}_{Y,s} - \mu_Y(\hat{\theta}))$ and the other medians are correspondingly approximated by sample medians across simulation draws. We record these quantities separately in each commuting zone, and Figure IV in the main text shows their distribution across commuting zones.

Note that in all cases we focus on unconditional performance measures, both because we think these are of substantive interest in this application, and because computing and reporting conditional quantities (e.g. conditional coverage probabilities $Pr\{\mu_Y(\hat{\theta}) \in CI | \hat{\theta} = \tilde{\theta}\}$) is difficult in this setting given the prohibitively large size of $|\Theta|$, which is equal to $|\mathcal{T}|$ choose $\lfloor |\mathcal{T}|/3 \rfloor$.

G.2 Fixed-Length Projection Intervals

We report confidence intervals and, where applicable, point estimates for the conventional, projection, conditional, and hybrid approaches. Following Chetty et al. (2020) we also report results based on an empirical Bayes approach. We defer discussion of the empirical Bayes approach to the next section, while the conventional and conditional approaches are as described in the main text. Due to the large size of Θ , however, we need to modify the projection (and thus hybrid) approach for this application.

Specifically, the projection approach considered in the main text ensures that the width of the projection interval for $\mu_Y(\hat{\theta})$ is proportional to $\sqrt{\Sigma_Y(\hat{\theta})}$. To compute these intervals, we need to approximate the critical value c_α , which we do by repeatedly drawing $\xi \sim N(0, \Sigma_Y)$ and setting

⁶Computed based on the 2018 Picture of Subsidized Housing dataset from the US Department of Housing and Urban Development, (US Department of Housing and Urban Development, 2018).

c_α equal to the $1 - \alpha$ quantile of $\max_\theta |\xi(\theta)| / \sqrt{\Sigma_Y(\theta)}$. In the neighborhoods application, however, solving the optimization problem $\max_\theta |\xi(\theta)| / \sqrt{\Sigma_Y(\theta)}$ requires a numerical search over Θ , and is computationally prohibitive. To sidestep this computational challenge, in this application we consider *fixed length* projection intervals which set $CI_P = [Y(\hat{\theta}) - c_\alpha^*, Y(\hat{\theta}) + c_\alpha^*]$.

We term these fixed length intervals since their length does not depend on $\hat{\theta}$. The critical value c_α^* corresponds to the $1 - \alpha$ quantile of $\max_\theta |\xi(\theta)|$, again for $\xi \sim N(0, \Sigma_Y)$. Unlike the original critical value c_α , we can easily approximate c_α^* by simulation. In particular, to compute $\max_\theta |\xi(\theta)|$ it suffices to independently draw $\xi_t \sim N(0, \sigma_t^2)$ in each tract t , for σ_t the Opportunity Atlas standard error for that tract. We then sort the tract-level noise draws ξ_t and select either the top or bottom third, whichever yields a larger average in absolute value, and take $\max_\theta |\xi(\theta)|$ equal to the resulting (absolute) average. Validity of the resulting projection confidence interval follows by the same argument as before, as does validity of hybrid estimators and intervals based on this version of projection.

G.3 Empirical Bayes and Winner’s Curse

Chetty et al. (2020) focus on what they term forecast-unbiased estimates. These correspond to posterior means from a correlated random effects model which treats mobility as normally distributed conditional on a set of observable tract characteristics, with a mean that changes linearly in the tract characteristics and a constant variance. Specifically, for W_t the characteristics of tract t , these estimates correspond to posterior means under the prior π that takes μ_t independent across tracts, with

$$\mu_t | W_t \sim N(W_t' \beta, \omega^2). \tag{A.44}$$

They then plug in estimates of ω and β , so in our simulations we do the same.

If we take the model (A.44) seriously and abstract from estimation of ω and β (for instance because the number of tracts is large and we plug in consistent estimates), Bayesian posterior means solve the winner’s curse problem under the prior. Specifically, note that the posterior mean for μ_t given the vector of estimates $\hat{\mu}$ is simply the mean given $\hat{\mu}_t$, $E_\pi[\mu_t | \hat{\mu}] = E_\pi[\mu_t | \hat{\mu}_t]$. The law of iterated expectations implies, however, that $E_\pi[\mu_t | \hat{\mu}]$ is unbiased for μ_t conditional on $\hat{\mu}$, so for any set \mathcal{E} such that $Pr_\pi\{\hat{\mu} \in \mathcal{E}\} > 0$,

