Appeared in Journal of Labor Economics, Volume 35, Number S1, pp. S99-S147

LOCAL INSTRUMENTS, GLOBAL EXTRAPOLATION: EXTERNAL VALIDITY OF THE LABOR SUPPLY-FERTILITY LOCAL AVERAGE TREATMENT EFFECT

James Bisbee
Rajeev Dehejia
Cristian Pop-Eleches
Cyrus Samii

January 16, 2017

Abstract

We investigate the external validity of local average treatment effects (LATEs), specifically Angrist and Evans' (1998) use of same sex of the two first children as an instrumental variable for the effect of fertility on labor supply. We estimate their specification in 139 country-year censuses using Integrated Public Use Micro Sample International data. We compare each country-year's actual LATE to the extrapolated LATE from other country-years. We find that, with a sufficiently large reference sample, we extrapolate the treatment effect reasonably well, but the degree of accuracy depends on the extent of covariate similarity between the target and reference settings.

1. Introduction

In this paper we examine the extent to which we can extrapolate an instrumental variables (IV) local average treatment effect (LATE) from a quasi-experimental evidence base to new contexts of interest; i.e., the external validity of IV. In light of Imbens and Angrist (1994), Angrist, Imbens, and Rubin (1996) and the subsequent local average treatment effect literature, the empirical relevance of a LATE is limited not just to the time and place of the data used, but also to the subpopulation of compliers. In this sense, it is doubly local. Yet it is natural for social scientists to hope that LATEs are causal effects that are relevant beyond their specific context to other settings of policy interest.

We address the tension between the localness of the empirical analysis and the desire to generalize the results by extending the Angrist and Evans (1998) IV strategy from their original data set (the United States in 1980 and 1990) to 139 country-year censuses from the Integrated Public Use Micro Sample International (IPUMS-I; Minnesota Population Center 2015) data. The censuses span the world geographically (as listed in Appendix Table A-1) and cover five decades from 1960 to 2010. We use results from Abadie (2003) and Angrist and Fernadez-Val (2010) to characterize the complier population in each country-year sample in terms of covariates. We use these characteristics to extrapolate the treatment effect from a given country-year or -years (the "reference" context) to a country-year of hypothetical interest (the "target" country-year). We do so by calibrating estimates from the reference context to the complier distribution of covariates in the target context. Our approach extends current methods for extrapolation by incorporating macro-level covariates along with micro-level covariates. In principle one could extrapolate a reference LATE to various subpopulations in the target country. We focus on the complier subpopulation because we can directly identify an internally valid benchmark (namely the target country-year LATE) against which to compare the extrapolation. Using our approach, one could also extrapolate both the target and reference LATEs to average treatment effects (ATE's), hence compare an extrapolated reference ATE to the target ATE (see Appendix D).

The exercise connects to three interrelated literatures within labor and development economics. First, it relates to Lalonde (1986) and the papers that followed from it (see Heckman et al. 1997, 1998, 1999 and Dehejia and Wahba 1998, 2002 *inter alia*). By using an external reference sample to estimate the treatment effect in a setting where we already possess a plausibly internally valid estimate of the treatment effect, we adopt Lalonde's basic template. In addition, we are interested in characterizing when an externally extrapolated result is likely to provide a reliable estimate of the treatment effect (in the spirit of Heckman, Ichimura, Smith, and Todd 1998).

Second, a handful of papers has applied the Angrist-Evans design to data from other countries. Cruces and Galiani (2007) study fertility and labor supply in Mexico and Argentina, and find that the US results generalize to these contexts. Ebenstein (2009) studies Taiwan and concludes that, due to stronger sex preferences, the instrumental variables estimate is closer to the ATE than it is in the United States.

Third, our work also connects to a small but growing literature that has begun to grapple with issues of external validity in randomized controlled trials (RCT's); see *inter alia* Allcott (2014), Dehejia, Pop-Eleches, and Samii (2015), Gechter (2015), Prichett and Sandefur (2013), and Vivalt (2015). In no small part spurred by Lalonde (1986) and the ensuing literature, RCT's have been used extensively, indeed globally, to estimate the causal impact of a broad range of policy interventions. A tension similar to that of instrumental variables resonates in this exercise.

Each RCT evaluation is also a local average treatment effect, perhaps in the Imbens-Angrist (1994) sense if issues of non-compliance arise, but also in the broader sense of evaluating an intervention in a specific time and place on a not always representative set of experimental subjects. At the same time, there is an intellectual agenda in which an accumulation of experimental evidence might allow one to reach more general conclusions regarding the efficacy of certain policies or the relevance and validity of specific economic models (an issue which we examine in Dehejia, Pop-Eleches, and Samii 2015).

Following Dehejia, Pop-Eleches, and Samii (2015), we test the performance of our approach using four different steps: first, we document substantial effect heterogeneity in both the first stage and instrumental variable estimates; second, we characterize which differences between a target and reference context most strongly predict errors in our method; third, we look at how performance improves as we accumulate a larger evidence base from which to extrapolate; and fourth we compare the errors associated with our method to the errors produced by using an unidentified OLS estimate from within the target context.

The paper begins by outlining our identifying assumptions and empirical approach in Section 2 with simulation results included in Appendix B. We then examine the ability of our method to accurately extrapolate causally-identified relationships to new settings in Sections 3 through 6. We include a range of robustness checks in Section 7 that focus on potential violations of internal validity while our eighth section concludes.

2. Methods

We begin by defining the conditions and methods for extrapolating from an instrumental variables estimate to the causal effect in a target population.[1] We focus on using covariate-adjusted local average treatment effects (LATEs) estimated from a set of reference contexts to identify and estimate the LATE in a target context, where the LATEs are the complier average causal effects in the respective populations (Angrist et al. 1996). The IPUMS-I data allow us to estimate LATEs in different populations defined by country and year. As such, we can use these data to conduct benchmarked comparisons between actual LATEs in a given context and what we would obtain by extrapolation from other contexts. Nonetheless, the methods that we apply here are straightforward to generalize for extrapolation to other types of populations. Angrist and Fernandez-Val (2010) and Hartman et al. (2015) provide useful discussions of defining targets for extrapolating causal effects.

We suppose that a sample of units is drawn and distributed over a sample of locations. Building on Angrist and Fernandez-Val (2010), a randomly sampled unit i falling in location c is characterized by a treatment variable, $D_{ic} \in \{0,1\}$, and potential outcomes, $Y_{ic}(1)$ and $Y_{ic}(0)$, that would obtain depending on whether the unit is assigned to treatment ($D_{ic} = 1$) versus control ($D_{ic} = 0$).[2] In our application, the treatment is an indicator for whether a mother has more than two children, restricting consideration to the subpopulation of women with at least two children. Observed outcomes for unit i in context c are given by,

[1] This section and the analysis in this paper draw extensively on Dehejia, Pop-Eleches, and Samii (2015), where we discuss external validity issues in the context of experiments.

[2] Our notation differs slightly from Angrist and Fernandez-Val (2010) in that we use the subscripts, i and c, to denote terms associated with random draws of individuals and locations, whereas terms without subscripts denote population functionals/parameters or indices.

$$Y_{ic} = \alpha_c + r_{ic}\, D_{ic} + \eta_{ic} \tag{1}$$

where α_c = E[$Y_{ic}(0)$], $\eta_{ic} = Y_{ic}(0) - \alpha_c$, and $r_{ic} = Y_{ic}(1) - Y_{ic}(0)$ is the unit-level causal effect of $D_{ic.}$ Expectations are taken with respect to distributions in location c.

We also define an instrument, $Z_{ic} \in \{0,1\}$, that affects treatment assignment. Thus, we have potential treatment assignments, $D_{ic}(1)$ and $D_{ic}(0)$, that correspond to the treatment values that would obtain for a unit under Z_{ic} = 1 versus Z_{ic} = 0, respectively. In our application, the instrument is an indicator for whether the sexes of the first two children are the same. The realized treatment for unit i in location c is given by,

$$D_{ic} = \gamma_c + p_{ic}Z_{ic} + v_{ic} \tag{2}$$

where γ_c = E[$D_{ic}(0)$], $v_{ic} = D_{ic}(0)$- γ_c , and $p_{ic} = D_{ic}(1) - D_{ic}(0)$. Again, expectations are taken with respect to distributions in location c.

Units are characterized by unit-level covariates, X_{uic} , as well as covariates that characterize their location, X_{pc} , and we collect these covariates into the vector $X_{ic} = (X'_{uic}, X'_{pc})'$.[3] These covariates play a central role in extrapolation. We assume throughout that conditions required for identifying the conditional LATEs in the different locations hold (Angrist and Fernandez-Val 2010, p. 7). These include, for all values of x that might occur in the populations and locations to which we want to extrapolate,

C1(a) *Conditional independence and exclusion*:

[3] One may wonder how our approach relates to conventional meta-analysis (Glass, 1976). In our view, it differs in two important ways: (1) our primary interest is in extrapolating to specific target populations, rather than describing the distribution of effects, although we do conduct some descriptive analyses below; and (2) we incorporate micro-level data in our analysis rather than only working with macro-level data that varies at the same level as the effect estimates themselves.

$$(Y_{ic}(1), Y_{ic}(0), D_{ic}(1), D_{ic}(0)) \perp\!\!\!\perp Z_{ic} | X_{ic} = x,$$

C1(b) *Valid conditional first stage*:

$$E[p_{ic}|X_{ic} = x] \neq 0 \textit{ and } 0 < \Pr[Z_{ic} = 1|X_{ic} = x] < 1, \text{ and}$$

C1(c) *Conditional monotonicity*:

$$\Pr[D_{ic}(1) \geq D_{ic}(0)|X_{ic} = x] = 1 \textit{ or } \Pr[D_{ic}(1) \leq D_{ic}(0)|X_{ic} = x] = 1.$$

Without loss of generality, we focus on the case where $\Pr[D_{ic}(1) \geq D_{ic}(0)|X_{ic} = x] = 1$.

We also define an indicator, W_{ic}, for whether a unit is in the target population to which we want to extrapolate, $W_{ic} = 1$, or whether it is from a reference population for which we have LATE estimates, $W_{ic} = 0$.

The covariate-specific LATE in the target population is defined as

$$\Delta_Z(x_u, x_p, 1) = E[r_{ic}|D_{ic}(1) > D_{ic}(0), X_{uic} = x_u, X_{pc} = x_p, W_{ic} = 1], \tag{3}$$

and in a reference set population as

$$\Delta_Z(x_u, x_p, 0) = E[r_{ic}|D_{ic}(1) > D_{ic}(0), X_{uic} = x_u, X_{pc} = x_p, W_{ic} = 0]. \tag{4}$$

Substantively, equations (3) and (4) characterize the average treatment effect in the target and reference populations for compliers with $X_{ic} = (x_u', x_p')'$.

Given covariate-specific LATEs for a target location ($W_{ic} = 1$), for which the location-level covariates equal x_1, the marginal LATE for the target location is given by (Froelich 2007)

$$\Delta_Z(1) = \int_{\mathcal{X}_u} \Delta_Z(x_u, x_1, 1) dF(x_u|D_{ic}(1) > D_{ic}(0), W_{ic} = 1), \tag{5}$$

where $\mathcal{X}_u$ is the support for unit-level covariates and $F(a|b)$ is the distribution function for *a* conditional on *b*.

We now state the assumptions needed to allow for a LATE from a reference population to be transported to the target population conditional on covariates. Our analysis builds on Hotz et al. (2005), although our setup is different in that we consider the accumulation of locations and therefore we incorporate macro-level covariates into the identification analysis.

Proposition 1 (Identification): *Suppose C1 holds across all populations. Furthermore, suppose*

C2 *Unconfounded location:* $(Y_{ic}(1), Y_{ic}(0)) \perp\!\!\!\perp W_{ic} | D_{ic}(1) > D_{ic}(0), X_{ic} = x$*; and*

C3 *Covariate overlap:* $0 < \Pr[W_{ic} = 0 | D_{ic}(1) > D_{ic}(0), X_{ic} = x] < 1$

for all x in the support of X_{ic} *in the target population. Finally, suppose we have data to estimate* $\Delta_z(x_u, x_p, 0)$ *for all x in the support of* X_{ic} *in the target population. Then,* $\Delta_z(1)$ *is identified and can be estimated from the data.*

Proof: Under C1-C3, we have

$$\Delta_z(x_u, x_p, 1) = \Delta_z(x_u, x_p, 0) \tag{6}$$

for all $x = (x_u', x_p')'$ in the support of X_{ic} in the target population, in which case

$$\Delta_Z(1) = \int_{X_u} \Delta_Z(x_u, x_1, 0) dF(x_u | D_{ic}(1) > D_{ic}(0), W_{ic} = 1) \tag{7}$$

■

To calibrate estimates from reference contexts to target contexts, we adopt an estimation approach based on interacted regressions.

Proposition 2 (Complier-centered interaction estimation): *Suppose conditions C1-C3 hold and that we have for the reduced form relationship*

$$Y_{ic} = \beta_0^Y + \beta_1^Y Z_{ic} + \sum_{k=1}^{K} (X_{ick}^* \phi_k^Y + Z_{ic} X_{ick}^* \lambda_k^Y) + \epsilon_{ic}^Y, \tag{8}$$

and the first stage relationship

$$D_{ic} = \beta_0^D + \beta_1^D Z_{ic} + \sum_{k=1}^{K} (X_{ick}^* \phi_k^D + Z_{ic} X_{ick}^* \lambda_k^D) + \epsilon_{ic}^D, \tag{9}$$

with X_{ick}^* being the value of the covariate X_{ick} centered on the sample complier mean in the target context (W$_{ic}$ = 1), and $E[Z_{ic}, \epsilon_{ic}^R] = E[X_{ick}, \epsilon_{ic}^R] = 0$ for $R = Y, D$. Let $\tilde{Z}_{ic}$ be the residuals from a linear regression of the sample Z_{ic} values onto the matrix of sample $(X_{ic1}^*, \dots, X_{icK}^*, Z_{ic}X_{ic1}^*, \dots, Z_{ic}X_{icK}^*)$ values. Over C contexts, each with samples of N_c units and with $W_{ic} = 0$, as C and N_c grow large,

$$plim \frac{\sum_{c=1}^{C} \sum_{i=1}^{N_c} \frac{(Y_{ic} - \bar{Y})\tilde{Z}_{ic}}{\tilde{Z}_{ic}^2}}{\sum_{c=1}^{C} \sum_{i=1}^{N_c} \frac{(D_{ic} - \bar{D})\tilde{Z}_{ic}}{\tilde{Z}_{ic}^2}} = \Delta_z(1). \tag{10}$$

Proof: By standard results for centered regression with heterogeneous effects (e.g., Imbens and Wooldridge 2009, pp. 28-30), we have for units with $W_{ic} = w$

$$\begin{aligned} & plim \sum_{c=1}^{C} \sum_{i=1}^{N_c} \frac{(R_{ic} - \bar{R})\tilde{Z}_{ic}}{\tilde{Z}_{ic}^2} \\ & \qquad = \int \frac{Cov(R_{ic}, \tilde{Z}_{ic} | X_{ic} = x)}{Var(\tilde{Z}_{ic} | X_{ic} = x)} dF(x | \, D_{ic}(1) > D_{ic}(0), W_{ic} = w). \end{aligned} \tag{11}$$

for $R_{ic} \in \{Y_{ic}, D_{ic}\}$, $\bar{R}$ the sample mean of the R_{ic} values, and $\tilde{Z}_{ic}$ formed as residuals off of X_{ic}^* values centered on the means for compliers with $W_{ic} = w$. The result then follows from the consistency of the IV estimator for the LATE (Imbens and Angrist 1994).

■

The key assumption for this estimation strategy is that we can define a linear series in covariates to account for unit-level heterogeneity in both outcomes and treatment take-up given variation in the instrument. When the covariates are limited to indicators for an exhaustive set of micro-level strata, estimation via a centered interaction regression is algebraically equivalent to the stratification reweighting used by Angrist and Fernandez-Val (2010) (see, e.g., Miratrix et al. 2012).

Proposition 2 shows that we can use 2SLS with interactions centered on the target population complier means to extrapolate from reference data to the LATE in the target population. This requires that we can estimate the means of covariates among compliers ($D_{ic}(1) > D_{ic}(0)$). For macro-level covariates, of course, compliers have the same location-level value as all other units. For the unit-level covariates, by Theorem 3.1 of Abadie (2003), we can estimate the complier means via "kappa weighting." Specifically, for our target population with $W_{ic} = 1$, we have

$$E[X_{uic}|D_{ic}(1) > D_{ic}(0), W_{ic} = 1] = \frac{E[\kappa_{ic}(X_{uic}, 1)X_{uic}\,]}{E[\kappa_{ic}(X_{uic}, 1)]}, \tag{12}$$

where

$$\kappa_{ic}(x_u, 1) = 1 - \frac{D_{ic}(1 - Z_{ic}\,)}{1 - E[Z_{ic}|X_{uic} = x_u, W_{ic} = 1]} - \frac{(1 - D_{ic})Z_{ic}}{E[Z_{ic}|X_{uic} = x_u, W_{ic} = 1]} \tag{13}$$

We use the macro-level covariates and the sample analog of expression (12) to compute the X_{ik}^{c} terms for the complier-centered interaction regression. In our applications below, we take the X_{ik}^{c} terms as fixed and therefore apply standard 2SLS inference.

The success of extrapolation rests on the unconfounded location assumption (C2) and on covariate overlap (C3). We corroborate the validity of these assumptions in our data below, and also examine the implications of violations of (C3) for extrapolation.

3. A world of LATEs: same-sex, more kids, and mothers' labor supply

3.1 The same-sex instrumental variable and IPUMS-I data

Angrist and Evans (1998) used two instrumental variables for a mother's incremental fertility: the first two children having the same sex (i.e., boy-boy or girl-girl) and a twin birth. In this paper we focus on the first. They examine the sample of married women between age 18 and 34 with two or more children from the 1980 and 1990 US IPUMS. A preference for a gender mix of children encourages mothers with the first two children of the same sex to have an additional child. The IV strategy uses that variation to look at the effect of increased fertility on labor supply. The identified local average treatment effect is the effect of fertility on labor supply for those women who have an extra child when their first two children are the same sex but would not otherwise. For the 1980 (1990) sample, same-sex leads to a 0.068 (0.070) increase in the probability of the third child (relative to approximately 0.5 of the sample that has a third child). The reduced-form effect of same-sex on whether the mother worked for pay is -0.0080 (for 1980) and -0.0053 (for 1990), with an IV estimate of -0.120 (for 1980) and -0.104 (for 1990), relative to 0.528 of the 1980 sample and 0.667 of the 1990 sample who work.

We use the IPUMS-I data (Minnesota Population Center 2015) to take the Angrist-Evans strategy to the world. The IPUMS-I data provide harmonized measures that are comparable across countries and years. Data are available for a maximum of 139 country-years, although our sample is smaller for some specifications due to missing data. Individual covariates include the mother's age, her age at the birth of her first child, her education (coded as 1=illiterate, 2=primary, 3=secondary, and 4=college or higher), and her spouse's education (coded similarly).