$$E_\pi[\mu_t - E_\pi[\mu_t | \hat{\mu}_t] | \hat{\mu} \in \mathcal{E}] = 0.$$

Likewise, since we model $\hat{\mu}_t$ as normally distributed conditional on μ_t , the posterior mean is also the posterior median, so

$$Pr_\pi\{E_\pi[\mu_t | \hat{\mu}_t] > \mu_t | \hat{\mu} \in \mathcal{E}\} = \frac{1}{2},$$

and $E_\pi[\mu_t | \hat{\mu}_t]$ is median-unbiased under the prior conditional on the event $\{\hat{\mu} \in \mathcal{E}\}$. Note, however,

that selection of a particular set of target tracts can be written as an event $\{\hat{\mu} \in \mathcal{E}\}$, so this argument implies that Bayesian posterior means are immune to the winner’s curse under the prior. This depends crucially on the prior, however, since if we calculate the outer probability with respect to some other distribution of effect sizes $\tilde{\pi} \neq \pi$, we typically have

$$Pr_{\tilde{\pi}}\{E_{\pi}[\mu_t|\hat{\mu}_t] > \mu_t|\hat{\mu} \in \mathcal{E}\} \neq \frac{1}{2}.$$

G.4 Additional Figure for Movers Application

Figure 9 plots conventional, conditional, and projection intervals for the Opportunity Atlas application described in the main text.

In our main analysis we compare a simple average over targeted tracts to a weighted average over tracts in the same commuting zone, weighting by the number of voucher-recipient households with children. It may also be of interest to treat the targeted and non-targeted tracts more symmetrically, comparing a simple average over the target tracts to a simple average over the commuting zone as a whole. Formally, this corresponds to the alternative target parameter $\bar{\mu}_Y(\theta) = \frac{1}{|\theta|} \sum_{t \in \theta} \mu_t - \frac{1}{|\mathcal{T}|} \sum_t \mu_t$. Figures 10 and 11 report results for this alternative target in the Opportunity Atlas data. These results are qualitatively similar to our baseline results, with the notable exception that the coefficients are substantially smaller. The smaller coefficients from under this weighting reflects that voucher recipient households are located in neighborhoods with below-average mobility at baseline. Hence, a simple average of mobility across tracts in a commuting zone yields a larger average economic mobility than the weighted average we consider in our main results, and thus a smaller contrast with the targeted tracts

G.5 Evaluating the Normality Assumption in Empirical Bayes

As discussed in Appendix G.3, Bayesian approaches are immune to the winner’s curse in settings where the distribution of true effects is correctly described by the prior. In Section 7 of the main text we saw that empirical Bayes credible sets display large coverage distortions in simulations calibrated to the Opportunity Atlas data, which suggests that the normal approximation is not entirely reliable in this application. This appendix formally investigates the quality of the normal approximation in this application and its relationship to the coverage of empirical Bayes credible sets.

We begin by considering formal tests for normality in each of the 50 CZs used in our simulation. Specifically, the normal prior used by empirical Bayes specifies that

$$\mu_t|W_t \sim N(W_t' \beta, \omega^2),$$

for W_t a set of (observed) track-level covariates. We observe tract-level estimates $\hat{\mu}_t \sim N(\mu_t, \sigma_t^2)$

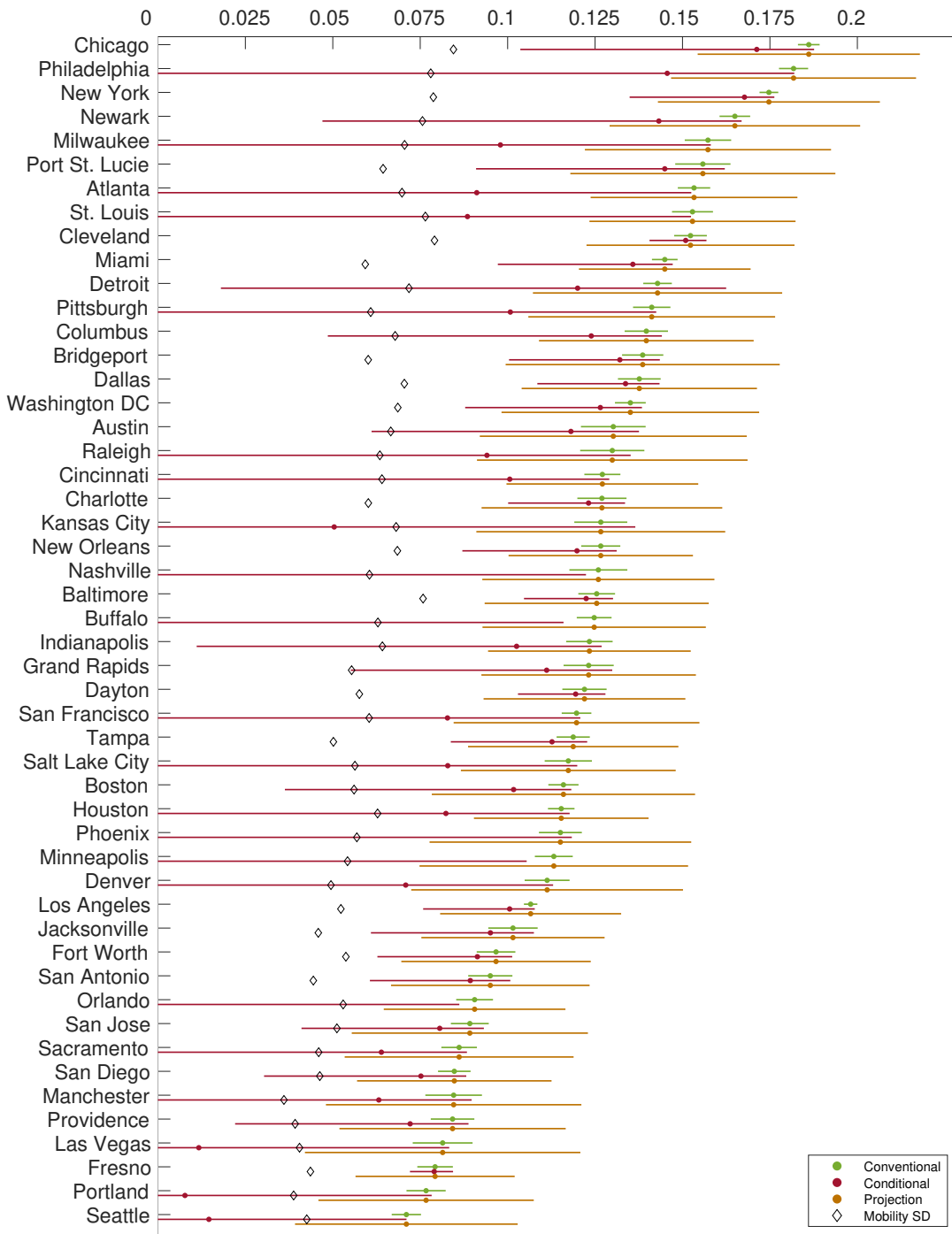


Figure 9: Estimates and confidence intervals for average economic mobility for selected census tracts based on the Chetty et al. (2020) Opportunity Atlas, relative to the within-CZ average, weighted by number of voucher recipient households with children. CZs are ordered by the magnitude of the conventional 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.