Summary statistics are presented in Table 1. The average age of mothers at the time of the survey in the global sample is 30.05, and the average mother's education is 1.92.

In our investigation of external validity we explore country-year level differences and whether they explain the pattern of extrapolation error. Our country-year covariates include GDP per capita, female labor force participation, the sex ratio imbalance (the number of male children divided by the number of female children minus 0.5), the total fertility rate, and the pairwise geographical and temporal distances between country-year samples. Summary statistics of these dyadic absolute differences are presented in column 2 of Table 1. We will demonstrate that these differences strongly predict the magnitude of the extrapolation error.

We focus on the same-sex instrumental variable rather than a twin birth because it is more likely that two non-twinned children can be born in the same calendar year in high-fertility countries than in the US. Same-sex has its own challenges when used as an instrument on a global scale. The first concern is sex selection. While this is not believed to be an issue with US data, for some countries in our sample (such as China) it is clearly a concern. We address this by treating sex-selectivity as a country-year covariate and examining whether it affects the IV extrapolation from reference to target. In Section 7.2, we also show that dropping potentially sex-selective countries does not significantly affect the results.

A related concern regarding instrument validity is the total fertility rate of a country. The decision to have a third child is based on preferences for sex heterogeneity is clearly less salient in countries where most families have more than three children. We confirm that our results persist when we drop countries in the 25th and 75th percentiles of the total fertility rate measure.

Similarly, we note that there are some country-year observations in which the female labor force participation rate was effectively zero. In these cases, the relationship between an additional child and the economic activity of the mother is necessarily null, regardless of the instrumentation strategy or first-stage strength. Our results are robust to the omission of these observations.

A final concern relates to exclusion restriction violations, especially for low-income countries. Butikofer (2011) has presented suggestive evidence that the gender mix of the first two children in low-income countries can directly influence a mother's labor supply through the cost associated with having a third child (see also Huber 2015 for evidence of instrument validity for US data). We address this in a similar fashion, first by examining the extent to which GDP per capita affects extrapolation error and second, in Section 7.3, by using the Huber-Mellace (2015) test to detect and drop country-years that are unlikely to satisfy instrument validity.

3.2 IPUMS-I data: first stage, compliers, and IV estimates

In this section we provide a graphical summary of the variation in the first stage and IV estimates from IPUMS-I data. Figure 1, panel a, plots the first-stage effect of same-sex on an incremental child against the standard error of the estimate. Each point represents a country-year IPUMS-I sample, and different markers are indicators for where the country-year falls in the distribution of fertility rates. Figure 1, panel b, plots the density of first-stage effects. We note that all but a handful of first stage estimates are positive: the preference for a gender mix of children seems to be global. From panel b, we note that the average first-stage effect is approximately 0.04 across countries and years. Both panels highlight the heterogeneity of the first-stage strength, an issue we return to in the cumulative analysis. The strongest first stage relationships are found among countries in the first three quartiles of the fertility rate measure (between 1.4 and 3.1 children per mother). Intuitively, the countries with small first-stage coefficients are predominantly those with the highest fertility rates above 3.1 children per mother.

Figure 2 recreates the analysis of Figure 1 but replaces the first-stage results with the IV estimates of the effect of an additional child on the mother's work status. We indicate observations not by where they fall in the fertility rate distribution but rather by where they fall in the labor force participation rate distribution. The striking difference between Figures 1 and 2 is that the IV estimates are both negative and positive. While we can see in Figure 2, panel b, that the average affect across countries and years is negative (-0.129, compared to -0.120 and -0.104 for the US in 1980 and 1990), there is a genuine mix of positive and negative coefficients. In panel a, the most significant negative IV estimates are among countries with higher levels of female labor force participation. However, as with the total fertility rate results presented above, the pattern is not systematic.

In Figure 3, we examine how the population of individuals who comply with the instrument differs from the overall population. Points are coded by geographic region. In the top two panels, we look at mother's age at the time of the survey and at the time of the first birth. The top-left panel indicates that, on average, the complier population is younger than the overall population for most country-years at the time of the survey. At the same time, there are regions (notably North America and Western Europe) where these two distributions are similar. This pattern is reversed in the top-right panel, which charts the comparison for mother's age at first birth. Here we note that compliers are consistently older than the overall population across all country-years in our data set. In the bottom two figures, we find that complier mothers and their spouses are more likely to have secondary or tertiary education than the overall population although these differences are far less pronounced.

The above analysis highlights a key aspect of our approach: heterogeneity in differences between complier and raw populations is the dimension along which we calibrate the IV treatment effects in the reference country to extrapolate to the target of interest. While discussed formally in the methodology section above, it bears emphasizing that this heterogeneity lies at the heart of what external validity means in an IV context. If differences in complier populations affect the relationship of interest, external validity may be compromised. By calibrating our reference estimates to approximate the target complier population, we remove this threat to external validity. The efficacy of this technique hinges crucially on whether the observable covariates that we can measure fully capture the latent characteristics that would otherwise undermine external validity.

4. Extrapolation error regressions

In this section, we examine the extent to which extrapolation error from reference to target country-year can be explained by covariate differences between the two contexts. Our strategy is to create all possible pairwise combinations of the country-year samples, with one country serving as the target and the other as the reference. We use the complier characteristics in the target country to calibrate the conditional-on-*x* LATEs in the reference country. Since for any two country-years, the extrapolation differs depending on which is the target country and which is the reference country, our dyads consist of all $n \times (n-1)$ pairwise permutations. For each dyad, we record the extrapolation error, E_{ij} (the target country-year *c* LATE estimate minus the extrapolated treatment effect from the reference country-year *d*), its standard error, and covariate differences between reference and target (for simplicity we assume these can be summarized simply as $\delta_{cd} = X_c - X_d$).

As in Dehejia, Pop-Eleches, and Samii (2015) we use this setup to estimate an external validity function that characterizes the extrapolation error:

$$E_{cd} = \beta' \delta_{cd} + \epsilon_{cd},$$

where we weight the regression by the inverse of the variance of the extrapolation error, constructed as the sum of variances of the component LATE estimates. In the spirit of the Heckman, Ichimura, Smith, and Todd (1998) bias function, which characterizes selection bias as a function of covariates, our goal is to characterize the reference-to-target country-year extrapolation error, while maintaining the assumption of a valid instrumental variables strategy (hence an internally valid target country-year LATE). Note that, in addition to within country-year micro covariates, δ_{ij} includes country-year level macro covariates as well, including GDP per capita, labor force participation, and total fertility rate. We do not (indeed, we cannot) use these population-level covariates in our extrapolation procedure in the dyadic estimation as they are constant for units from a single target context. Nevertheless, we are interested in exploring whether these differences between target and reference are significant predictors of extrapolation error for the cumulative analysis that follows.

Results are presented in Figure 4 and Table 2. We begin in Figure 4 by examining the univariate relationship between covariate differences and extrapolation error. For most covariates, we find that differences between target and reference context are significant predictors in the expected direction: greater reference-target differences are associated with increased extrapolation error. Since the majority of these estimates are from log-log regressions, the coefficients can be directly compared in terms of percent changes, meaning that a ten percent increase in the difference between reference and target in mother's education, spouse's education, total fertility rate, and per capita GDP is associated with a 0.02, 0.017, 0.01, and 0.007 increase in extrapolation error, respectively (relative to an average world LATE of -0.129).

Temporal and geographic distances are presented in standardized measures, implying that a 1 standard deviation increase in geographic distance (roughly 4,650km) corresponds to a 2 percent increase in extrapolation error although this is not significant at conventional levels. The effect of reference-target temporal distance is noisily estimated but seemingly negligible.

When we include all covariates simultaneously, we find in Table 2 that differences in spouse's education and total fertility rate remain significant in the full sample. The magnitudes are considerable for these covariates. A ten percent increase in the difference in mean levels of mother's age corresponds to a 0.011 increase in the absolute difference between the target and extrapolated estimates (again relative to an average world LATE of -0.129). Overall the results underline an intuitive but important result: when the target and reference countries are more similar, extrapolation error tends to be smaller. It is worth noting that these coefficients are relatively stable as we restrict the sample to successively stronger first stage targets in columns (2) to (4).

5. Accumulation of evidence

While the dyadic regressions discussed above highlight the importance of covariate differences between reference and target countries, they do not allow us to deduce how close the extrapolation comes to the target country LATE. Furthermore, while the dyadic setup is useful to explore the external validity function, it uses only a single country-year reference to predict the target, whereas in fact the available pool of reference countries is much larger in all but the first time period of our data set. In extending the available reference pool, we can also expand the covariates used in extrapolation to include both individual-level and population-level covariates. We center the cumulative reference distribution of the macro-level covariates on the target values.

There are two similar ways to conceptualize how evidence accumulates. The first uses all available data prior to the target year for any given unit. To highlight the improvement of our procedure as we increase the reference sample size, we estimate the extrapolation error associated with each reference dataset accumulated up to different calendar years. For example, consider Ecuador in 1974 as our target of interest. We calculate the prediction error associated with a reference sample comprised of only the available data in 1970, 1971, 1972, and 1973. In 1970, the reference sample size contains less than 200,000 observations from only a handful of countries. The sample increases with each additional year although we are careful to exclude Ecuador itself. (Our exclusion of the target country data in earlier periods constitutes a more difficult test of our extrapolation technique by removing the most similar reference data from the accumulated pool of observations.) We assess the prediction error in this manner for all available target country-years, averaging by year and depicting two standard deviations with vertical bars. Figure 5 depicts these results.

Several patterns become evident. First, as more evidence becomes available through an increased reference set, prediction error typically decreases. Second, for target countries later in the sample for which a larger reference set is available, prediction error tends to converge to zero. Third, the improvements in extrapolation performance can be understood in terms of both bias and variance, as indicated by mean error estimates that converge to zero with increasing precision. Finally, the earliest years exhibit dramatically imprecise extrapolation estimates. In these cases, the available reference data is too sparse to ensure unconfounded location. Recognizing that errors in excess of an absolute value of 1 are theoretically impossible, we bound the y-axis accordingly and indicate extreme errors with X's.

We adopt a different approach to visualizing the cumulative benefits of extrapolation in Figure 6 which examines cumulative improvements for each target country-year. Here, we are interested in estimating how extrapolation error changes over time, not measured by calendar year, but by temporal distance from each target. Returning to our Ecuador example, the reference populations for 1970 through 1973 would be treated as 4, 3, 2, and 1 years distant from our target. However, using the cumulative data up to 2000 when assessing Chile in 2004 would also be treated as 4 and the two prediction errors are averaged together. As in Figure 5, we indicate the increasing reference sample size with shaded points and two standard deviations with vertical bars.

While the overall results reinforce the findings discussed above, Figure 6 highlights the importance of choosing a reference set that is not too temporally distant from the target. Specifically, we find quite poor extrapolated estimates in cases where the reference population is more than 10 years separated from our target, as indicated by errors that increasingly exceed the theoretically plausible bounds of 1 and -1. However, we still document a steady decline in mean absolute error as we move closer to the target in time.

Overall, Figures 5 and 6 show that with a sufficiently large reference set, the extrapolated LATE is able to systematically replicate the actual country-year LATE with considerable precision. Given the internal validity of the IV estimates, this in turn serves as a test of the validity of our key identifying assumption of uncounfounded location. However, additional research would be needed to explore whether the weak results for earlier years are merely due to smaller reference sample sizes or reflect a change in the underlying causal relationship between fertility and female labor supply among the countries in our sample.

6. Extrapolation vs. interpolation

Is there any reason to believe that errors associated with extrapolation are systematically larger or smaller than biases associated with endogeneity? This section presents a series of comparisons between the extrapolated LATE estimates measured using different criteria and the OLS estimates within the target country. The thought experiment is trading off two possible biases: extrapolation error from the extrapolated LATE versus endogeneity bias from regressing a women's labor force status on an endogenous indicator of incremental fertility.

It is worth noting the artificiality of the exercise at the outset. We know that OLS, whether biased or not, is estimating the average treatment effect, whereas the extrapolated LATE is replicating the LATE for the target country. So even without bias, we would not expect these two to be the same. Nonetheless, we argue that the choice is not entirely artificial: a policy maker could indeed be faced with the choice of two potential biases.

The results are presented in Figure 7, where the x-axis depicts the LATE extrapolation error and the y-axis depicts the difference between the OLS estimate and the estimated LATE in the target. The left-panel averages over the extrapolated error in all available dyad pairs for each target country-year, weighting by the reference sample size. The right panel uses the accumulated evidence up to the prior year, excluding data from the target country itself. We choose the most recent cumulative extrapolated results for each target country-year under the assumption that this represents both the largest reference dataset as well as the most accurate extrapolated prediction on average. The average OLS error is indicated by the horizontal solid line while the average extrapolation error is indicated by the vertical solid line. The dotted lines represent zero average bias.

As illustrated in the left-panel, simply choosing a reference context at random yields larger errors than relying on the OLS estimate calculated in the target country-year. However, the right-panel indicates superior extrapolation performance when using the accumulated evidence as the reference. This improvement is due partly to the ability to include macro-level covariates in the extrapolation procedure using accumulated data, which we can't do in a dyadic extrapolation context.

In Appendix C, we present additional results that compare OLS estimates to a variety of different extrapolation methods, including pairing the target with the reference closest in terms of the Mahalnobis distance, geographic distance, or using the reference dyad with the best predictive accuracy in previous years for a given target country. The results suggest that, while extrapolation error remains a concern, the endogeneity bias of within country-year OLS is generally larger.

7. Robustness checks and extensions

7.1 Extrapolating using the number of children as the endogenous variable

Angrist and Evans (1998) present results using both the number of children and an indicator for more than two children as the endogenous variable. In our main results, we focus on the latter. Here we present results using the former. Figure 8 presents cumulative extrapolation error results which are similar to those presented in Figure 5. When using reference countries separated by twenty or more years from the target, estimates are noisy and tend to bounce around from year to year. But as we move closer in time to the target, estimates converge on, and are not statistically significantly different from, zero extrapolation error.

7.2 Dropping sex-selectors

The most direct challenge to the validity of the IV assumptions in our application is the practice of sex selection in some of the countries in our sample (most notably China under the one-child policy). In our main results we control for the degree of sex selection within country-years via the gender ratio measure. In this section we instead drop countries where sex selection and potential non-exogeneity of the same-sex variable is a concern (in particular, India, China, Nepal and Vietnam). Figure 9 presents the cumulative extrapolation error results which are qualitatively and quantitatively similar to those in Figure 7. After an initial "burn in" period, extrapolation error converges to zero.

7.3 Dropping country-years with invalid IV's

As an extension of Section 7.2, we rely on recent work by Kitagawa (2008) and Huber and Mellace (2014) who exploit the implications of the LATE assumptions to derive systematic tests of IV validity. Unlike Section 7.2, where our rationale for dropping sex selectors is based on indirect evidence and case study research on cultural determinants of gender heterogeneity (see Rosenzweig and Wolpin 2000 for evidence from India and Edlund and Lee 2013 for evidence from South Korea), here we employ a data-driven test for violations of the LATE assumptions. As described in Huber (2015), the LATE assumptions require that, for all y in the support of Y_{ic}:

$$f(y, D_{ic} = 1|Z_{ic} = 1) \geq f(y, D_{ic} = 1|Z_{ic} = 0),\ f(y, D_{ic} = 0|Z_{ic} = 0) \geq f(y, D_{ic} = 0|Z_{ic} = 1),$$

lest the joint densities of the compliers be less than zero. Violations of these inequalities are not enough to identify which LATE identifying assumptions fail, but they do provide evidence that Z_{ic} is not randomly assigned, defiers exist in the data and dominate the compliers, or both.

We use the procedure outlined by Huber (2015) to identify which country-years fail to satisfy the identifying assumptions necessary for the same-sex instrument to be valid.[4] One benefit we enjoy thanks to our large dataset is that our finite sample power is high enough that we are unlikely to commit Type II errors. Nevertheless, as stressed by Huber (2015), failures to reject the null cannot be taken as evidence of instrument validity. Table 4 lists the country-years with partial p-values smaller than 0.4, representing a conservative test for IV validity. We rerun our cumulative analysis on the restricted data and present the results in Figure 11, represented by light gray circles. The results are not meaningfully different from those presented above.

As a final check, we stratify the data over coarsened covariates in an attempt to see whether the IV test fails for any subset of the population. We use three bins for the educational attainment of the mother and her spouse (0 = less than high school, 1 = high school, 2 = more than high school) and a binary variable indicating whether the mother is in her 20s or her 30s at the time of the survey. We then run the IV validity test on each sub-population in each country-year, yielding as many as 18 separate p-values for evaluation (although many country-years do not have full coverage for all possible strata). Figure 10 lists the results for all 139 available country-years in the dataset, ranked by the minimum partial p-value across all available strata. With 18 possible violations for each country-year, we elevate our threshold for removal to the 95% level of confidence and drop any country-year with at least one p-value less than 0.05 from

[4] We are grateful to Martin Huber for graciously providing his original R code.

our analysis, resulting in the omission of 29 country-years for our robustness check, listed in dark font at the top of the y-axis.

Again, our conclusions are robust to the omission of these country-years. Figure 11 overlays the main results with the same results calculated after dropping the invalid country-years defined by sex-selectors and the simple and stratified IV tests. We indicate extreme errors with small black icons. The convergence is still striking with the IV test results outperform the main analysis in the earliest years.

8. Conclusion

In this paper we have investigated the degree to which LATEs from one context can be extrapolated to another. We set out to develop and demonstrate an extrapolation procedure built on the assumption that effect heterogeneity can be captured by observable covariates. We demonstrated the validity of our methodology using real world data on the relationship between having a third child and a mother's labor force participation. Despite heterogeneity in both the first-stage strength of our IV strategy in certain country-years, as well as variation the overall effect, our procedure performs well, particularly when we exploit the accumulated data.

Returning to our twofold motivation in pursuing this exercise – namely informing both the external validity of instrumental variables estimates and of the growing body of policy-relevant evidence from natural and randomized experiments – our findings are both optimistic and cautious. We find that external validity improves when the reference data and target data are from similar settings and given sufficient data. Even with a small set of covariates, average extrapolation error is close to zero when extrapolating LATEs from one country-year to another using the richest reference sample sizes. Furthermore, the resulting extrapolation error is usually less than the endogeneity bias of using within-target OLS.

At the same time, extrapolation error increases considerably with reference-target differences in both micro- and macro-level covariates. Covariate differences of 10 percent between reference and target settings lead to extrapolation error ranging from 5 to 20 percent of the overall treatment effect. While it is difficult to offer a specific quantitative guideline, our results suggest the importance of a close match between covariate profiles in reference and target settings. This echoes findings in the program evaluation literature such as Heckman, Ichimura, Smith, and Todd (1998) and our own related work on this theme (Dehejia, Pop-Eleches, and Samii 2015).

Given the increasing number of internally valid, albeit local estimates that are becoming available to assess the impact of policy interventions, our results suggest that there is some hope to reach externally valid, generalizable conclusions from this stream of evidence. However, we also demonstrate that the quality of extrapolation depends crucially on a sufficient body of quasi-experimental evidence from contexts that resemble the policy environment of interest. Finally, we note an important qualification: our results are narrowly relevant only to the application we have considered. Further replications of this exercise for other instrumental variables and natural and field experiments are necessary to develop a more systematic understanding of the opportunities for, and limits to, externally valid knowledge.

References

Abadie, Alberto. (2003). “Semiparametric Instrumental Variable Estimation of Treatment Response Models.” *Journal of Econometrics 113*: 231-263.

Allcott, Hunt (2014), “Site Selection Bias in Program Evaluation,” manuscript, New York University.

Angrist, Joshua, and William Evans. (1998). “Children and their Parents’ Labor Supply: Evidence from Exogenous Variation in Family Size.” *American Economic Review 99*(3): 450-477.

Angrist, Joshua, and Ivan Fernandez-Val. (2010). Extrapolating: External Validity and Overidentification in the LATE Framework. NBER Working Paper 16566.

Angrist, Joshua, Guido W. Imbens, and Donald B. Rubin. (1996). “Identification of Causal Effects Using Instrumental Variables.” *Journal of the American Statistical Association 91*: 444-472.

Butikofer, Aline. (2010). “Sibling Sex Composition and Cost of Children.” Manuscript.

Cruces, Guillermo, and Sebastian Galiani (2007), “Fertility and Female Labor Supply in Latin America: New Causal Evidence,” *Labour Economics 14*: 565-573.

Dehejia, Rajeev, and Sadek Wahba (1999). “Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs.” *Journal of the American Statistical Association 94*(448): 1053-1062.

Dehejia, Rajeev, and Sadek Wahba (2002). “Propensity Score Matching Methods for Non-Experiemental Causal Studies.” *Review of Economics and Statistics 84*: 151-161.

Dehejia, Rajeev, Cristian Pop-Eleches, and Cyrus Samii (2015). “From Local to Global: External Validity in a Fertility Natural Experiment.” National Bureau of Economic Research, Working Paper No. 21459.

Ebenstein, Avraham (2009), “When Is the Local Average Treatment Effect Close to the Average? Evidence from Fertility and Labor Supply,” *Journal of Human Resources, Volume 44*(4): 955-975.

Edlund, Lena, and Chulhee Lee (2013). “Son Preference, Sex Selection, and Economic Development: The Case of South Korea.” National Bureau of Economic Research, Working Paper No. 18679.

Gechter, Michael (2015), “Generalizing the Results from Social Experiments: Theory and Evidence from Mexico and India,” Manuscript.

Glass, Gene V. (1976), “Primary, Secondary, and Meta-Analysis of Research,” *Educational Researcher 5*(10): 3-8.

Froelich, Markus. (2007). “Nonparametric IV Estimation of Local Average Treatment Effects with Covariates.” *Journal of Econometrics 139*(1): 35-75.

Hartman, Erin, Richard Grieve, Roland Ramsahal, and Jasjeet Sekhon. (2015). “From Sample Average Treatment Effect to Population Average Treatment Effect on the Treated:

Combining Experimental with Observational Studies to Estimate Population Treatment Effects." *Journal of the Royal Statistical Society, Series A* (in press).

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. (1998). "Characterizing Selection Bias Using Experimental Data. *Econometrica 66*(5): 1017-1098.

Heckman, James, Robert LaLonde, and Jeffrey Smith. (1999). "The Economics and Econometrics of Active Labor Market Programs," in Orley Ashenfelter and David Card (eds.), *Handbook of Labor Economics*, Volume 3A. Amsterdam: North-Holland.

Heckman, James, Jeffrey Smith, and Nancy Clements. (1997). "Making the Most Out of Program Evaluations and Social Experiments: Accounting for Heterogeneity in Program Impacts." *Review of Economic Studies 64*(4): 487-535.

Hotz, V. Joseph, Susan Williams McElroy, and Seth G. Sanders. "Teenage childbearing and its life cycle consequences exploiting a natural experiment." *Journal of Human Resources* 40.3 (2005): 683-715.

Huber, Martin (2015). "Testing the Validity of the Sibling Sex Ratio Instrument." *Labour 29*(1): 1-14.

Humber, Martin, and G. Mellace (2015). "Testing Instrument Validity for LATE identification based on inequality moment constraints." *Review of Economics and Statistics 98*(2): 398-411.

Imbens, Guido W., and Joshua Angrist. (1994). "Identification and Estimation of Local Average Treatment Effects." *Econometrica 62*(2): 467-475.

Imbens, Guido W., and Jeffrey Wooldridge. (2009). "Recent Developments in the Econometrics of Program Evaluation." *Journal of Econometric Literature 47*(1):5-86.

Kitagawa, T. (2008). "A Bootstrap Test for Instrument Validity in the Heterogeneous Treatment Effect Model." Manuscript.

Lalonde, Robert. (1986). "Evaluating the Econometric Evaluation of Training Programs with Experimental Data." *American Economic Review 76*(4): 604-620.

Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 6.4 [dataset]. Minneapolis: University of Minnesota, 2015. http://doi.org/10.18128/D020.V6.4.

Miratrix, Luke, Jasjeet S. Sekhon, and Bin Yu. (2013). "Adjusting Treatment Effect Estimates by Post-Stratification in Randomized Experiments." *Journal of the Royal Statistical Society 75*(2): 369-396.

Pritchett, Lant, and Justin Sandefur. (2013). "Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix." Center for Global Development Working Paper No. 336.

Rosenzweig, Mark, and Kenneth Wolpin (2000). "Natural 'Natural Experiments' in Economics." *Journal of Economic Literature 38*(4): 827-874.

Rossi, Peter (1987). "The Iron Law of Evaluation and Other Metallic Rules." *Research in Social Problems and Public Policy* (4): 3-20.

Vivalt, Eva (2015), “How Much Can We Generalize from Impact Evaluation Results?”, manuscript, New York University.

Appendix A: Full Summary Statistics

Appendix B: Simulations

We use a simulation to demonstrate the properties of IV extrapolation under assumptions C1-C3. To construct an empirically plausible simulation, we start with the covariate, instrument (that is, the "same sex of first two kids" indicator), and treatment data (that is, the "more than two kids" indicator) from one percent subsamples of the IPUMS census data for Cuba in 2002 (yielding 215 observations) and the United States in 1990 (yielding 2,976 observations). The covariates that we use include the gender of the first and second born children, the woman's age coarsened into three-year bins, the woman's education level coarsened into four bins (less than primary, primary, secondary, and university completed), and her spouse's education level coarsened into the same four bins. We use these covariates to generate potential outcomes under treatment and control for women in each simulation replicate sample. We generate observed treatments and outcomes on the basis of the actual instrument and treatment values in the data. As such, we have realistic potential outcome distributions but we can also compute the actual LATE for each simulation replicate sample.

The simulations were run on a 1% random sample of the full IPUMS data in which we generated potential outcomes conditional on the covariate profile and observation "type". In this context, observation type refers to whether the unit was a complier, an always-taker, or a never-taker. We deterministically set half of the population to compliers and a quarter each to always- and never-takers. By construction, defiers are removed, thus ensuring we meet the conditional monotonicity assumption C1c discussed in Section 2.

Because potential outcomes are defined in terms of the micro-covariates only, C2 holds. We restrict attention to the portions of the Cuba and United States samples that overlap in their covariates such that C3 holds. We determine that C1b holds by observation: for the US in 1990, the first stage coefficient in the sample is 0.060 (robust s.e.=0.015), while for Cuba in 2002, the first stage coefficient in the sample is 0.090 (robust s.e.=0.048). We then assume C1a based on arguments of Angrist and Evans (1998). To generate the extrapolations, we use the complier-centered interactions 2SLS model defined above.

Combining the covariate profile with type yielded an extended covariate vector V_{ic} which has length k and, when stacked on observations, yields matrix $V_{n\times k}$. To generate the potential outcomes, we multiplied $V_{n\times k}$ by a k-length vector of coefficients B to yield the $Y(1)$ vector. B was drawn from a multivariate normal distribution with mean -0.1 and standard deviation 0.01, capturing the observed effect estimates in the real data. Without loss of generality, we set $Y(0) = 0$.

Having defined our potential outcomes as such – and in so doing, guaranteeing compliance with the necessary assumptions – we simulated the instrument assignment in a manner that varied with the covariate profile X_i. (Note that instrument assignment varies with X_i and not V_i since the latter would violate the conditional random assignment assumption of the instrument.) Specifically, we used a logit specification to determine the probability that the instrument $Z_{ic} = 1$ conditional on X_{ic} using the following specification:

$$\Pr\left[Z_{ic} = 1 | X_{ic}\right] = \frac{1}{1 + e^{-(X_{ic}A)}}, \tag{B.1}$$

where A is a k' length vector of coefficients, again drawn from a multivariate normal distribution for the sake of simplicity. Note that $k' = k - 3$ to account for the removal of unit type (again, compliers, always-, and never-takers) from the extended covariate vector V_{ic}. The simulation was run 1,000 times, using the centered-interactions technique specified above to extrapolate the target estimate from the reference population.

Figure B.1 displays results from the simulation exercise. In this context, Cuba was the target and the US was the reference sample. We see in Figure B.1 that the distribution of extrapolations is centered on the true LATE, even as we reduce the size of the reference population. The extrapolation distribution (depicted with the dotted lines) is also more precise than the IV estimates fit on the target population data (depicted with the light gray shading). This is because the reference sample (2,976 observations) is much larger than the target population sample (215 observations).

We also document evidence of "super" external validity in which extrapolated estimates are measured more precisely than the target estimate. This phenomenon is not merely the result of a larger reference sample size. In order for this result to obtain, it must be that the covariate strata of compliers in the reference population is large. Furthermore, the unit-level effects must have the largest variance in strata that are well-represented in the reference population.

In summary, the simulation illustrates our extrapolation strategy, and confirms that when the identifying assumptions are satisfied, the method indeed works (in the sense that extrapolated LATEs on average replicate the target LATE). Note, however, that the quality of the extrapolation depends on the sample size of the reference context and also on the degree of reference-target covariate overlap.

Appendix C: Dyadic Extrapolation Results

In the analysis above, we explored the comparison between our extrapolation technique and using an internally invalid OLS estimate. Our extrapolation technique used all potential reference populations for each target country-year and averaged the extrapolation estimates over them. Some dyadic pairings may be very inappropriate matches due to large differences in the individual and country-level covariates, making the above analysis a harder test for the comparison to OLS estimates. In the analysis that follows, we run a variety of different comparisons to (1) confirm the substantive interpretation provided above and (2) highlight the variety of approaches one can take in selecting the appropriate reference population.

Figure C.1 uses the same framework to compare the OLS estimates against the dyadic prediction error where the reference country-year is chosen according to different intuitive heuristics. The simplest heuristic, depicted in the top-right panel, is to simply use the largest available reference dataset. As illustrated, sample size along is not enough to generate accurate extrapolated estimates.

An alternative approach is to use the reference country-year that most effectively predicted the target in previous years. The errors associated with this technique are depicted in the top-right panel although, again, they are marginally inferior to the biased OLS estimate. A more effective heuristic is to use the reference country that is geographically closest to the target. However, although the average bias is smaller, the bottom-left panel exhibits greater variance associated with the extrapolated estimates than associated with the OLS.

Finally, in the bottom-right panel, we choose the reference that is closest to the target in covariate space. As discussed in Section 4, differences in the covariate profile significantly predict the errors between target and extrapolated estimates. By minimizing the Mahalanobis distance, we are effectively reducing the total impact of these differences in choosing the best dyadic pair.

While heuristics do suggest modest improvements in extrapolation accuracy, the best approach is to use the accumulated data instead of relying on a single dyadic comparison. By including both micro and macro level covariates in the centered interactions, cumulative extrapolation is better able to capture the necessary variation to ensure that the unconfounded location assumption holds.

Appendix D: Extrapolating reference ATE to target ATE

In the extrapolation exercise presented above, we are assuming that the characteristics of the complier population in the target setting are known and can be used to reweight the local average treatment effect in the reference country. There is a potential circularity here in the sense that knowledge of the target complier population implies the existence of micro data on the instrument and treatment variables in the target. Our argument in favor of the exercise is that the target complier population is simply one possible policy-relevant subpopulation in the target, specifically the only target subpopulation for which we have an internally valid estimate of the average treatment effect.

An alternative approach is to use the procedure discussed in Section 2 to extrapolate to the ATE in the target location from both the target LATE and the reference LATEs. Rather than reweight the reference LATEs based on the characteristics of target compliers, we reweight based on the characteristics of the full target sample, and likewise reweight the target LATE to the target ATE using the characteristics of the full target sample. As above, after extrapolating, we then measure "bias" as the extrapolated ATE from the reference data minus the extrapolated ATE from the target data. A snapshot of the results for this approach is presented in Figure D.1.

Figure D.1 depicts the average extrapolation error, where the set of reference countries evolves along the x-axis as they become available in years up to and including the year of observation of the target. The pattern is similar to Figure 5 in Section 5. Prior to 1980, the extrapolation tends to be noisy with values that exceed the theoretically sensible thresholds of 1 and -1. But as additional reference country-years become available, extrapolation error approaches zero in magnitude and is not significantly different from zero despite being reasonably stable.

An advantage of extrapolating ATE's is that these are directly comparable to OLS estimates within the target. In Figure D.2 we revisit our extrapolated reference IV to target OLS comparison for this case.

Figure D.2 compares extrapolated reference IV average treatment effects to OLS, selecting the reference based on three different heuristics as well as using the accumulated data. Again, we see superior performance associated with the cumulative approach although it bears noting that the OLS estimate performs quite well.

Appendix E: Using prior information to improve predictions

In many evaluation contexts prior information exists that can be used to improve the extrapolation. For example, one might begin with the prior of a zero treatment effect (perhaps motivated by Rossi's [1987] "Iron Law"). In the context of our application, prior information, if available, can readily be incorporated into the extrapolation by appropriately weighting the reweighted reference LATE with the prior. Here we present the simplest case of taking a convex combination of the reweighted LATE and the prior of a zero treatment effect.

Figure E.1 summarizes the weight on the prior that minimizes the root mean squared error of the cumulative extrapolation for each reference sample size (where the reference sample size increases with the increasing availability of reference country-years over time). The optimal weights range from 0.9 to 0.5. The fact that the optimal weights tend to put substantial weight on the prior reflects the fact that in this application many LATEs are in fact close to zero, i.e., that the prior of zero treatment effect is, ex post, a good one. The optimal weight also reflects the fact that the zero prior reduces posterior estimation variability. The importance of the latter diminishes as sample size increases, which is reflected in the downward trend of the optimal weight toward 0.5 for the full reference sample. In general, of course, the prior cannot be chosen with the benefit of hindsight, and absent extremely strong prior information weights in this range are unlikely.

At the same time, even a small weight put on a prior of a zero treatment effect tends to improve the root mean squared error of the extrapolation as depicted in Figure E.2. As weight on the prior increases, root mean squared error decreases essentially linearly. Again, while a very high weight on the prior is implausible, even a low weight on a prior of zero is beneficial. As illustrated in Table E.1, a weight of 0.01 on the prior reduces mean prediction error by 0.031 and root mean squared error by 0.14. These values are statistically significant after controlling for reference population size and the standard error of the extrapolated estimate. Prior information, if it is available and proves to be correct, is a valuable input to improving external predictions.

Table 1: Main variables summarized by observations and dyadic absolute differences.

Variable Name	*Country-Year Level*	*Dyadic Level*
	Raw statistics	Abs. Differences
Average Education (mother)	1.88	0.64
N = 134 / 11,278,326	(0.85) [0.53]	(0.47)
Average Education (spouse)	2.02	0.59
N = 134 / 9,715,191	(0.98) [0.49]	(0.44)
Age (mother @ survey)	30	0.93
N = 141 / 11,749,847	(3.6) [0.63]	(0.69)
Age (mother @ 1st birth)	20.65	1.1
N = 141 / 11,749,847	(3.11) [0.8]	(0.86)
Ratio (male::female)	0.01	0.01
N = 141 / 11,749,847	(0.3) [0.01]	(0.01)
GDP per capita	8927.8	10434.9
N = 141 / 11,749,847	(10182.84) [-]	(9400.21)
Total Fertility Rate (children per mother)	2.77	0.74
N = 139 / 11,616,325	(0.59) [-]	(0.54)
Labor Force Participation Rate	0.54	0.24
N = 125 / 10,140,694	(0.22) [-]	(0.17)
Year	1990.7	13.19
N = 141 / 11,749,847	(10.33) [-]	(10.21)
Geographic Distance	-	7910.03
N = - / 15,034	(-) [-]	(4638.53)
2SLS Variables		
Economically active Mother (Y)	0.46	0.27
N = 127 / 10,259,072	(0.5) [0.25]	(0.2)
More Kids (D)	0.6	0.22
N = 141 / 11,749,847	(0.49) [0.15]	(0.16)
Number of Children (D)	3.11	0.6
N = 141 / 11,749,847	(1.26) [0.45]	(0.43)
Two children of same sex (Z)	0.51	0.01
N = 141 / 11,749,847	(0.5) [0.01]	(0.01)
Two boys (Z)	0.26	0.01
N = 141 / 11,749,847	(0.44) [0.01]	(0.01)
Two girls (Z)	0.24	0.01
N = 141 / 11,749,847	(0.43) [0.01]	(0.01)

Notes: Standard deviations calculated on country year means presented in parentheses. Average household standard deviations presented in brackets. The three 2SLS variables are dummies. More kids is coded zero if the mother has only 2 children and one if the mother has more than 2 children. Same sex is coded zero if the first two children are of different genders and coded one if the first two children are of the same gender. Economically active mother is coded zero if the mother is not economically active and coded one if the mother works for pay. Total observations recorded as total country-years / total census-level observations.

Table 2: Multivariate regression of absolute extrapolation error on absolute covariate differences in dyadic data.