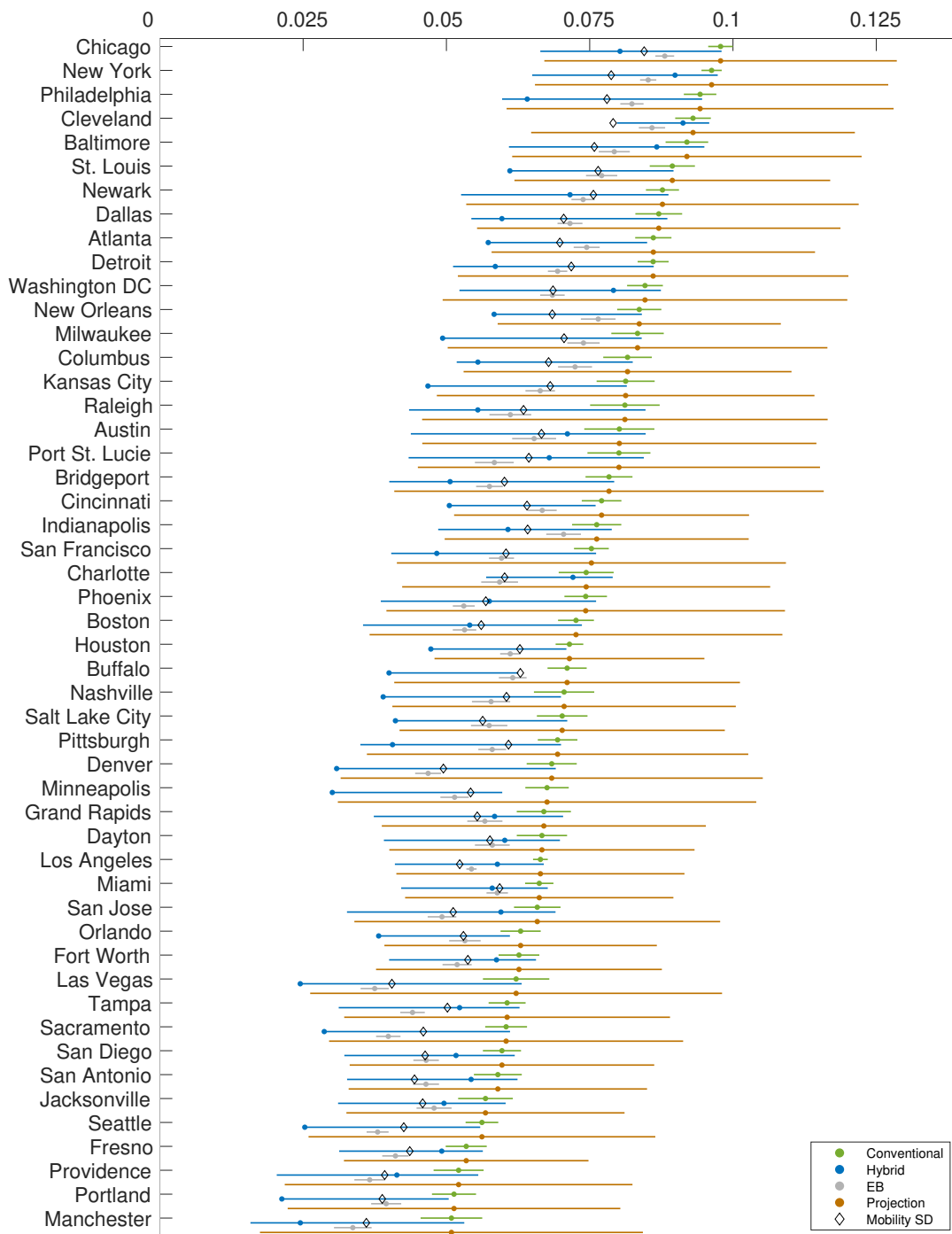


Figure 10: Estimates and confidence intervals for average economic mobility for selected census tracts based on the Chetty et al. (2020) Opportunity Atlas, relative to the within-CZ unweighted average. CZs are ordered by the magnitude of the conventional estimate. A coefficient of 0.1 implies that the 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.

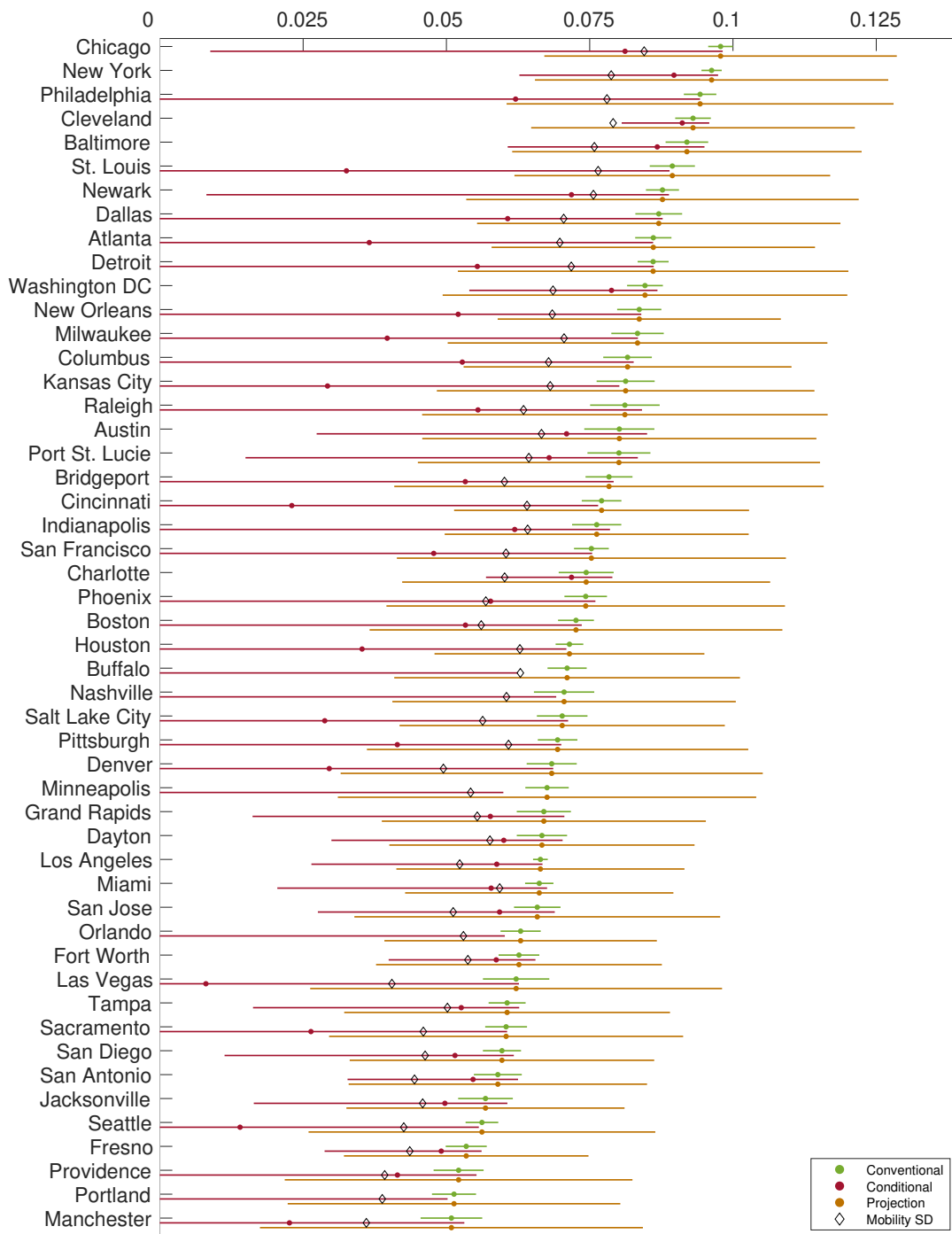


Figure 11: Estimates and confidence intervals for average economic mobility for selected census tracts based on the Chetty et al. (2020) Opportunity Atlas, relative to the within-CZ unweighted average. CZs are ordered by the magnitude of the conventional estimate. A coefficient of 0.1 implies that the 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.

for σ_t^2 known. If we knew β and ω we could thus compute

$$\frac{\hat{\mu}_t - W_t' \beta}{\sqrt{\omega^2 + \sigma_t^2}},$$

which would follow a standard normal distribution if the normal prior were correct. Since β and ω are in fact unknown, within each CZ we construct an estimate $\hat{\beta}$ by regressing $\hat{\mu}_t$ on W_t , and construct an estimate $\hat{\omega}^2$ as the average squared residual minus the average of σ_t^2 (or zero in the case where $\hat{\omega}^2$ would be negative). We conduct a Kolmogorov-Smirnov (KS) test for whether

$$\frac{\hat{\mu}_t - W_t' \hat{\beta}}{\sqrt{\hat{\omega}^2 + \sigma_t^2}}$$

follows a standard normal distribution, where to account for estimation of the parameters we compare the KS statistic to a bootstrap critical value, obtained via parametric bootstrap imposing the normal model. The resulting KS test rejects normality in 37 CZs at the 5% level, and in 27 CZs in the 1% level.

We further examine the relationship between the quality of the normal approximation and the performance of empirical Bayes. Specifically, we calculate the rank correlation between the coverage of empirical Bayes credible sets and the KS statistic across the 50 CZs, obtaining a correlation of -0.15 (with a p-value of 0.13). This correlation increases to -0.21 (with a p-value of 0.033) when we recenter the KS statistic in each CZ around the mean of its bootstrap distribution. Hence CZs with a larger KS statistic tend to have worse coverage for empirical Bayes credible sets, though the low correlation suggests that the KS statistic does not fully capture the features of the distribution most important for determining the coverage of empirical Bayes credible sets.