	(1) Full Sample	(2) FS t-stat > 2	(3) FS t-stat > 5	(4) FS t-stat > 10
Mother's Education (log)	.092	.092	.092	.102
	(.058)	(.058)	(.058)	(.063)
Spouse's Education (log)	.107*	.106*	.106*	.107*
	(.060)	(.060)	(.060)	(.063)
Mother's Age @ Survey (log)	.032	.031	.031	.019
	(.038)	(.038)	(.038)	(.041)
Mother's Age @ First Birth (log)	.000	.001	.002	.008
	(.031)	(.031)	(.031)	(.033)
GDP pc (log)	.009	.009	.009	.025
	(.029)	(.029)	(.029)	(.031)
Gender Ratio (log)	.052	.050	.048	.040
	(.037)	(.037)	(.037)	(.039)
Labor Force Part. Rate (log)	-.032	-.032	-.031	-.034
	(.036)	(.036)	(.036)	(.039)
Total Fert. Rate (log)	.079*	.080*	.080*	.077*
	(.041)	(.041)	(.041)	(.044)
Temporal Dist. (1SD = 10yrs)	-.005	-.005	-.004	-.006
	(.037)	(.037)	(.037)	(.040)
Geographic Dist. (1SD = 4,650km)	-.030	-.029	-.029	-.029
	(.038)	(.038)	(.038)	(.041)
Constant	-2.172***	-2.180***	-2.201***	-2.423***
	(.345)	(.347)	(.348)	(.367)
N	13362	10620	7730	3920
R^2	.047	.046	.046	.048

Notes: Heteroskedastic-robust standard errors presented in parentheses. Explanatory variables are measured by the log of the absolute difference between the target value and the reference. Gender ratio calculated as the ratio of boys to girls. Mother's education level coded as 1 = less than primary completed, 2 = primary completed, 3 = secondary completed, 4 = university completed. Temporal and geographic distances presented in standardized units. * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 3: List of country-years with partial P-values less than 0.40.

Country	*Year*	*Partial P-value*	*St. Diff*$_0$	*St. Diff*$_1$
Uganda	2002	0.094	0.017	0.002
India	1987	0.126	-0.052	0.016
Portugal	2001	0.157	0.024	-0.205
Panama	1960	0.163	-0.157	0.046
Israel	1995	0.212	0.004	0.021
Malaysia	1980	0.221	-0.059	0.025
Malaysia	1991	0.257	-0.057	0.014
Chile	1970	0.284	0.018	-0.097
India	1993	0.316	-0.014	0.009
Mali	1998	0.343	-0.005	0.008
Guinea	1996	0.36	-0.014	0.007
Guinea	1983	0.389	0.014	0.01
Costa Rica	1973	0.391	-0.017	0.01
India	1983	0.392	0.002	0.011

Notes: Invalid country-years ranked by partial P-values from Huber-Mellace (2014) test of IV validity, column 3. The p-values test whether $f(y, D = 1|Z = 1) \geq f(y, D = 1|Z = 0)$ and, similarly, $f(y, D = 0|Z = 0) \geq f(y, D = 0|Z = 1)$. These constraints can be rewritten as four point estimates $(\hat{\theta}_1, \ldots, \hat{\theta}_4)$ which must fall between the bounds of the mixed population. The fourth and fifth columns give the standardized point estimates in the form of $\frac{\max(\hat{\theta}_1, \hat{\theta}_2)}{SD(Y)}$ for the treated (*St. Diff*$_1$) and non-treated (*St. Diff*$_0$) subpopulations. Violations of the null are therefore positive values. Inference is applied to the test statistics using two-stage bootstrapping, the details of which can be found in Huber and Mellace (2015).

Table A-1: All country year statistics

Country	*Year*	*GDPpc*	*LFP*	*TFR*	*Sex Ratio*	*Educ.*	*Age*	*FS* β	*IV* β
Argentina	1970	7615	0.31	2.48	.508 (.31)	1.57 (.58)	30.2 (3.57)	.049 (.008)	-.064 (.124)
	1980	8487	0.28	2.67	.506 (.30)	1.66 (.62)	29.6 (3.76)	.044 (.003)	-.060 (.054)
	1991	7423	0.51	2.59	.506 (.30)	1.98 (.60)	30.1 (3.64)	.035 (.002)	-.157 (.069)
	2001	8552	0.58	2.52	.508 (.31)	2.2 (.66)	29.9 (3.7)	.026 (.003)	-.169 (.109)
Armenia	2001	2837	0.76	2.24	.528 (.31)	3.06 (.53)	30.2 (3.55)	.124 (.008)	-.157 (.057)
Austria	1971	16527	0.6	2.08	.511 (.32)	30.07 (3.35)	0 (0.01)	.045 (.184)	- (-)
	1981	22437	0.65	2.22	.513 (.33)	30.53 (3.25)	0.1 (0.01)	-.292 (.132)	- (-)
	1991	27956	0.67	2.07	.511 (.33)	30.89 (3.07)	0 (0.01)	-.311 (.156)	- (-)
	2001	33839	0.79	2.07	.515 (.33)	31.64 (2.85)	0.1 (0.01)	-.200 (.128)	- (-)
Belarus	1999	5678	0.82	1.76	.515 (.35)	2.95 (.55)	31.3 (3.08)	.027 (.004)	-.266 (.144)
Bolivia	1976	3255	0.22	3.06	.508 (.28)	1.26 (.52)	29.6 (3.67)	.014 (.006)	.113 (.400)
	1992	2755	0.52	3.14	.509 (.29)	1.71 (.74)	29.6 (3.68)	.018 (.005)	.359 (.361)
	2001	3134	0.49	2.84	.514 (.30)	1.83 (.77)	29.7 (3.71)	.014 (.005)	-.025 (.393)
Brazil	1960	2469	0.14	3.83	.510 (.27)	1.06 (.30)	29.3 (3.77)	.016 (.002)	-.021 (.095)
	1970	3845	0.19	3.71	.509 (.27)	1.11 (.41)	29.4 (3.77)	.021 (.002)	-.073 (.065)
	1980	6943	0.3	3.31	.507 (.29)	1.27 (.63)	29.5 (3.72)	.028 (.002)	-.019 (.060)
	1991	6117	0.43	2.85	.510 (.30)	1.49 (.78)	29.9 (3.64)	.040 (.002)	-.063 (.038)
	2000	6834	0.58	2.37	.510 (.31)	1.59 (.78)	29.9 (3.74)	.035 (.002)	-.057 (.044)
Cambodia	1998	888	0.83	3.57	.509 (.29)	1.2 (.43)	30.2 (3.44)	.024 (.004)	.087 (.131)
Chile	1970	4465	0.25	3.42	.504 (.28)	1.6 (.65)	29.5 (3.75)	.027 (.005)	.209 (.135)
	1982	4308	0.27	2.67	.508 (.31)	1.87 (.70)	29.9 (3.62)	.032 (.005)	.017 (.115)
	1992	6527	0.28	2.27	.511 (.32)	2.13 (.68)	30.3 (3.42)	.040 (.004)	.008 (.085)
	2002	9664	0.43	2.09	.509 (.33)	2.31 (.67)	30.9 (3.28)	.028 (.005)	-.216 (.154)
China	1982	624	0.88	2.9	.525 (.30)	1.56 (.54)	30.9 (2.82)	.077 (.001)	-.055 (.013)
	1990	1157	0.9	2.26	.529 (.30)	1.66 (.59)	30.7 (3.4)	.149 (.001)	-.017 (.006)
Colombia	1973	4089	0.31	3.77	.507 (.27)	1.35 (.55)	29.2 (3.88)	.021 (.003)	.134 (.118)
	1985	4962	0.45	2.91	.508 (.30)	1.71 (.73)	29.4 (3.78)	.036 (.003)	-.025 (.076)
	1993	5785	0.41	2.56	.509 (.31)	1.83 (.69)	29.8 (3.67)	.036 (.003)	.059 (.068)
	2005	6491	0.32	2.31	.513 (.31)	1.91 (.79)	29.7 (3.76)	.034 (.003)	.035 (.067)
Costa Rica	1973	7067	0.22	4.12	.509 (.27)	1.45 (.67)	29.1 (3.87)	.006 (.009)	.622 (1.649)
	1984	7075	0.27	2.98	.510 (.30)	1.8 (.68)	29.3 (3.76)	.049 (.009)	.001 (.141)
	2000	8870	0.38	2.44	.509 (.31)	2.03 (.72)	30.1 (3.64)	.039 (.008)	.031 (.170)
	2002	7624	0.47	1.44	.516 (.34)	2.47 (.63)	31.2 (3.07)	.039 (.005)	-.052 (.146)
Ecuador	1974	4067	0.17	3.74	.509 (.27)	1.31 (.54)	29.2 (3.82)	.011 (.005)	.740 (.506)
	1982	5074	0.25	3.39	.510 (.28)	1.6 (.69)	29.2 (3.8)	.017 (.005)	.050 (.224)
	1990	4429	0.33	3.03	.511 (.29)	1.88 (.78)	29.5 (3.7)	.034 (.005)	.181 (.129)
	2001	4824	0.36	2.49	.510 (.31)	1.98 (.76)	29.6 (3.77)	.029 (.005)	-.053 (.144)
Egypt	1996	3233	0.21	3.16	.525 (.29)	30.05 (3.73)	0 (0)	.014 (.030)	- (-)
France	1962	11116	0.47	2.21	.510 (.30)	1.53 (.63)	30.8 (3.18)	.031 (.004)	-.174 (.094)
	1968	14312	0.55	2.24	.512 (.30)	1.69 (.68)	30.7 (3.34)	.039 (.004)	-.132 (.078)
	1975	18472	0.71	2.13	.513 (.32)	1.88 (.79)	30.2 (3.4)	.033 (.003)	-.066 (.100)
	1982	21910	0.79	1.93	.514 (.32)	1.94 (.84)	31 (3.15)	.049 (.003)	-.267 (.063)
	1990	25766	0.85	1.88	.512 (.32)	2.08 (.90)	31.3 (2.93)	.048 (.004)	-.155 (.072)
	1999	28716	0.89	1.87	.512 (.33)	2.26 (.98)	31.9 (2.76)	.046 (.004)	-.144 (.075)
Ghana	2000	1478	0.85	2.67	.503 (.31)	1.4 (.60)	30.1 (3.52)	-.006 (.004)	.639 (.604)
Greece	1971	13129	0.25	1.95	.517 (.33)	1.74 (.66)	31 (3.3)	.074 (.006)	-.143 (.074)
	1981	16555	0.33	1.86	.517 (.33)	2.18 (.59)	30.4 (3.51)	.075 (.005)	-.008 (.057)
	1991	17768	0.44	2	.522 (.34)	2.39 (.64)	30.8 (3.28)	.064 (.005)	-.082 (.082)
	2001	21887	0.5	2.01	.517 (.34)	2.64 (.72)	31.6 (2.93)	.051 (.006)	.022 (.131)
Guinea	1983	740	0.52	2.21	.520 (.33)	1.09 (.39)	29.2 (3.79)	.005 (.007)	-2.13 (3.46)
	1996	727	0.73	2.72	.522 (.29)	1.07 (.30)	29.1 (3.78)	.008 (.005)	.461 (.620)

Notes: Standard deviations presented in parentheses. Standard deviations unavailable for per capita GDP, total fertility rate, and labor force participation rate due to measurement at higher levels of aggregation than the household.

Table A-1: All country year statistics (cont'd)

Country	*Year*	*GDPpc*	*LFP*	*TFR*	*Sex Ratio*	*Educ.*	*Age*	*FS β*	*IV β*
Hungary	1970	7779	1.88	0.51	.333 (1.97)	0.54 (30.62)	3.3 (0.04)	.007 (-)	- (-)
	1980	11255	1.78	0.51	.337 (2.41)	0.6 (29.95)	3.4 (0.05)	.006 (-)	- (-)
	1990	12489	0.75	1.76	.504 (.34)	2.58 (.58)	31.1 (3.22)	.051 (.006)	-.413 (.118)
	2001	13732	0.62	1.88	.513 (.33)	2.72 (.69)	31 (3.14)	.039 (.008)	-.447 (.190)
India	1983	1107	0.32	2.78	.533 (.29)	1.35 (.64)	29.4 (3.75)	.010 (.005)	-.340 (.503)
	1987	1261	0.3	2.75	.537 (.29)	1.42 (.71)	29.4 (3.74)	.016 (.005)	-.844 (.373)
	1993	1434	0.33	2.63	.544 (.30)	1.51 (.77)	29.6 (3.65)	.025 (.005)	-.205 (.214)
	1999	1910	0.3	2.64	.544 (.30)	1.59 (.82)	29.9 (3.59)	.036 (.005)	-.008 (.141)
Iraq	1997	2755	0.13	4.63	.512 (.26)	1.68 (.74)	29.3 (3.73)	.010 (.002)	-.044 (.181)
Israel	1972	13991	0.32	3.07	.518 (.31)	2.44 (.78)	28.5 (3.55)	.025 (.008)	-.119 (.276)
	1983	16123	2.88	0.51	.316 (2.30)	0.79 (29.20)	3.4 (0.02)	.007 (-)	- (-)
	1995	20790	0.59	2.78	.520 (.31)	2.5 (.83)	29.3 (3.34)	.006 (.007)	2.06 (2.89)
Italy	2001	29146	0.66	1.72	.514 (.34)	2.4 (.60)	30.6 (2.58)	.027 (.003)	.047 (.174)
Jordan	2004	3947	0.27	4.12	.522 (.27)	2.52 (.73)	30.1 (3.48)	.015 (.005)	.192 (.324)
Kenya	1989	1172	0.76	3.55	.507 (.28)	1.5 (.53)	28.7 (3.7)	-.004 (.003)	.768 (1.183)
	1999	1133	0.51	0.29	1.665 (.52)	29.12 (3.70)	- (-)	- (-)	- (-)
Kyrgyz Republic	1999	1597	0.78	2.89	.511 (.30)	2.99 (.46)	30.1 (3.49)	.068 (.006)	-.132 (.075)
Malaysia	1970	2065	0.43	3.55	.511 (.26)	1.22 (.43)	29.6 (3.73)	.012 (.008)	.162 (.878)
	1980	4250	0.53	3.28	.509 (.28)	1.57 (.54)	29.7 (3.51)	.026 (.009)	-.575 (.449)
	1991	6272	0.47	2.91	.515 (.29)	1.84 (.58)	30.4 (3.41)	.020 (.007)	-.558 (.389)
	2000	9474	0.5	2.85	.514 (.29)	1.9 (.73)	30.8 (3.27)	.035 (.007)	-.400 (.215)
Mali	1987	628	0.5	3.03	.514 (.29)	1.1 (.34)	29 (3.81)	-.001 (.005)	-3.67 (24.3)
	1998	768	0.39	3.33	.515 (.28)	1.06 (.25)	29.2 (3.81)	.005 (.004)	1.45 (1.42)
Mexico	1970	6848	0.17	3.94	.512 (.27)	1.24 (.48)	28.9 (3.88)	.009 (.004)	.188 (.451)
	1990	9427	0.27	3.18	.508 (.29)	1.72 (.71)	29.6 (3.68)	.030 (.001)	-.079 (.039)
	1995	9158	0.45	2.69	.505 (.30)	1.84 (.67)	29.5 (3.72)	.029 (.007)	-.079 (.259)
	2000	11380	0.35	2.74	.510 (.30)	1.88 (.70)	29.7 (3.66)	.033 (.001)	-.031 (.039)
Mongolia	1989	2740	3.38	0.51	.283 (2.22)	0.82 (29.82)	3.2 (0.01)	.008 (-)	- (-)
	2000	2219	0.78	2.59	.506 (.31)	2.59 (.70)	30.6 (3.26)	.052 (.009)	.066 (.136)
Nepal	2001	918	0.64	2.87	.532 (.29)	1.32 (.66)	29.5 (3.75)	.009 (.003)	-.583 (.335)
Pakistan	1973	991	0.54	0.28	1.133 (.41)	29.71 (3.85)	- (-)	- (-)	- (-)
	1998	1732	3.66	0.52	.280 (1.25)	0.55 (29.22)	3.8 (0)	.001 (-)	- (-)
Panama	1960	2142	0.38	2.98	.504 (.27)	1.42 (.59)	28.8 (3.85)	.016 (.016)	1.48 (1.93)
	1970	3419	0.34	3.39	.513 (.28)	1.56 (.65)	29 (3.84)	.011 (.009)	.056 (.942)
	1980	5200	0.41	3.2	.513 (.29)	1.82 (.73)	29.4 (3.75)	.009 (.009)	-.700 (1.25)
	1990	5531	0.37	2.59	.516 (.30)	2.1 (.77)	29.5 (3.74)	.041 (.009)	-.007 (.217)
	2000	6950	0.45	2.42	.513 (.30)	2.16 (.74)	29.8 (3.73)	.032 (.009)	.439 (.324)
Peru	1993	3855	0.33	2.93	.505 (.30)	1.84 (.89)	29.8 (3.66)	.027 (.003)	-.017 (.104)
	2007	6374	0.41	2.39	.509 (.32)	2.16 (.89)	30.1 (3.57)	.030 (.003)	.019 (.104)
Philippines	1990	2334	0.49	3.43	.513 (.28)	2.2 (.85)	29.8 (3.59)	.028 (.002)	-.133 (.062)
	1995	2365	3.12	0.51	.283 (2.36)	0.87 (29.94)	3.6 (0.03)	.002 (-)	- (-)
	2000	2464	2.99	0.52	.296 (2.33)	0.8 (30.09)	3.5 (0.03)	.002 (-)	- (-)
Portugal	1981	11369	0.64	2.09	.512 (.32)	1.27 (.64)	30.7 (3.34)	.052 (.008)	.042 (.159)
	1991	15661	0.76	1.8	.509 (.34)	1.53 (.77)	31.3 (3.04)	.024 (.008)	.138 (.361)
	2001	20095	0.86	1.63	.512 (.34)	1.88 (.85)	31.8 (2.94)	.032 (.009)	-.670 (.332)
Puerto Rico	1970	10418	3.05	0.5	.296 (2.09)	0.83 (29.29)	3.8 (0.02)	.026 (-)	- (-)
	1980	12556	2.76	0.51	.303 (2.45)	0.78 (29.96)	3.6 (0.05)	.011 (-)	- (-)
	1990	17870	0.5	2.26	.512 (.31)	2.73 (.69)	30.2 (3.58)	.057 (.012)	.169 (.208)
	2000	25284	0.56	1.89	.514 (.32)	2.89 (.62)	30.1 (3.69)	.056 (.012)	-.354 (.219)
	2005	26054	0.68	1.77	.514 (.32)	2.93 (.62)	30.4 (3.62)	.061 (.033)	-.191 (.545)

Notes: Standard deviations presented in parentheses. Standard deviations unavailable for per capita GDP, total fertility rate, and labor force participation rate due to measurement at higher levels of aggregation than the household.