Index	Treatment	Description
0	0	Control group with no matched donations
Match ratio		
1	1:1	An additional dollar up to the match limit
2	2:1	Two additional dollars up to the match limit
3	3:1	Three additional dollars up to the match limit
Match size		
1	\$25,000	Up to \$25,000 is pledged
2	\$50,000	Up to \$50,000 is pledged
3	\$100,000	Up to \$100,000 is pledged
4	Unstated	The pledged amount is not stated
Ask amount		
1	Same	The individual is asked to give as much as their largest past donation
2	25% more	The individual is asked to give 25% more than their largest past donation
3	50% more	The individual is asked to give 50% more than their largest past donation

Table 4: Treatment arms for Karlan and List (2007). Individuals were assigned to the control group or to the treatment group, in the ratio 1:2. Treated individuals were randomly assigned a match ratio, a match size and an ask amount with equal probability. There are 36 possible combinations, plus the control group. The leftmost column specifies a reference index used throughout this section for convenience.

H Application: Charitable Giving

Karlan and List (2007) partner with a political charity to conduct a field experiment examining the effectiveness of matching incentives at increasing charitable giving. In matched donations, a lead donor pledges to ‘match’ any donations made by other donors up to some threshold, effectively lowering the price of political activism for other donors.

Karlan and List (2007) use a factorial design. Potential donors, who were previous donors to the charity, were mailed a four page letter asking for a donation. The contents of the letter were randomized, with one third of the sample assigned to a control group that received a standard letter with no match. The remaining two thirds received a letter with the line “now is the time to give!” and details for a match. Treated individuals were randomly assigned with equal probability to one of 36 separate treatment arms. Treatment arms are characterized by a match ratio, a match size, and an ask amount, for which further details are given in Table 4. The outcome

Treatment	Average donation	Standard error	95% CI
(1,3,2)	1.52	0.35	[0.83,2.20]
(2,1,3)	1.51	0.46	[0.61,2.41]
(2,1,1)	1.42	0.39	[0.66,2.19]
(3,1,3)	1.40	0.36	[0.70,2.11]

Table 5: The average donations for the four best treatment arms according to the data, $n=50,083$. Treatments are indexed by the indicators for (Match ratio, Match size, Ask amount) defined in Table 4. The reported 95% confidence intervals are the conventional ones that do not take selection into account.

of interest is the average dollar amount that individuals donated to the charity in the month following the solicitation.

In total, 50,083 individuals were contacted, of which 16,687 were randomly assigned to the control group, while 33,396 were randomly assigned to one of the 36 treatment arms. The (unconditional) average donation was \$0.81 in the control group and \$0.92 in the treatment group. Conditional on giving, these figures were \$45.54 and \$44.35, respectively. The discrepancy reflects the low response rate; only 1,034 of 50,083 individuals donated.

Table 5 reports average revenue from the four best-performing treatment arms, along with standard errors and conventional confidence intervals. The point estimates for the best-performing arm suggest that a campaign that promises a dollar-for-dollar match up to \$100,000 in donations and asks individuals to donate 25% more than their largest past donation raises \$1.52 per potential donor, on average, with a confidence interval of \$0.83 to \$2.20. This estimate and confidence interval are clearly subject to winner’s curse bias, however: we are picking the best-performing arm out of 37 in the experiment, which will bias our estimates and confidence intervals upward.

Simulation Results To investigate the extent of winner’s curse bias and the finite-sample performance of our corrections, we calibrate simulations to this application. We simulate datasets by resampling observations with replacement from the Karlan and List (2007) data (i.e. by drawing nonparametric bootstrap samples). In each simulated sample we re-estimate the effectiveness of each treatment arm, pick the best-performing arm, and study the performance of estimates and confidence intervals, treating the estimates for the original Karlan and List (2007) data as the true values. The underlying data here are non-normal and we re-estimate the variance in each simulation draw. Hence, these results also speak to the finite-sample performance of the normal approximation. We report results based on 10,000 simulation draws.

Since revenue does not account for the cost of the fund-raising campaign, it is impossible for the solicitation to raise a negative amount. We therefore set the parameter space for $\mu(\hat{\theta})$

Winner											
(1,3,2)	(1,4,2)	(1,4,3)	(2,1,1)	(2,1,3)	(2,2,2)	(2,3,3)	(2,4,1)	(2,4,2)	(3,1,1)	(3,1,3)	(3,3,1)
16.0%	11.4%	1.3%	13.0%	18.9%	10.8%	1.3%	1.5%	2.8%	5.1%	10.0%	3.6%

Table 6: Frequency of simulation replications where each treatment is estimated to perform best in simulations calibrated to Karlan and List (2007). Treatments are indexed by the indicators for (Match ratio, Match size, Ask amount) defined in Table 4. 31 of the 37 treatments are best in at least one replication; those that won in at least 1% of simulated samples are reported.