Table A-1: All country year statistics (cont'd)

Country	*Year*	*GDPpc*	*LFP*	*TFR*	*Sex Ratio*	*Educ.*	*Age*	*FS β*	*IV β*
Romania	1977	5622	2.19	0.51	.318 (1.76)	0.79 (29.78)	3.6 (0.04)	.004 (-)	- (-)
	1992	5005	0.77	2.12	.508 (.32)	2.47 (.66)	30.4 (3.61)	.038 (.003)	-.039 (.078)
	2002	6575	0.56	1.9	.508 (.33)	2.43 (.67)	30.8 (3.29)	.039 (.004)	.041 (.107)
Rwanda	1991	768	0.97	3.98	.496 (.28)	30.27 (3.28)	0 (0)	-.590 (1.16)	- (-)
	2002	732	0.92	3.39	.496 (.29)	1.31 (.51)	30 (3.48)	.005 (.005)	-1.18 (1.22)
Saint Lucia	1980	5432	0.53	3.9	.504 (.31)	1.2 (.61)	28.9 (4.02)	.016 (.041)	.190 (4.73)
	1991	9052	0.54	3.05	.489 (.30)	1.76 (.93)	29.3 (3.62)	.002 (.038)	1.79 (31.6)
Senegal	1988	1251	0.24	3.2	.497 (.28)	1.14 (.38)	28.8 (3.79)	.003 (.004)	-.322 (1.42)
	2002	1276	0.33	3.25	.504 (.29)	1.16 (.42)	29.4 (3.76)	-.001 (.005)	2.88 (17.7)
Slovenia	2002	20432	0.92	1.94	.520 (.35)	2.89 (.51)	30.6 (2.42)	.008 (.010)	-.696 (1.24)
South Africa	1996	5477	0.72	2.6	.500 (.31)	2 (.70)	30.4 (3.42)	.025 (.003)	-.007 (.118)
	2001	5996	0.78	2.44	.497 (.31)	2.08 (.75)	30.6 (3.32)	.020 (.003)	.129 (.142)
	2007	7442	0.84	2.29	.498 (.32)	2.23 (.71)	30.5 (3.46)	.024 (.007)	.088 (.212)
Spain	1991	20715	0.49	2.3	.514 (.33)	2.12 (.71)	31.6 (2.87)	.057 (.004)	-.027 (.080)
	2001	26714	0.64	1.73	.516 (.34)	2.29 (.63)	32.3 (2.68)	.045 (.005)	-.067 (.140)
Switzerland	1970	29439	0.41	2.22	.513 (.32)	3 (.14)	30.9 (3.14)	.018 (.011)	-.395 (.542)
	1980	30010	0.44	1.92	.516 (.33)	3.01 (.19)	31.5 (2.89)	.055 (.010)	-.368 (.191)
	1990	34296	0.57	1.86	.514 (.33)	2.99 (.30)	31.6 (2.76)	.057 (.011)	-.143 (.193)
	2000	35788	0.64	2.19	.506 (.33)	2.95 (.44)	32.3 (2.51)	.051 (.012)	-.009 (.246)
Tanzania	1988	687	0.88	3.13	.500 (.29)	1.39 (.53)	29.1 (3.66)	-.013 (.003)	-.113 (.160)
	2002	790	0.76	2.91	.502 (.30)	1.75 (.53)	29.3 (3.61)	.003 (.002)	-.626 (.869)
Thailand	1970	1570	3.82	0.51	.274 (1.14)	0.44 (29.94)	3.6 (0.02)	.004 (-)	- (-)
	1980	2413	3.28	0.51	.292 (1.09)	0.37 (29.56)	3.6 (0.04)	.007 (-)	- (-)
	1990	4379	2.33	0.51	.319 (1.34)	0.71 (30.42)	3.4 (0.08)	.007 (-)	- (-)
	2000	5651	1.97	0.51	.338 (1.77)	0.68 (31.01)	3.1 (0.06)	.006 (-)	- (-)
Uganda	1991	582	0.68	3.09	.500 (.28)	1.29 (.47)	28.7 (3.68)	-.007 (.003)	.414 (.546)
	2002	884	0.59	3.4	.500 (.28)	1.39 (.55)	28.6 (3.76)	.000 (.002)	29.9 (2582)
United Kingdom	1991	22766	0.58	1.88	.511 (.32)	30.91 (3.22)	0.1 (0.01)	-.267 (.081)	- (-)
United States	1960	15388	0.44	2.49	.506 (.30)	2.54 (.64)	30 (3.62)	.040 (.003)	-.158 (.076)
	1970	20436	0.5	2.53	.509 (.30)	2.68 (.61)	29.9 (3.54)	.038 (.003)	-.173 (.089)
	1980	24985	0.63	2.17	.511 (.32)	2.81 (.60)	30.4 (3.42)	.060 (.002)	-.142 (.026)
	1990	31452	0.68	2.06	.511 (.32)	2.9 (.55)	30.7 (3.33)	.062 (.002)	-.131 (.024)
	2000	39644	0.73	1.86	.511 (.32)	2.93 (.58)	30.7 (3.45)	.057 (.002)	-.105 (.027)
	2005	42482	0.74	1.85	.511 (.32)	3.02 (.60)	31 (3.38)	.058 (.004)	-.011 (.066)
Venezuela	1971	9369	0.25	3.6	.507 (.27)	1.38 (.52)	28.9 (3.91)	.017 (.004)	.299 (.214)
	1981	9643	0.44	3.05	.511 (.29)	1.65 (.61)	29.1 (3.79)	.028 (.003)	.076 (.125)
	1990	8125	0.36	2.99	.510 (.29)	1.68 (.62)	29.4 (3.79)	.029 (.003)	-.137 (.108)
	2001	8681	0.41	2.4	.509 (.31)	1.93 (.54)	29.7 (3.75)	.074 (.003)	-.059 (.042)
Vietnam	1989	855	0.87	3.23	.514 (.29)	1.68 (.69)	30.3 (3.26)	.038 (.003)	-.075 (.052)
	1999	1466	0.83	2.5	.516 (.32)	1.83 (.68)	30.7 (3.24)	.077 (.003)	-.049 (.029)

Notes: Standard deviations presented in parentheses. Standard deviations unavailable for per capita GDP, total fertility rate, and labor force participation rate due to measurement at higher levels of aggregation than the household.

Table E.1: Shrinkage weights on measures of extrapolation error in dyadic data.

	Dependent variable:	
	Root Mean Squared Error (RMSE)	Mean Prediction Error
	(1)	(2)
Weight on Prior	-3.040^{***}	-1.061^{***}
	(0.137)	(0.083)
Target Size	-0.691^{***}	-0.446^{***}
	(0.137)	(0.083)
Target SE	5.025^{***}	1.622^{***}
	(0.137)	(0.083)
Constant	5.904^{***}	2.675^{***}
	(0.137)	(0.083)
Observations	12,625	12,625
R^2	0.130	0.044

Notes: Dependent variables given in column headers. RMSE calculated as $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\beta_{tar}-\beta_{ext_i})^2}$ for each target country-year. Mean prediction error (MPE) calculated as average of absolute difference between target estimate and each reference extrapolated estimate. Reference sample size and standard error of extrapolated estimate standardized to facilitate coefficient comparisons.

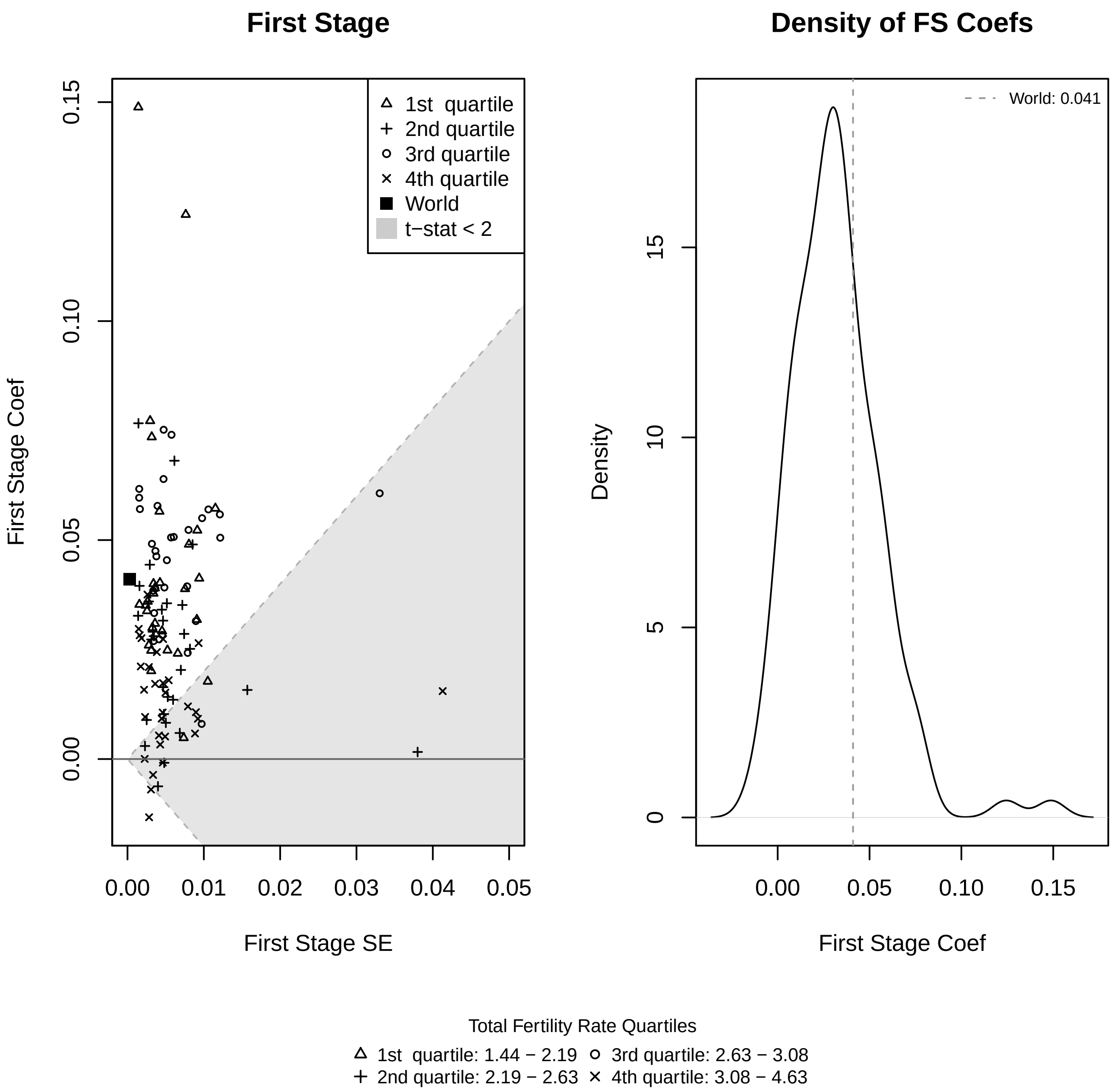

Figure 1: Summary of first-stage results of regressing *morekids* on *samesex* in full data. The left-panel is a scatter of all available country-years by region with the world coefficient indicated by a light gray circle. The gray cone covers the area in which the first-stage t-statistic is less than 2. The right-panel plots the density of the first-stage coefficients. Points are coded to reflect where each country-year falls in the overall distribution of fertility rates.

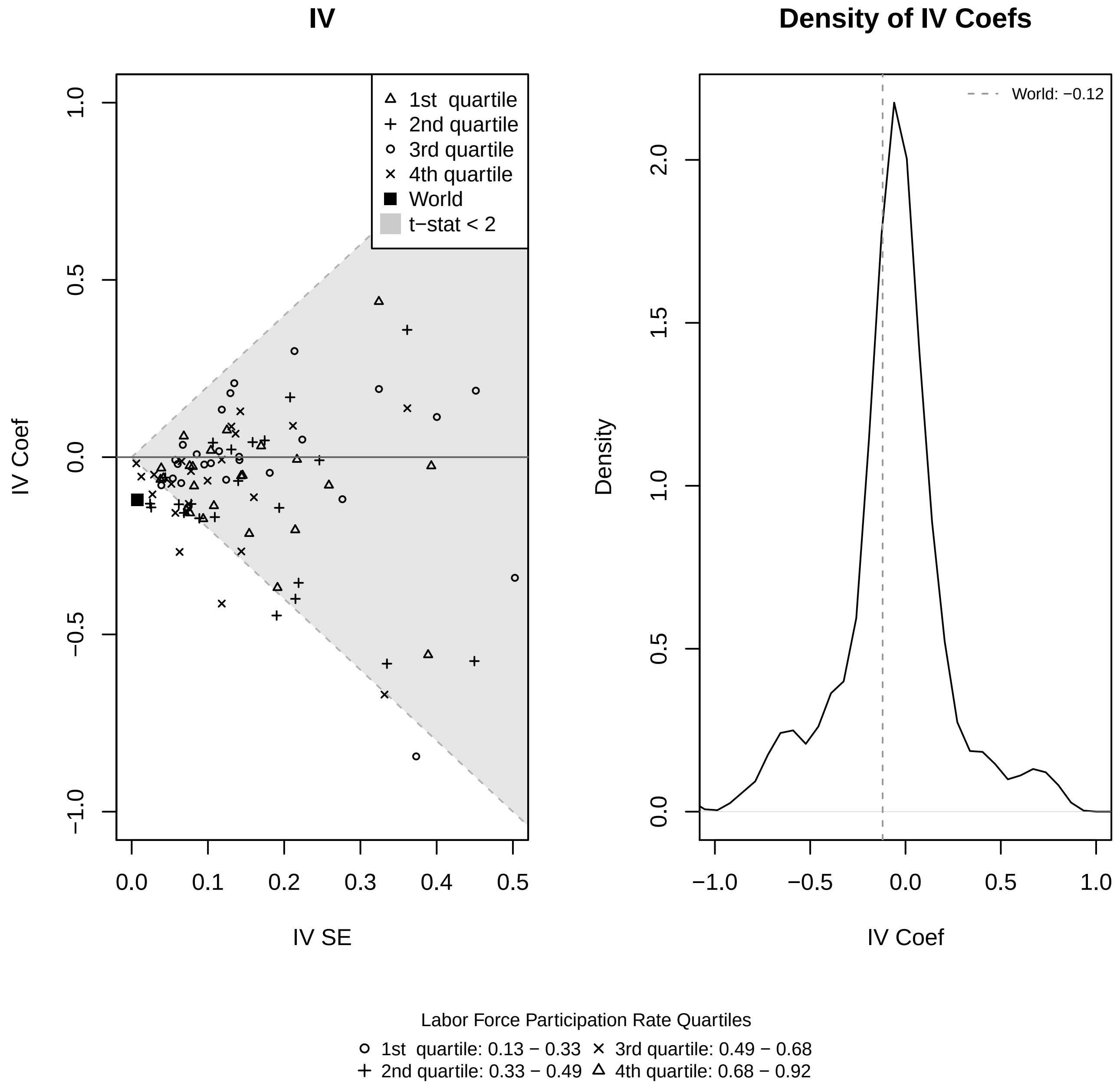

Figure 2: Summary of 2SLS results of regressing *econactivem* on *morekids* in full data. The left-panel is a scatter of all available country-years by region with the world coefficient indicated by a light gray circle. The gray cone covers the area in which the 2SLS t-statistic is less than 2. The right-panel plots the density of the 2SLS coefficients. Points are coded to reflect where each country-year falls in the overall distribution of female labor force participation.

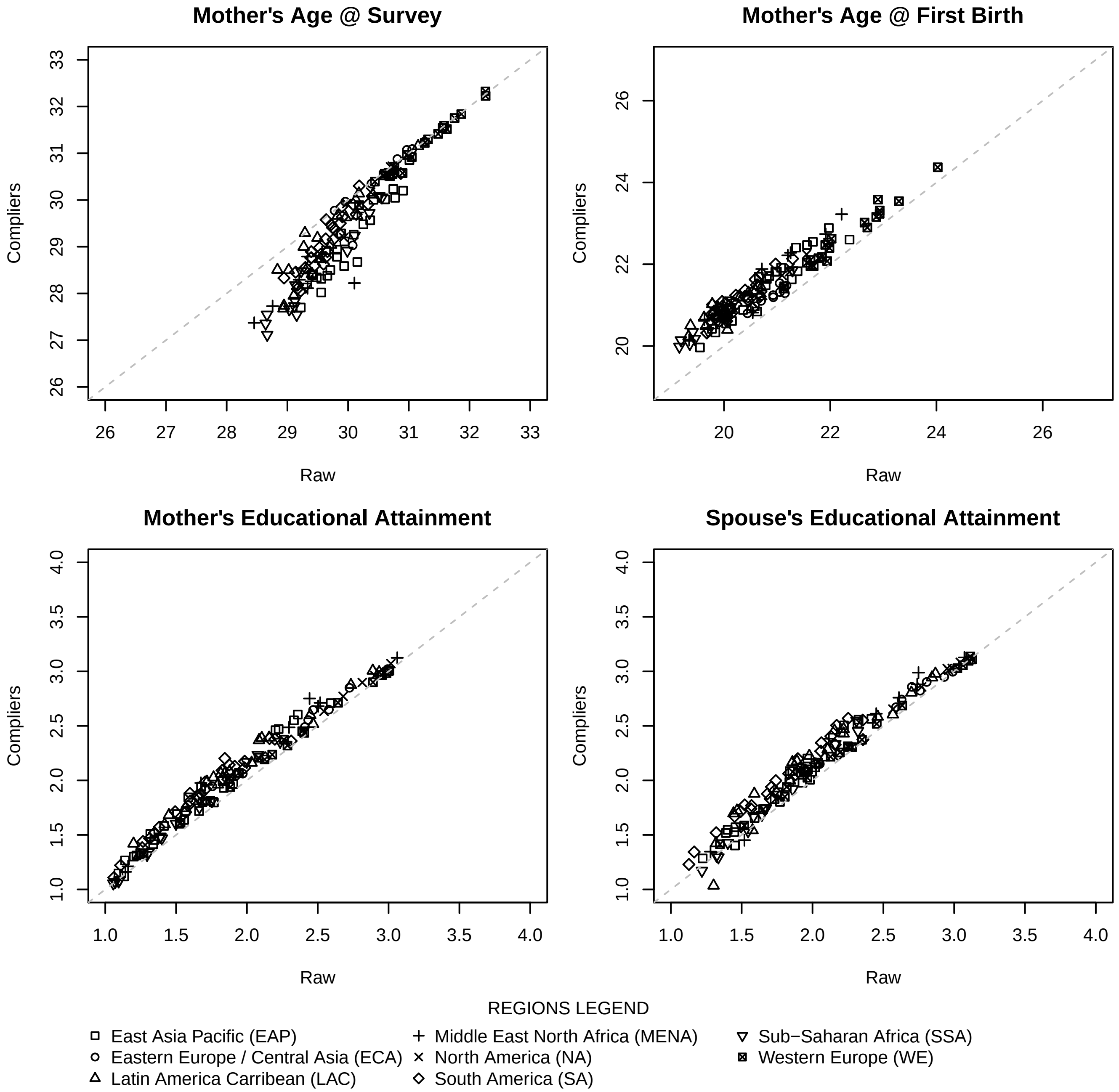

Figure 3: Scatter of complier means (y-axis) versus raw means (x-axis) of mother's age at survey (top-left), mother's age at first birth (top-right), mother's educational attainment (bottom-left), and spouse's educational attainment (bottom-right), by region. Points falling above (below) the 45° line indicate compliers who are older (younger) or more (less) educated than the full population.

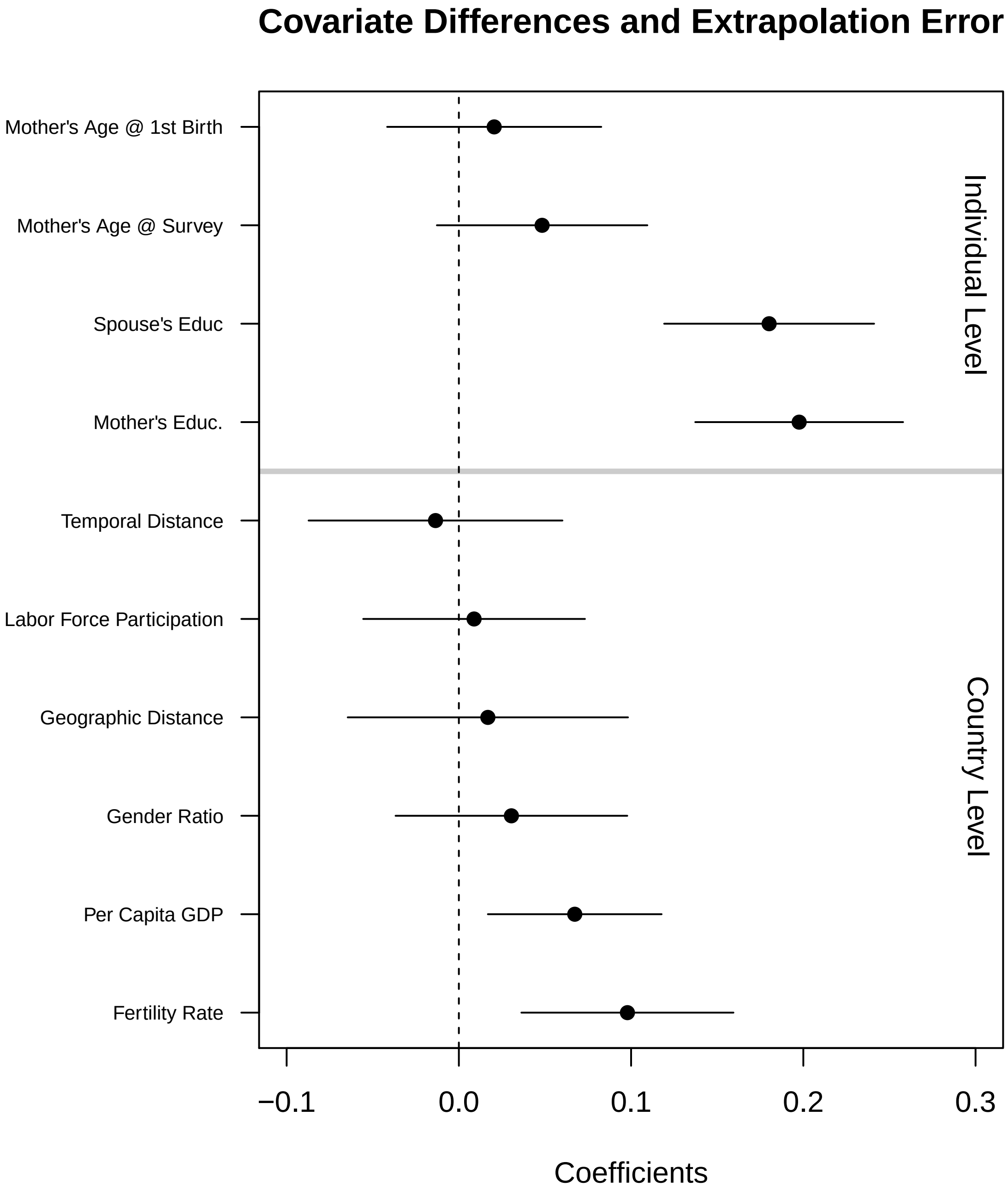

Figure 4: Results from a battery of bivariate regressions of absolute extrapolation error on the absolute covariate differences between reference and target contexts. Covariates include individual-level (education and age) as well as country-level (gender ratio, fertility rate, per capita GDP, labor force participation, and geographic and temporal distances).

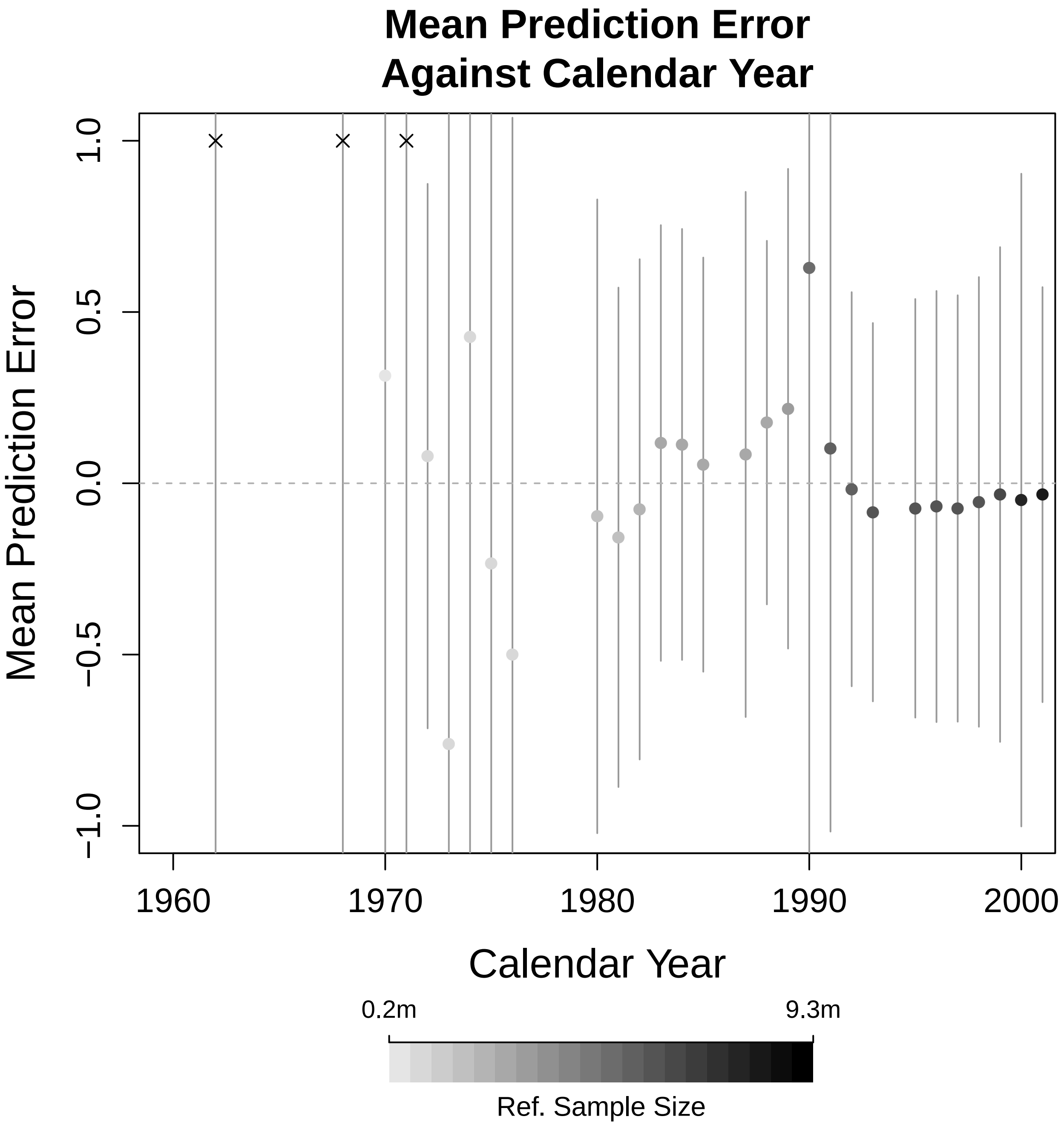

Figure 5: Cumulative predictions averaged across all targets and accumulated reference data by year. The dots are shaded according to accumulated sample size while the vertical bars represent 2 standard errors. Extrapolation errors greater than 1 in absolute value are indicated with black x's.

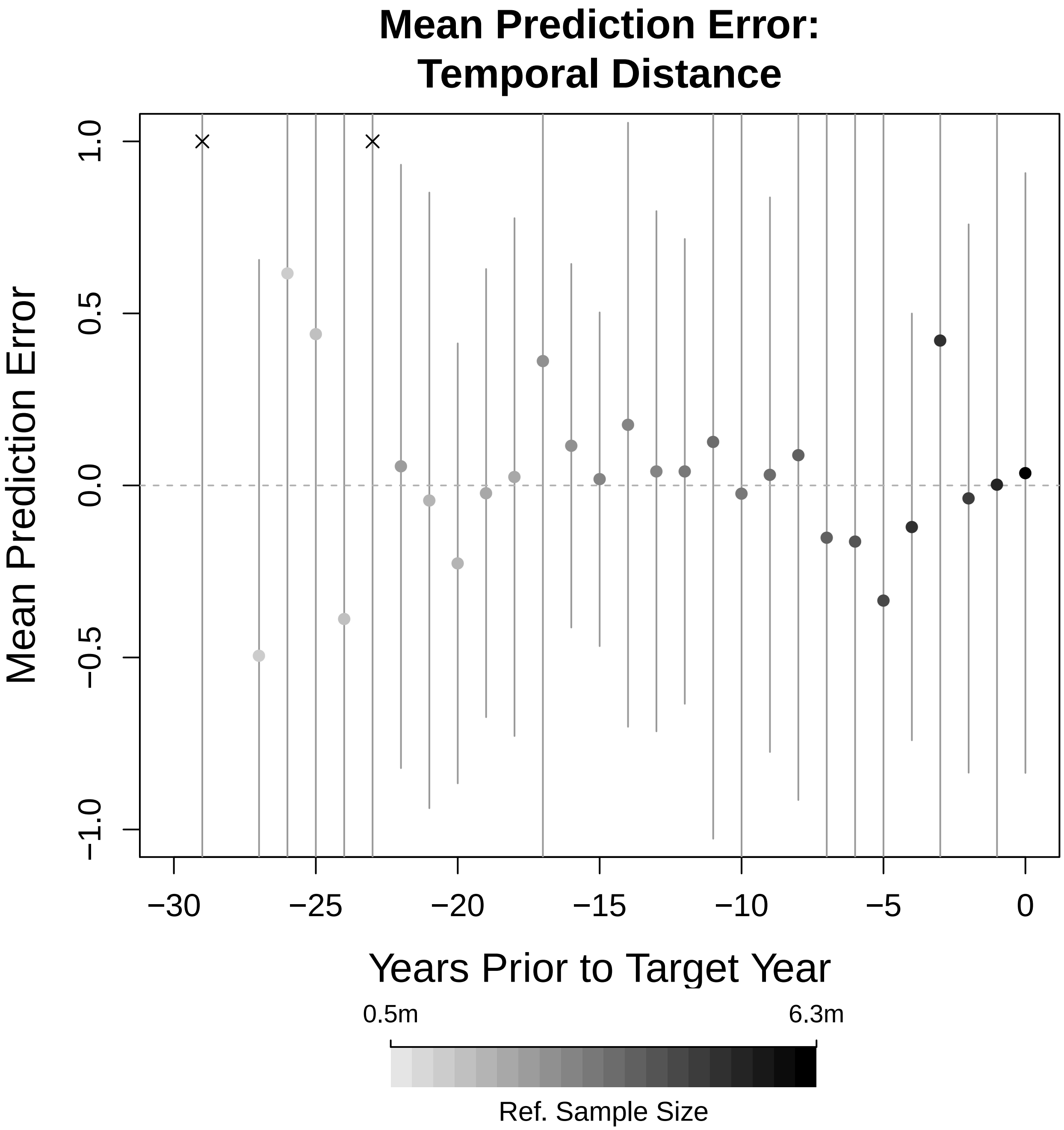

Figure 6: Average cumulative predictions across all target country-years. Averaging with t=0 being year of observation. The dots are shaded according to accumulated sample size while the vertical bars represent 2 standard errors. Extrapolation errors greater than 1 in absolute value are indicated with black x's.

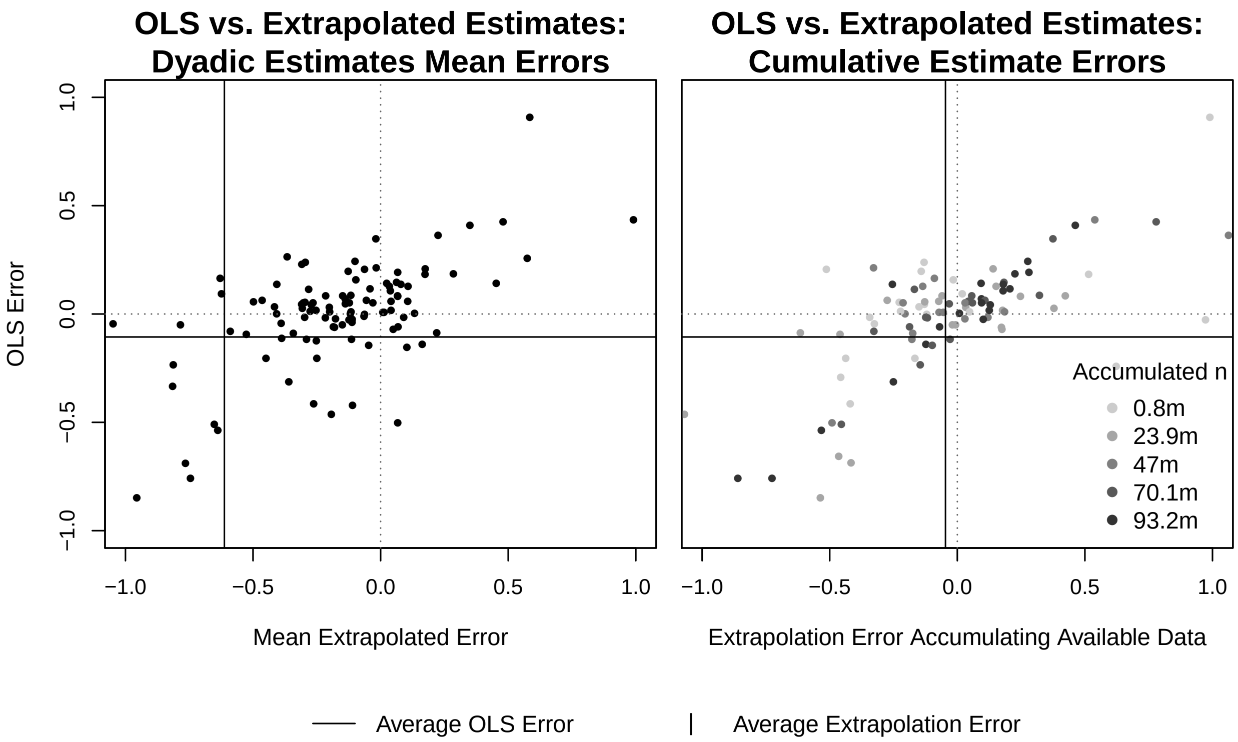

Figure 7: Scatter plot of prediction errors associated with OLS (y-axis) and extrapolation method (x-axis). The left panel averages extrapolation error over all possible dyads for each target country-year. The right panel chooses the largest available cumulative reference data and includes both micro and macro-level covariates in the reweighting.

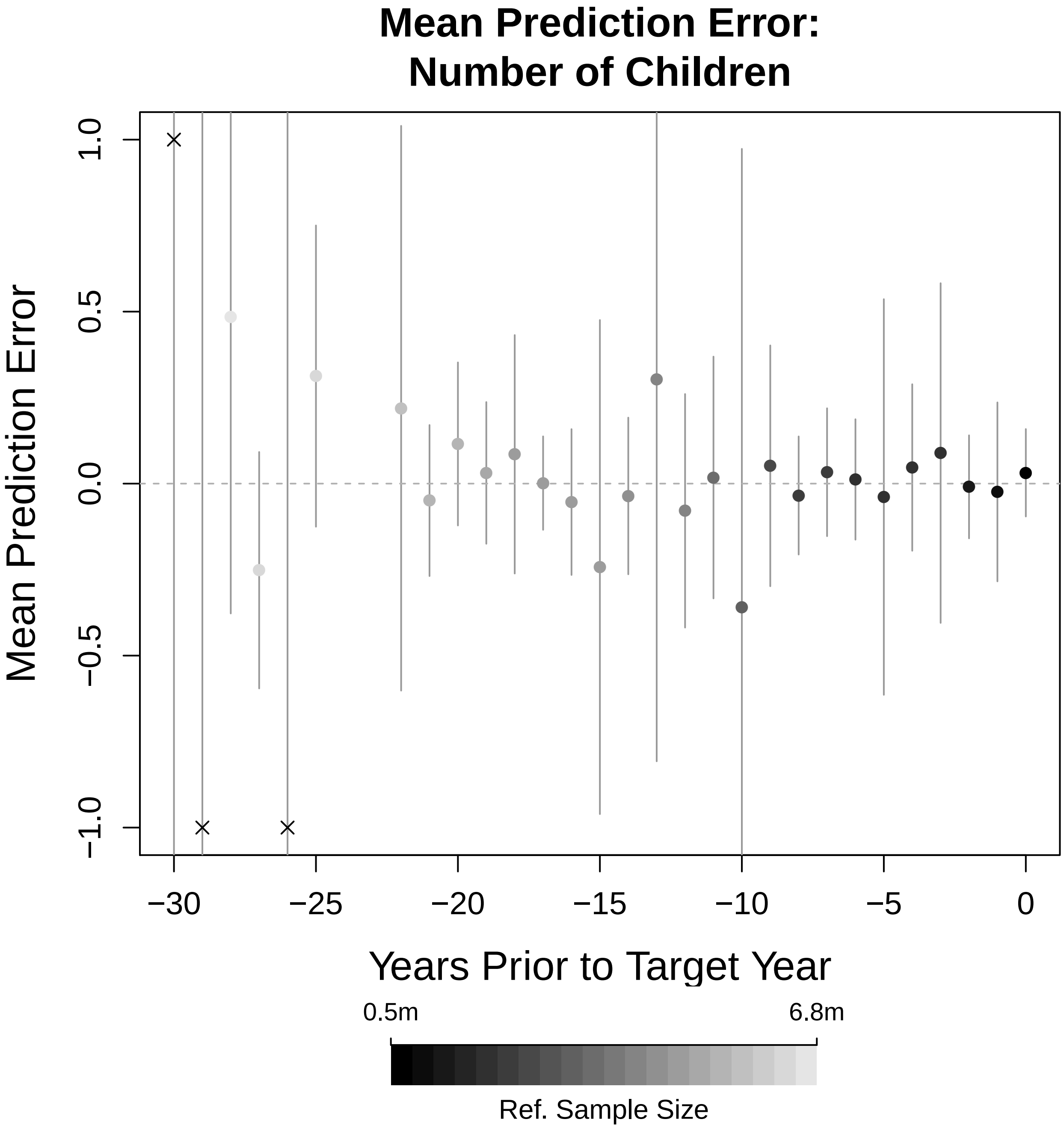

Figure 8: Average cumulative predictions across all target country-years. Averaging with t=0 being year of observation. The dots are shaded according to accumulated sample size while the vertical bars represent 2 standard errors. Extrapolation errors greater than 1 in absolute value are indicated with black x's. Treatment measured by number of children.