	Estimate		
	Conventional	Median unbiased	Hybrid
Median bias	0.61	-0.18	-0.18
Probability bias	0.50	-0.07	-0.07
Median absolute error	0.61	0.65	0.64

Table 7: Performance measures for alternative estimators in simulations calibrated to Karlan and List (2007). Probability bias is $Pr\{\hat{\mu}^{trim} > \mu(\hat{\theta})\} - \frac{1}{2}$.

to \mathbb{R}_+ , and trim the point estimators and the confidence intervals at zero, $\hat{\mu}^{trim} \equiv \max\{0, \hat{\mu}\}$ and $CS^{trim} = [0, \infty) \cap CS$. This trimming does not affect the coverage of the confidence intervals, and also preserves the α -quantile unbiasedness of the estimators so long as the true value $\mu(\hat{\theta})$ is greater than zero.

There is substantial variability in the “winning” arm: 31 of the 37 treatments won in at least one simulation draw and 12 treatment arms won in at least 1% of simulated samples. Table 6 lists these 12 treatments. The variability of the winning arm suggests that there is scope for a winner’s curse in this setting.

Table 7 examines the performance of conventional, median unbiased, and hybrid estimates, reporting (unconditional) median bias, probability bias ($Pr\{\hat{\mu}^{trim} > \mu(\hat{\theta})\} - \frac{1}{2}$), and median absolute error. Trimming the estimators at zero does not affect the reported performance measures. Conventional estimates suffer from substantial bias in this setting: they have a median bias of \$0.61, and over-estimate the revenue generated by the selected arm 100% of the time, up to rounding. The median unbiased and hybrid estimators substantially improve both measures of bias, though given the finite-sample setting they do not eliminate it completely and are both somewhat downward biased, though to a lesser degree.⁷ All three estimators perform similarly in terms of median absolute error.

⁷This is a particularly challenging setting for the normal approximation, as the outcomes distribution is highly skewed due to the large number of zeros. In particular, there are on average only 20 nonzero outcomes per non-control treatment (out of approximately 930 observations in each treatment group).

Tables 8 and 9 report results for confidence intervals. Specifically, we consider the conventional, projection, conditional, and hybrid confidence intervals with nominal coverage 95%. Table 8 reports unconditional coverage and median length, while Table 9 reports conditional coverage probabilities given $\hat{\theta}$ values among the 12 treatments listed in Table 6. Conventional confidence intervals slightly undercover unconditionally, with coverage 92%. Their conditional coverage varies depending on which treatment is the winner. If the winning treatment is one of the six best-performing treatments, the conditional coverage is at least 95%, while otherwise the conventional confidence intervals under-cover with coverage probability as low as 65%. Projection

	Unconditional coverage	Median length	
		Trimmed	Untrimmed
Conventional CS	0.92	1.88	1.88
CS_P	1.00	3.08	3.08
CS_{ET}	0.97	2.69	5.91
CS_{ET}^H	0.97	2.52	2.56

Table 8: Unconditional coverage probabilities of the confidence intervals in simulations calibrated to Karlan and List (2007). Unconditional median lengths are reported for the trimmed and untrimmed confidence intervals.

Treatment θ	Average donation $\mu(\theta)$	Conditional coverage			
		Conventional CS	CS_P	CS_{ET}	CS_{ET}^H
(1,3,2)	1.52	0.95	1	0.98	0.98
(2,1,3)	1.51	0.97	1	0.97	0.97
(2,1,1)	1.42	0.94	1	0.97	0.97
(3,1,3)	1.40	0.95	1	0.97	0.97
(2,2,2)	1.34	0.96	1	0.97	0.98
(1,4,2)	1.27	0.99	1	0.97	0.97
(3,3,1)	1.26	0.84	1	0.96	0.97
(3,1,1)	1.24	0.89	1	0.97	0.97
(2,4,2)	1.22	0.79	1	0.99	0.99
(2,3,3)	1.12	0.65	1	0.98	0.98
(2,4,1)	1.10	0.81	1	0.97	0.97
(1,4,3)	1.03	0.78	1	0.96	0.97

Table 9: Conditional coverage probabilities, $Pr\{\mu(\hat{\theta}) \in CS^{trim} | \hat{\theta} = \theta\}$, of the confidence intervals for each of the 12 treatments in Table 6. The treatments are sorted according to the average donation.