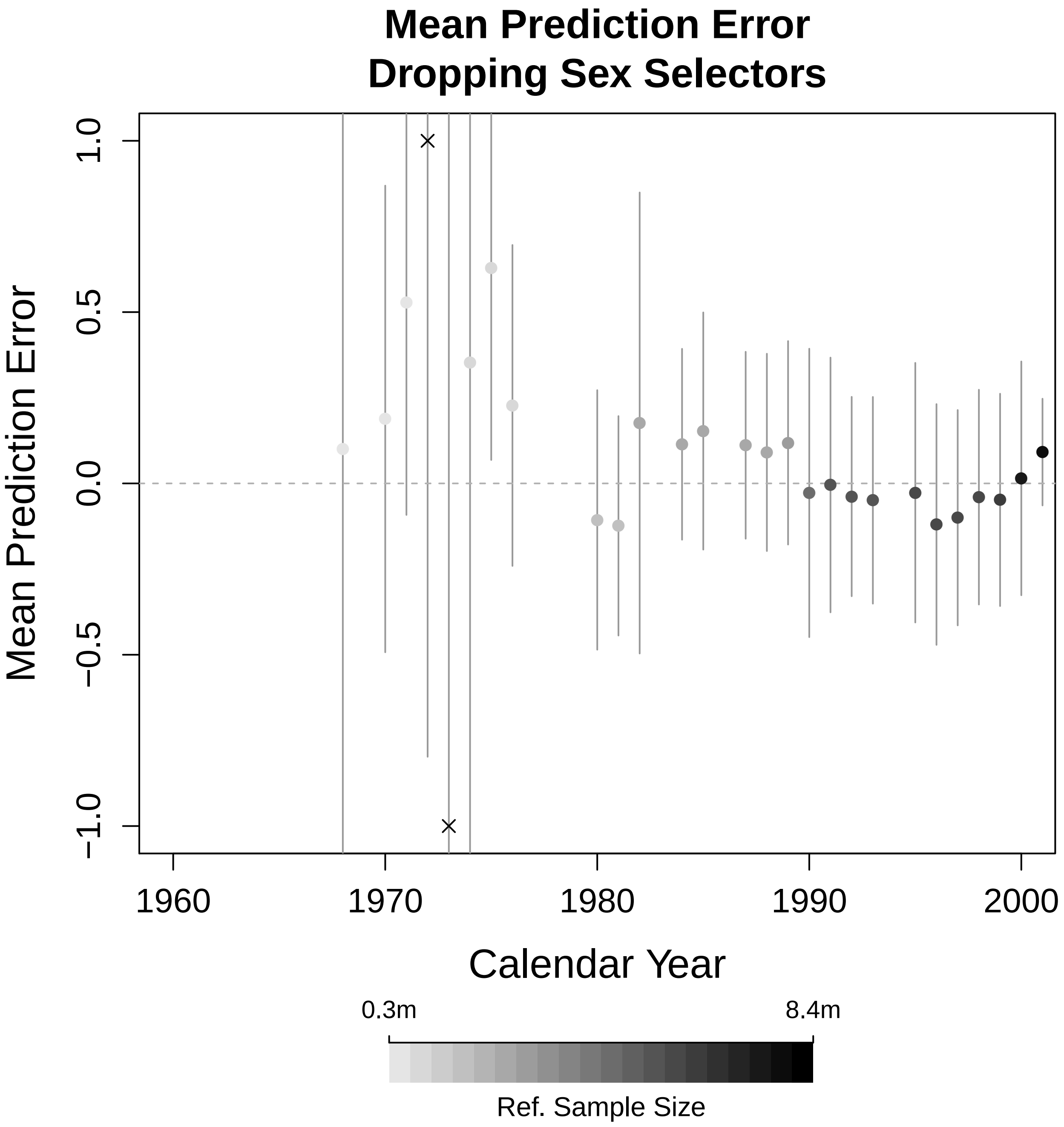

Figure 9: Cumulative predictions averaged across all targets and accumulated reference data by year. The dots are shaded according to accumulated sample size while the vertical bars represent 2 standard errors. Extrapolation errors greater than 1 in absolute value are indicated with black x's. Sex selectors dropped from sample include China, India, Nepal, and Vietnam.

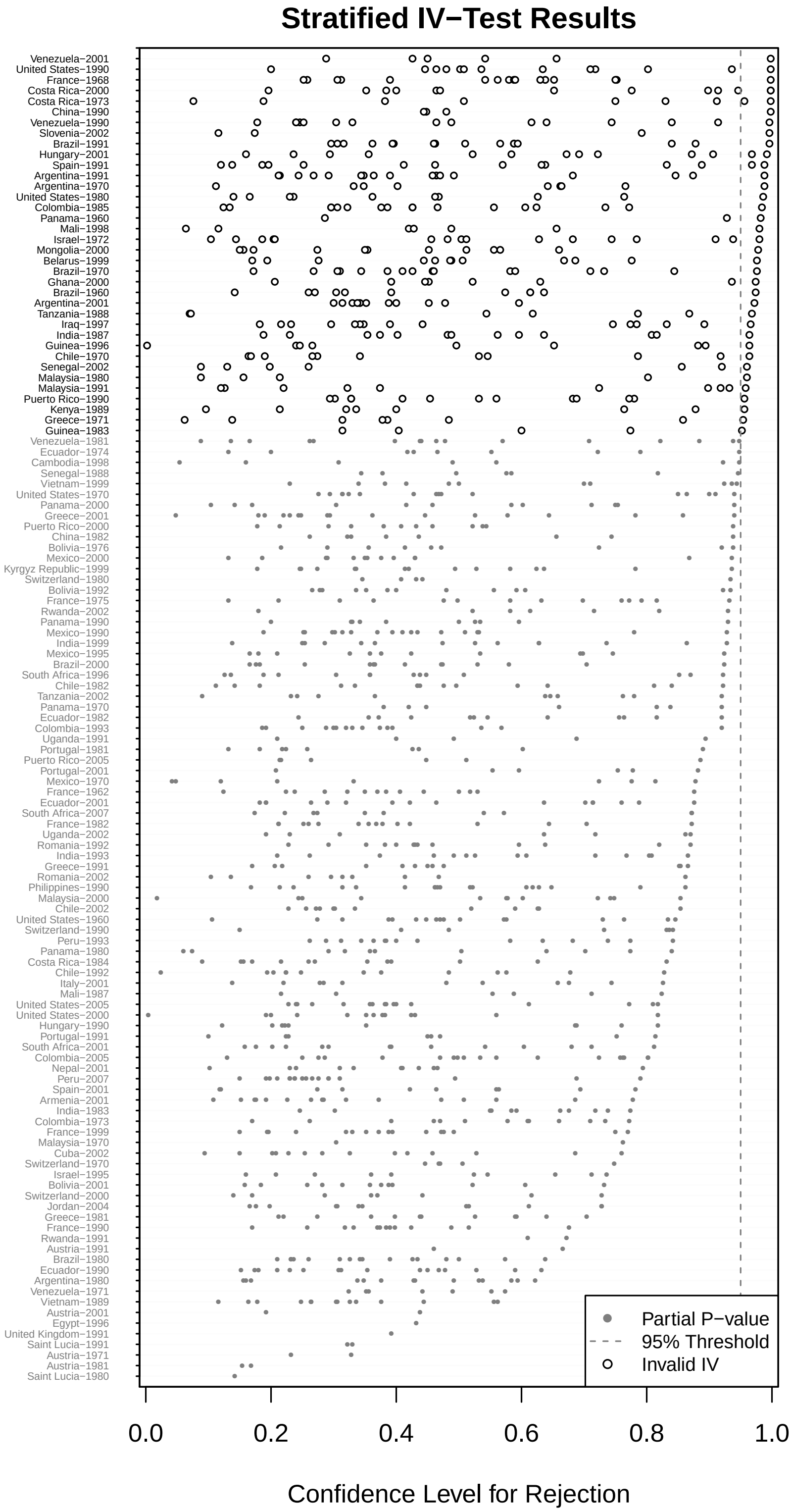

Figure 10: Partial p-values for each of 18 possible strata, ranked by maximum confidence level for rejection of IV validity. Country-years with at least one strata failing the validity test at the 95% level of confidence (depicted by hollow-circles) are dropped.

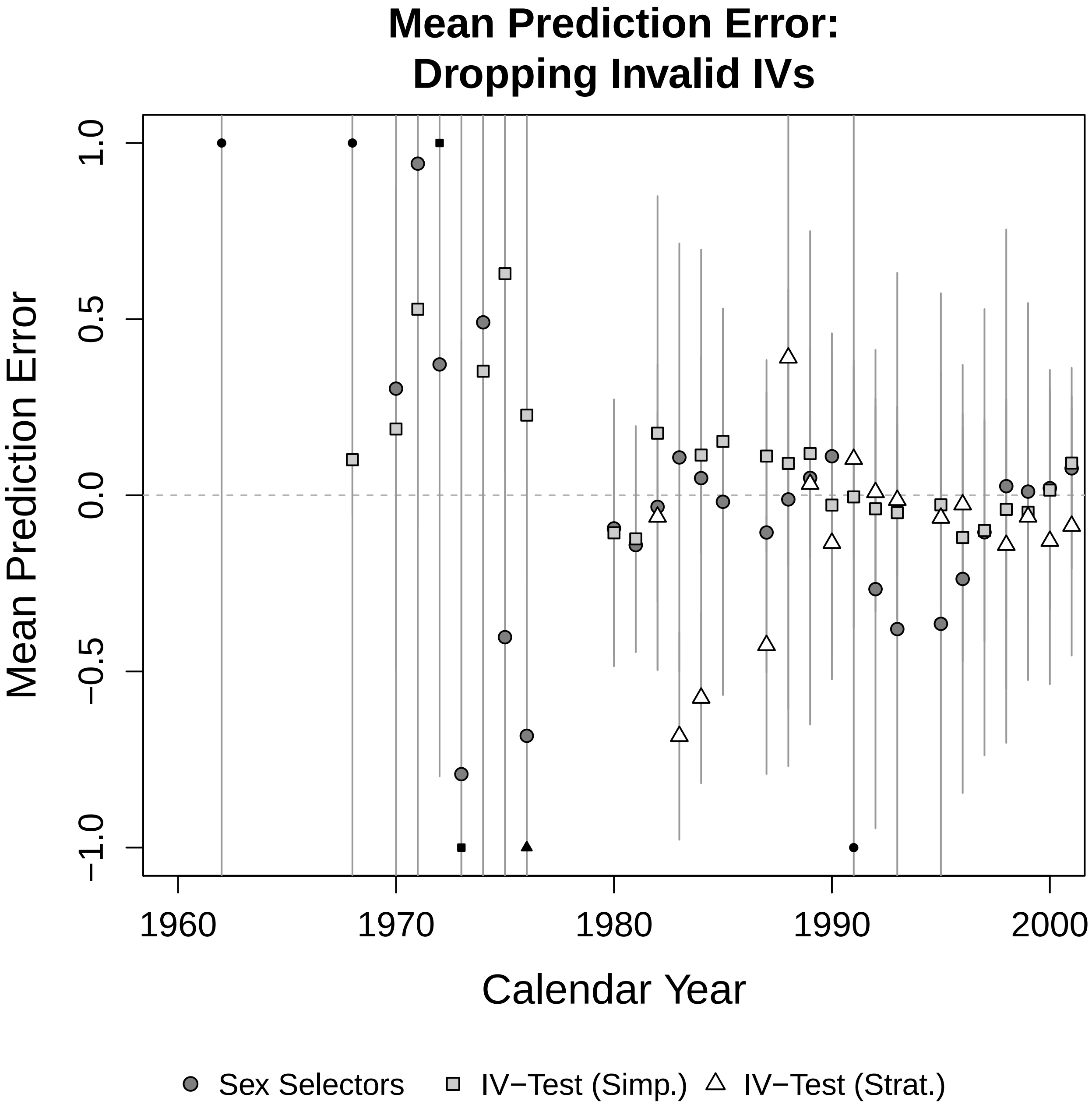

Figure 11: Cumulative predictions averaged across all targets and accumulated reference data by year. The dots indicate different robustness checks according to iv-tests. Extrapolation errors greater than 1 in absolute value are indicated with smaller black icons.

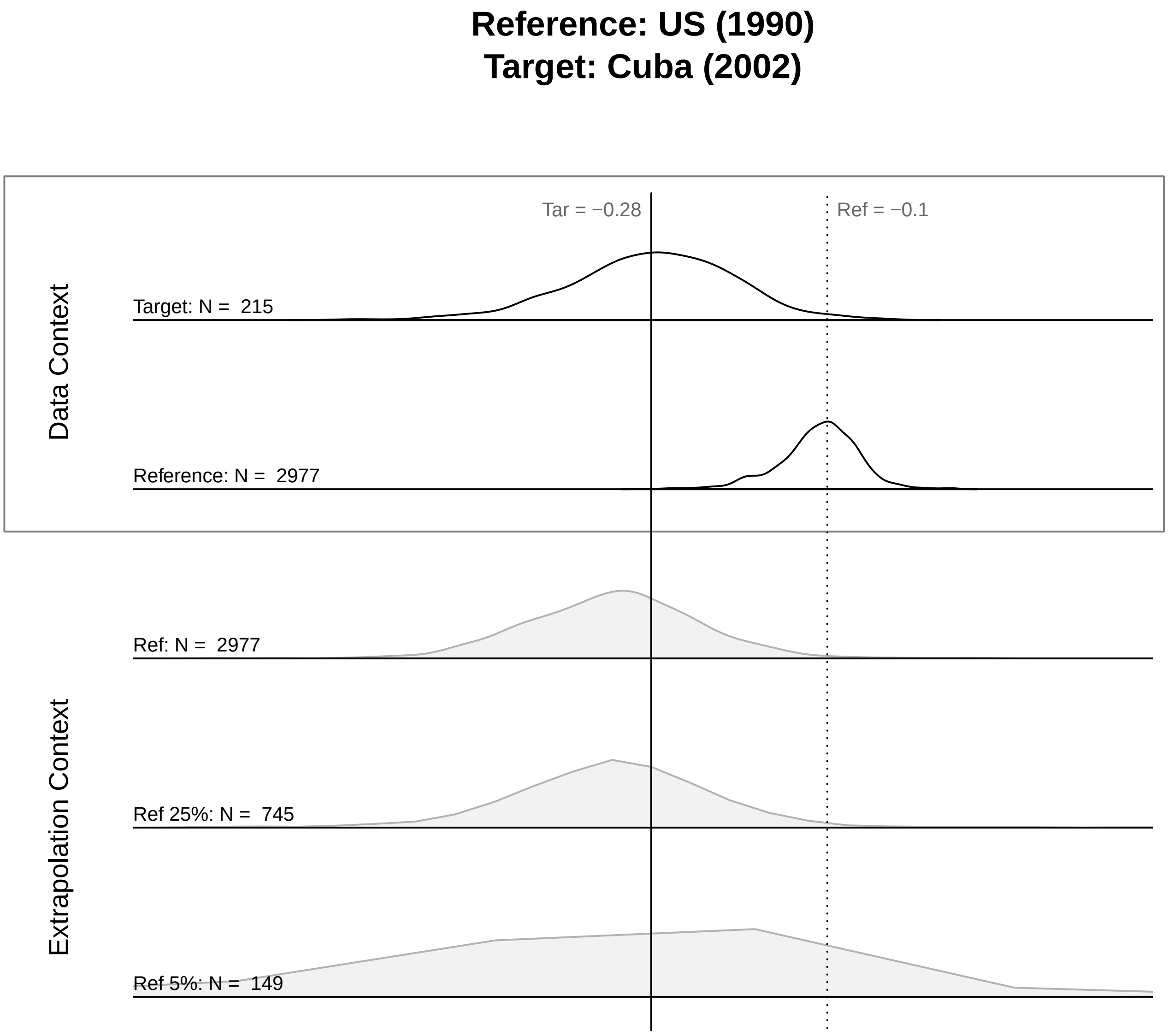

Figure B1: Simulation results for recovering LATE estimated in Cuba in 2002 using data from the United States in 1990. The x-axis measures the simulated LATEs with densities reflecting the distribution over 1,000 simulation runs. The top two densities depict the target LATE and the unadjusted reference LATE. The bottom three figures plot the adjusted reference LATEs after applying the extrapolation procedure.

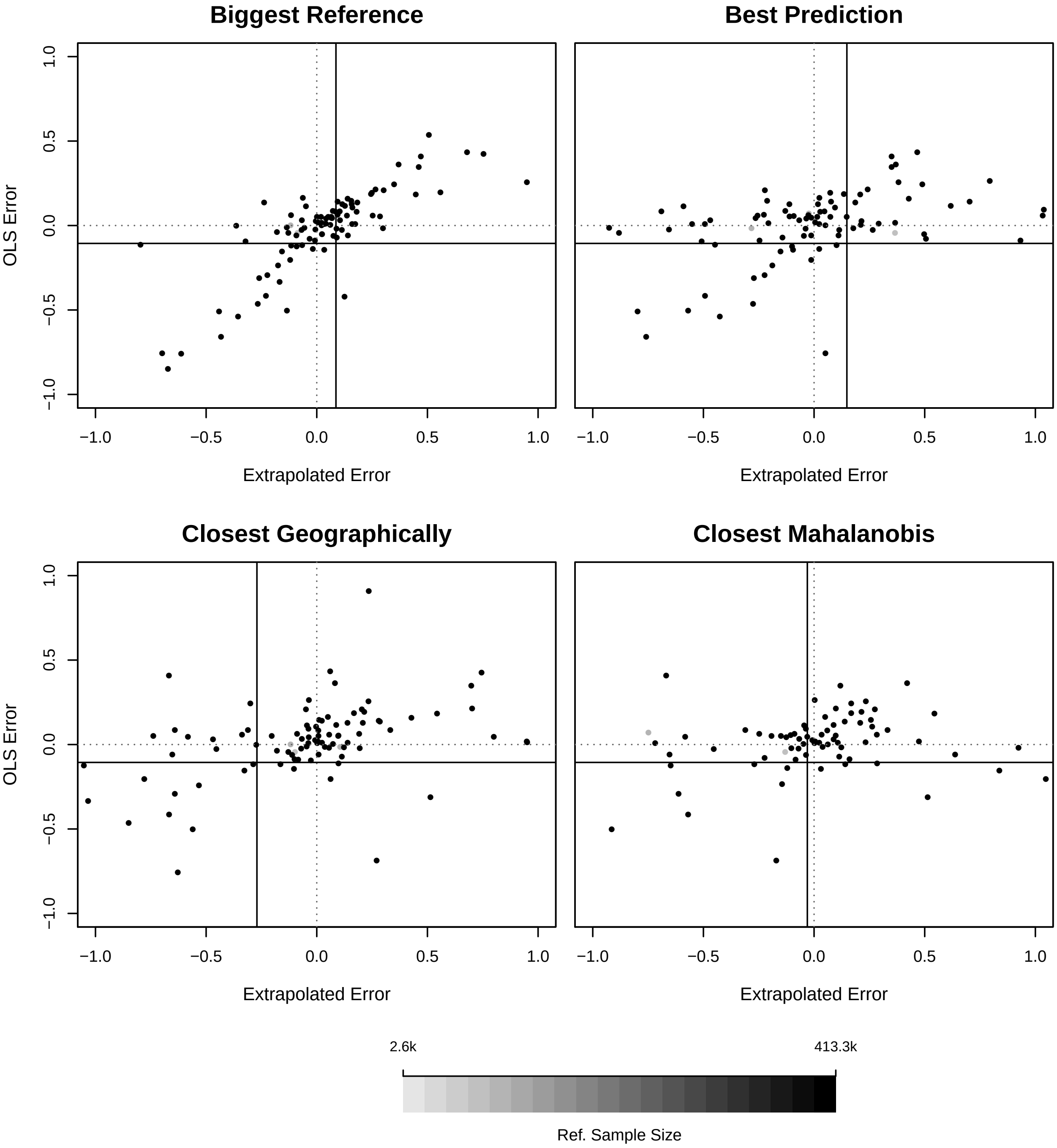

Figure C1: Scatter plot of prediction errors associated with OLS (y-axis) and extrapolation method (x-axis). The top-left panel extrapolates using the largest available reference dyad. The top-right panel chooses the reference dyad best able to extrapolate to the target in the prior year. The bottom-left panel chooses the dyad that is geographically closest to the target. And the bottom-right panel chooses the dyad that minimizes the Mahalanobis distance to the target country-year. The Mahalanobis distance is calculated on mother's age at survey, mother's age at first birth, mother's educational attainment, spouse's educational attainment, labor force participation rate, total fertility rate, and per capita GDP.

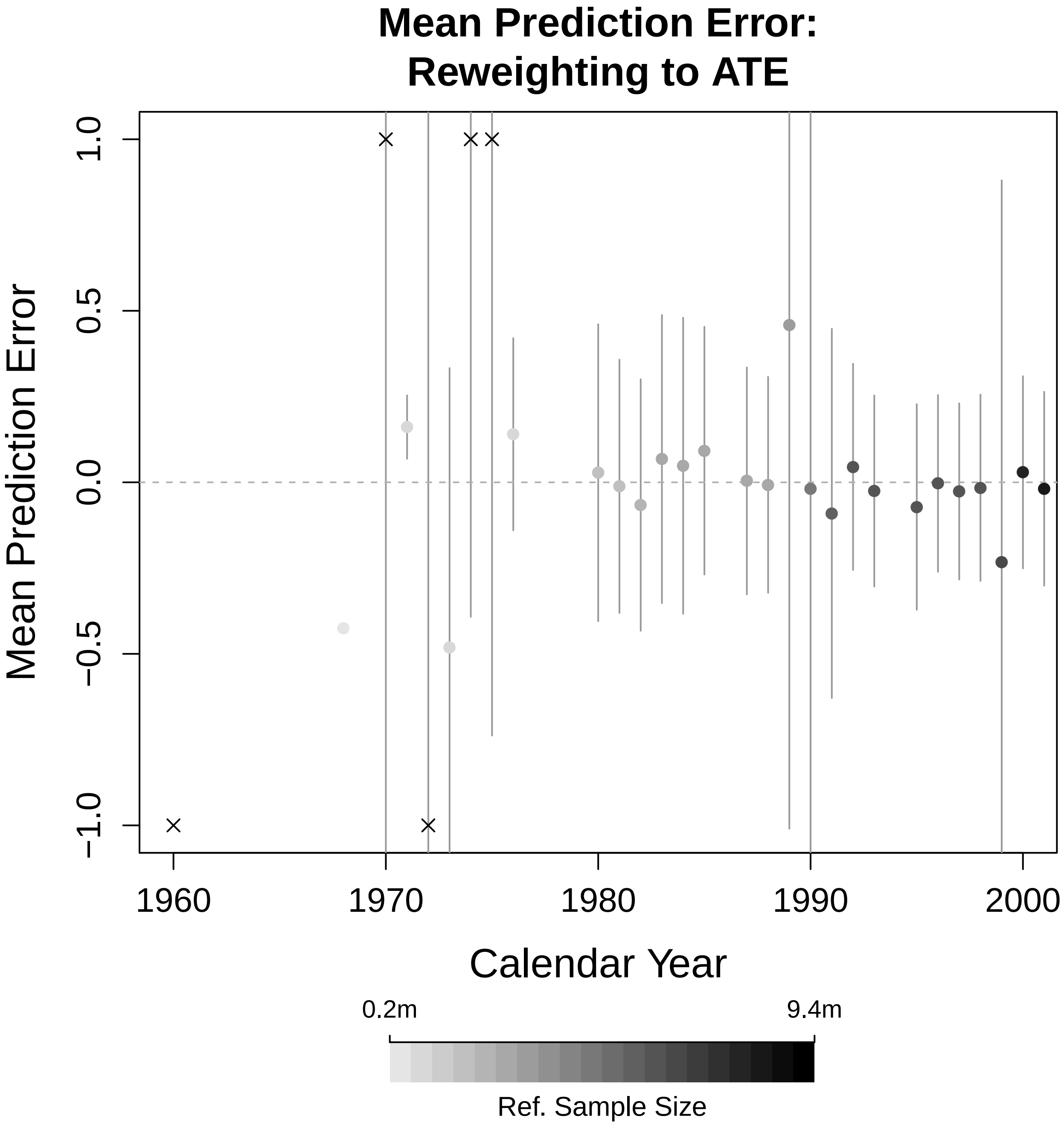

Figure D1: Cumulative predictions averaged across all targets and accumulated reference data by year. The dots indicate different robustness checks according to iv-tests. Extrapolation errors greater than 1 in absolute value are indicated with black x's. Extrapolation results are targeting ATE in target instead of LATE.

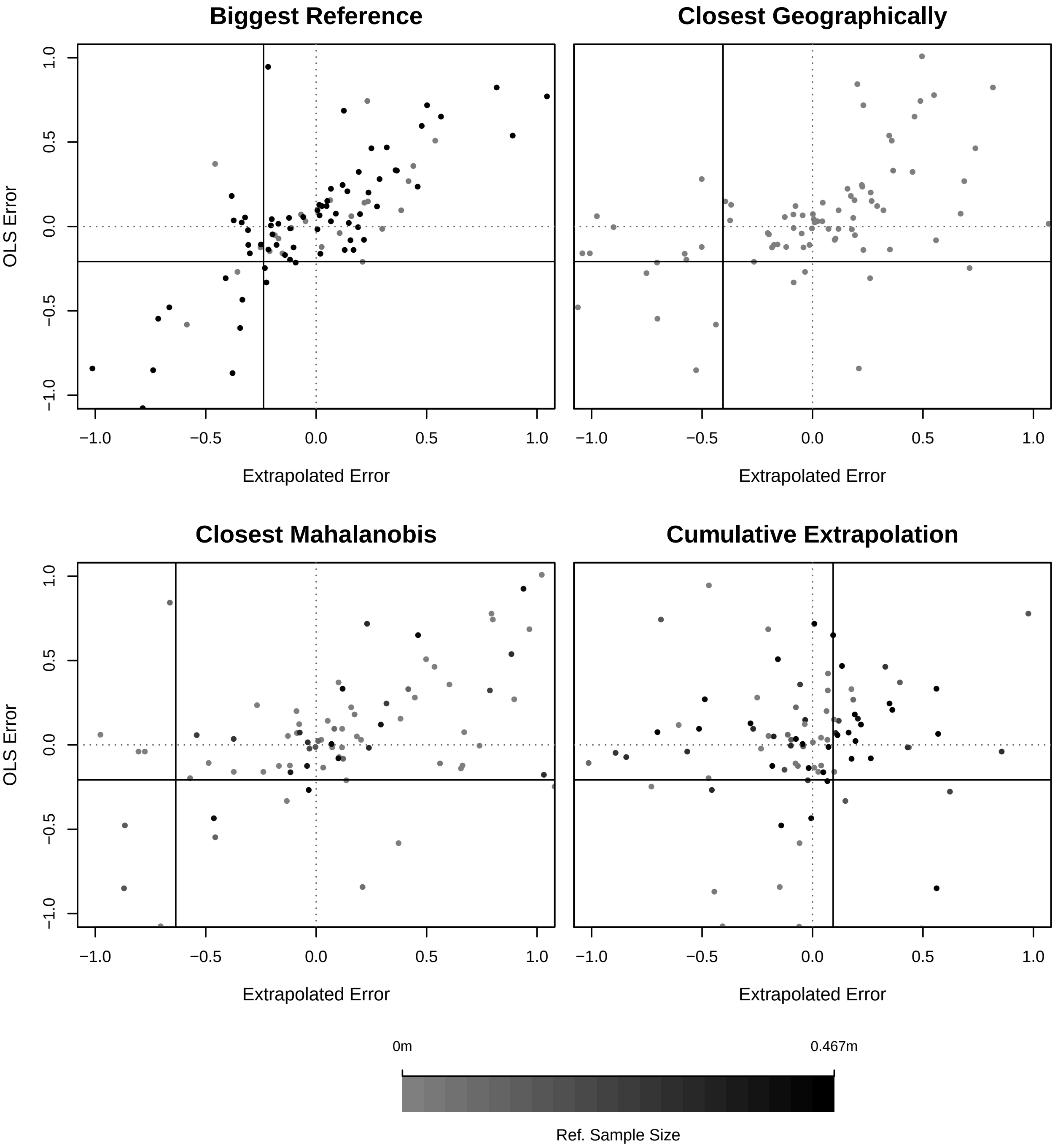

Figure D2: Scatter plot of prediction errors associated with OLS (y-axis) and extrapolation method (x-axis). The top-left panel extrapolates using the dyad that is geographically closest to the target. The top-right panel chooses the reference dyad best able to extrapolate to the target in the prior year. The bottom-left panel chooses the dyad that minimizes the Mahalanobis distance to the target country-year. The Mahalanobis distance is calculated on mother's age at survey, mother's age at first birth, mother's educational attainment, spouse's educational attainment, labor force participation rate, total fertility rate, and per capita GDP. And the bottom-right panel chooses the largest available cumulative reference data and includes both micro and macro-level covariates in the reweighting. Extrapolation technique reweighted to target ATE instead of LATE.

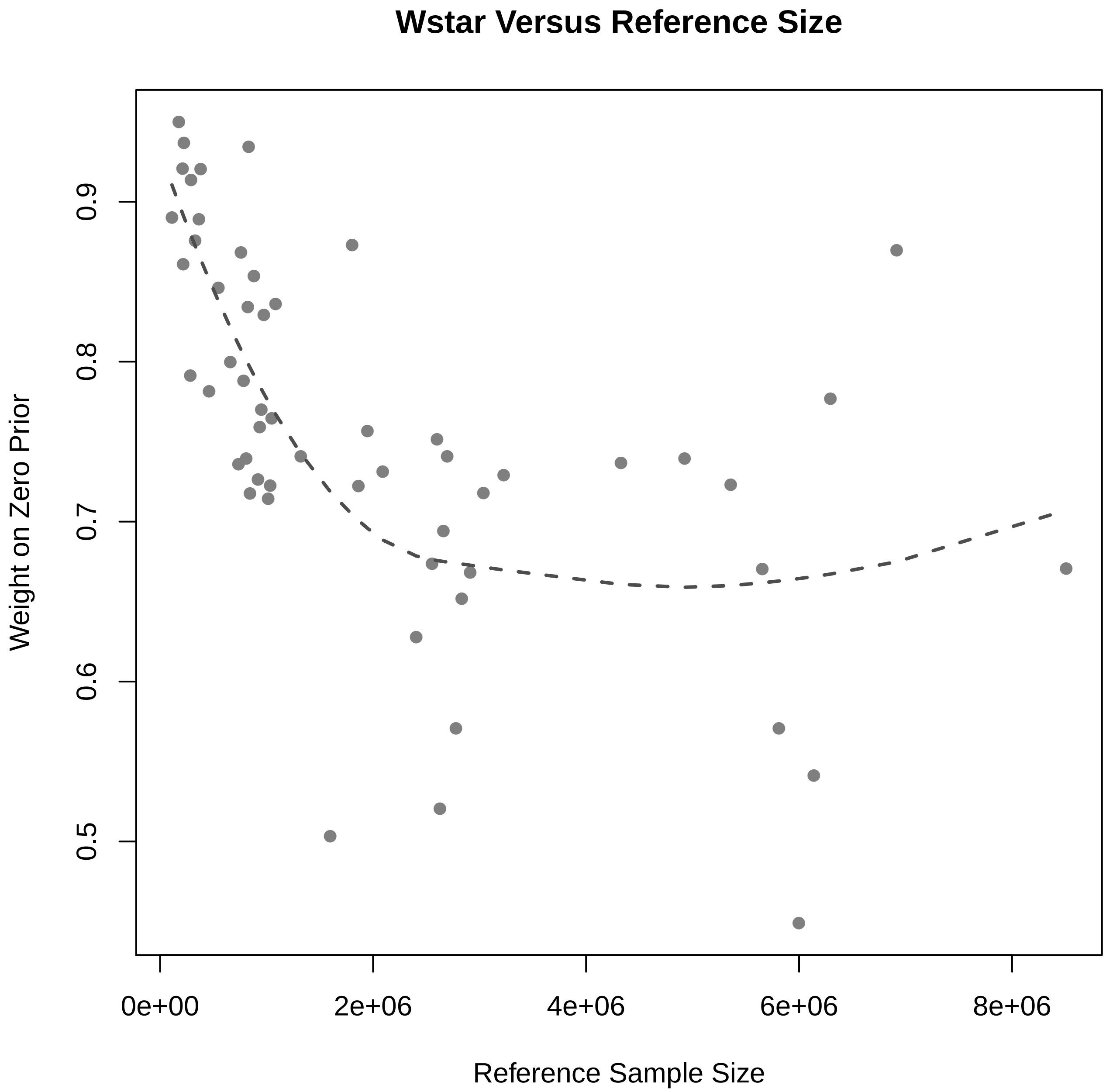

Figure E1: Optimal weights on zero prior in full data on y-axis, calculated as binned averages over range of reference sample sizes. Dotted line represents loess smoother with α parameter of 0.9.

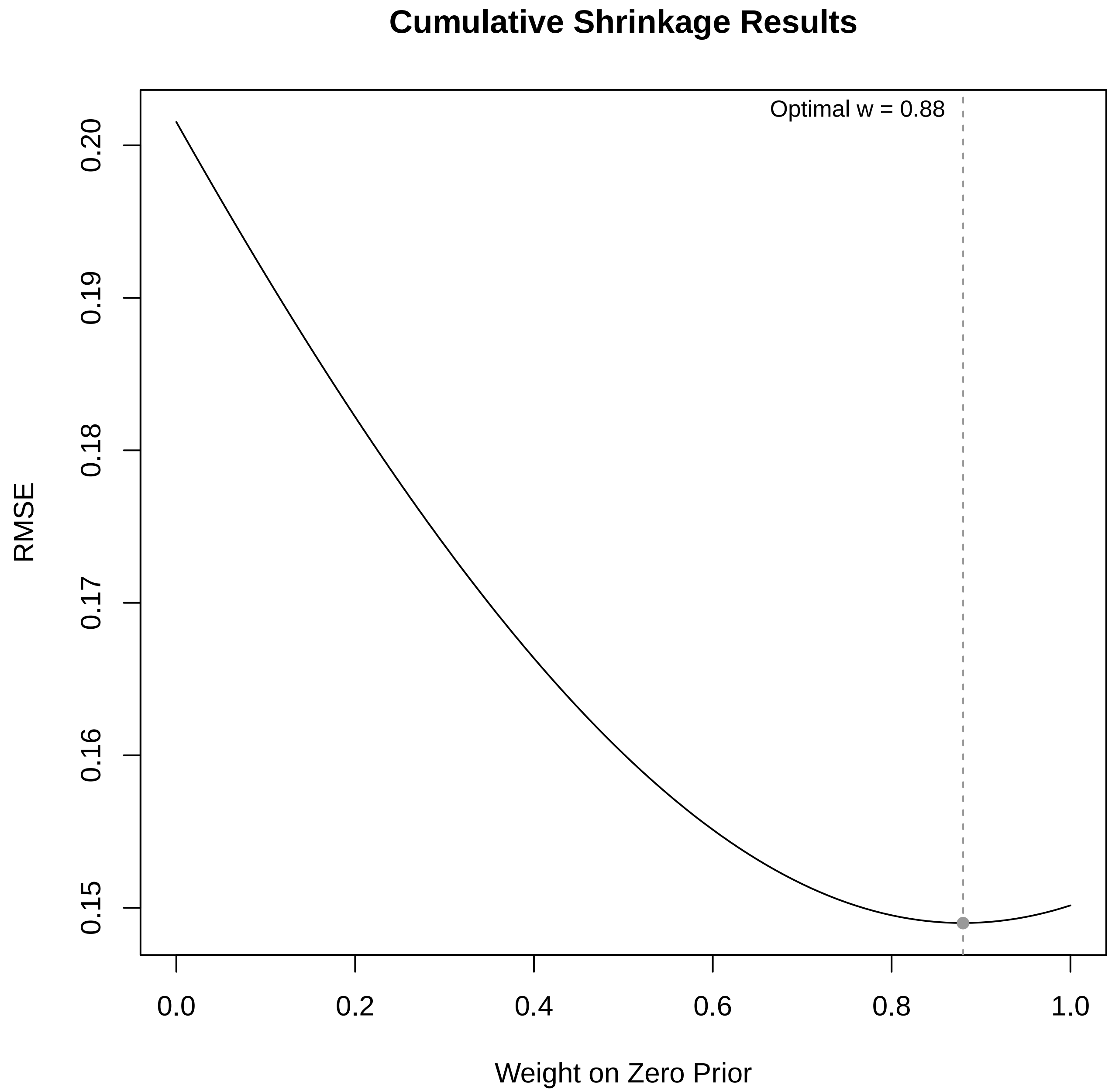

Figure E2: Cumulative shrinkage results for all data, calculated on most recent cumulative reference population for each target country-year. Weight on zero prior (w) that minimizes RMSE for all target country-years across all targets measured on x-axis, such that higher values on x-axis reflect greater weight on zero prior. RMSE on y-axis. w is a function of extrapolated estimate variance as follows: $\beta_{sh} = \left(\frac{(1-w)*Var(\beta_{ext})}{Var(\beta_{ext})}\right) * \beta_{ext} for w \in [0,1]$. RMSE is calculated as follows: $\sqrt{\frac{1}{N}\sum_{i=0}^{1}(\beta_{tar} - \beta_{sh_i})^2}$ where i indexes shrinkage weights for the full data. Minimizing shrinkage value given by gray circle.