Treatment (1,3,2)		Estimates	Equal-tailed CI
Conventional		1.52	[0.83,2.20]
Projection		–	[0.40,2.63]
Conditional	– trimmed	0	[0, 1.42]
	– untrimmed	-7.49	[-47.66,1.42]
Hybrid		0.20	[0.19,1.47]

Table 10: Conventional and bias-corrected estimates and confidence intervals for best-performing treatment in Karlan and List (2007) data.

confidence intervals over-cover unconditionally and conditionally for these treatments, with coverage 100%. Conditional and hybrid confidence intervals slightly over-cover, with unconditional and conditional coverage about 97%, and have unconditional median (trimmed) length around 35% larger than conventional intervals and around 20% shorter than projection intervals. It is important to emphasize, however, that the conditional coverage for projection and hybrid intervals is particular to the data generating process considered here: as illustrated in Figure IV of the main text, these intervals do not ensure conditional coverage in general.

The median length of conditional intervals more than doubles if we leave their lower bound untrimmed. In contrast, the median length of the hybrid confidence intervals is basically unaffected by trimming. This is because despite the similarity of their upper bounds, the lower bound of the conditional confidence intervals tends to be negative and substantially lower than the lower bound of the hybrid confidence intervals. In other words, if the parameter space is unconstrained, the hybrid confidence intervals are substantially shorter than conditional confidence intervals. The good performance of the hybrid approach in applications with unconstrained parameter space is encouraging, and in line with the results in Section 2.

Empirical results Returning to the Karlan and List (2007) data, Table 10 reports corrected estimates and confidence intervals for the best-performing treatment in the experiment. We repeat the conventional estimate and confidence interval for comparison. The median unbiased estimate makes an aggressive downwards correction to the conventional estimate, suggesting negative revenue (-\$7.49) from the winning arm if not trimmed. The conditional confidence interval is tight, ranging from 0 to \$1.42, if trimmed at zero, and otherwise extremely wide, ranging from -\$47.66 to \$1.42. The hybrid estimate also shifts the conventional estimate downwards, but much less so. Moreover, the hybrid confidence interval is no wider than the conventional interval, and excludes both zero and the conventional estimate. These results suggest that future fundraising campaigns deploying the winning strategy in the experiment are likely to raise some revenue, but substantially less than would be expected based on the conventional estimates.

Conditional inference seems potentially natural in this application. The data highlight an interpretable combination of treatment parameters (1:1 match, \$100,000 pledged, with an ask 25% above an individual’s highest past donation) as best-performing, raising the question of what we can conclude about this particular treatment, given that it was the best in the experiment. This is precisely the question answered by the conditional approach. By contrast, while the hybrid approach ensures correct coverage on average across different “winning” treatments which could arise, it offers no guarantees given the particular winner observed in the Karlan and List (2007) data.

References

- Andrews, D., X. Cheng, and P. Guggenberger (2020). Generic results for establishing the asymptotic size of confidence sets and tests. *Journal of Econometrics* 218, 496–531.
- Andrews, I., T. Kitagawa, and A. McCloskey (2021). Inference after estimation of breaks. *Journal of Econometrics* 224, 39–59.
- 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.
- Eaton, M. L. (1967). Some optimum properties of ranking procedures. *Annals of Mathematical Statistics* 38(1), 124–137.
- 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.
- 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.
- Lehmann, E. and H. Scheffé (1955). Completeness, similar regions, and unbiased estimation: Part ii. *Sankhyā: The Indian Journal of Statistics* 15(3), 219–236.
- Lehmann, E. L. (1966). On a theorem of bahadur and goodman. *Annals of Mathematical Statistics* 37(1), 1–6.
- Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses* (Third ed.). Springer.
- Pfanzagl, J. (1994). *Parametric Statistical Theory*. De Gruyter.

Tibshirani, R. J., A. Rinaldo, R. Tibshirani, and L. Wasserman (2018). Uniform asymptotic inference and the bootstrap after model selection. *Annals of Statistics* 46(3), 1255–1287.

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

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.