NBER WORKING PAPER SERIES

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

James Bisbee
Rajeev Dehejia
Cristian Pop-Eleches
Cyrus Samii

Working Paper 21663
http://www.nber.org/papers/w21663

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2015

We are grateful to Josh Angrist, David Card, John DiNardo, and conference participants at A Festschrift in Honor of Robert Lalonde for invaluable comments and suggestions. In addition to the cited references, this paper draws extensively on the framework developed in Dehejia, Pop-Eleches, and Samii (2014), extending the external validity framework developed there in the context of experiments to instrumental variables estimates. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Local Instruments, Global Extrapolation: External Validity of the Labor Supply-Fertility Local Average Treatment Effect
James Bisbee, Rajeev Dehejia, Cristian Pop-Eleches, and Cyrus Samii
NBER Working Paper No. 21663
October 2015
JEL No. C26,J01,J1,J13,J22

ABSTRACT

We investigate whether local average treatment effects (LATE’s) can be extrapolated to new settings. We extend the analysis and framework of Dehejia, Pop-Eleches, and Samii (2015), which examines the external validity of the Angrist-Evans (1998) reduced-form natural experiment of having two first children of the same sex on the probability of an incremental child and on mother’s labor supply. We estimate Angrist and Evans's (1998) same-sex instrumental variable strategy in 139 country-year censuses using data from the Integrated Public Use Micro Sample International. We compare each country-year's LATE, as a hypothetical target, to the LATE extrapolated from other country-years (using the approach suggested by Angrist and Fernandez-Val 2010). Paralleling our findings in Dehejia, Pop-Eleches, and Samii (2015), 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. Our results suggest that – at least for our application – there is hope for external validity.

James Bisbee
Department of Politics
New York University
19 West 4th Street
New York, NY 10012
jameshbisbee@gmail.com

Rajeev Dehejia
Robert F. Wagner Graduate School
of Public Service
New York University
295 Lafayette Street, 2nd floor
New York, NY 10012
and NBER
rajeev@dehejia.net

Cristian Pop-Eleches
The School of International and Public Affairs
Columbia University
1401A International Affairs Building, MC 3308
420 West 118th Street
New York, NY 10027
and NBER
cp2124@columbia.edu

Cyrus Samii
Department of Political Science
New York University
19 West 4th Street, 2nd Floor
New York, NY 10012
cds2083@nyu.edu

1. Introduction

Angrist and Evans (1998) use the sex composition of the first two children as an instrument for the effect of fertility on labor supply. In light of Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) and the subsequent local average treatment effect literature, the immediate empirical relevance of the Angrist-Evans finding is limited not just to the United States in 1980 and 1990 but also to the subpopulation of compliers - i.e. those mothers whose fertility was increased as a result of having had two first children of the same sex. In this sense, it is doubly local.

At the same time, it is natural as social scientists to view these results in a more general light: they hopefully reflect the underlying relationship between family size and a woman's labor supply. As an economic issue, the connection between these two variables is of broad interest. Within the context of developed, low-fertility countries, increases in fertility could lead to reductions in female labor supply and labor force participation. In developing countries, where fertility rates are higher but declining, the reverse effect is relevant: reducing fertility could significantly increase female labor force participation and spur economic growth.

In this paper we address the tension between these two perspectives. Using the Integrated Public Use Micro Sample International (IPUMS-I) data set, we implement the Angrist-Evans same-sex instrumental variable strategy in 139 country-year censuses. The censuses span the world geographically (as listed in Appendix Table A-1) and cover five decades from 1960 to 2010. In particular, we use results from Abadie (2003) and Angrist-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 (the "reference" country) to a country-year of hypothetical interest (the "target" country). We perform this exercise on each country-year pair and use the resulting 139 × 138 dyads to examine the extent to which the local average treatment effect (LATE) can be reliably extrapolated from reference to target. We also conduct the analysis cumulatively, using all available reference country-years prior to time *t* to extrapolate to a target country-year at time *t*.

We think of the complier population in the target country as the population of hypothetical policy interest. In principle one could extrapolate a reference LATE to various subpopulations in the target country. But for the complier subpopulation in the target we can directly identify an internally valid benchmark (namely the target country-year LATE) against which to compare the extrapolation. Using our approach, we can also extrapolate both the target and reference LATE's to average treatment effects (ATEs), hence compare an extrapolated reference ATE to the target ATE; we consider this approach in Section 8.1 below.

The exercise connects to two 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 use 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). Furthermore, our exercise represents the first step in an empirical induction in a setting where general theorems are not possible. For Lalonde, the question was the extent to which non-experimental estimators could replicate an experimental benchmark. As his paper showed, even in the context of a single data set, a thoughtful attack on this question could push the literature to replicate and extend his findings and eventually reach a broader understanding. We hope that, in a modest way, our paper could provoke further investigation into the relationship between local average treatment effects and their potential to extrapolate treatment effects globally.

Second, 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 of an intervention 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 and on a specific and 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 and in Section 6, below).

In this paper we take up the challenge of examining the extent to which we can extrapolate a LATE from a quasi-experimental evidence base to new contexts of interest. Using the Angrist and Fernandez-Val (2010) framework, we assume that heterogeneity in the local average treatment effect is driven by differences in observable characteristics of the complier population. By characterizing the complier population in each country-year sample, we can calibrate the local average treatment effect from a reference country to match the complier population distribution of covariates in a target country of interest. Of course, such an exercise will succeed in reliably extrapolating the treatment effect only if the identifying assumptions are approximately true in the application and data we are studying. It is precisely in this sense that our exercise uses Lalonde (1986) as a motivation. Our data provide internally valid estimates of the target LATE's. Most immediately, this allows us to test the validity of the identifying assumption that LATE heterogeneity is driven by observable covariates. But more substantively, we are interested in characterizing which differences between an experimental reference and a target context are most likely to lead to failures in external validity. This latter goal represents a step toward systematizing the discussion of issues related to external validity.

To preview our findings, our exercise has five steps (which parallel our approach in Dehejia, Pop-Eleches, and Samii 2015). First, using simulated data, which satisfies our identifying assumptions by construction, we confirm the basic insight that calibrating the treatment effect in a reference country does yield a reliable estimate of the treatment effect in a target country. Indeed, we show that if the identifying assumptions are true, not only is external validity possible but also "super-external validity": if only a small sample size is available in the target, it can be preferable to use an externally extrapolated estimate from a reference country where the treatment effect is more precisely estimated (i.e., with a larger sample size).

Second, we graphically document that there is considerable heterogeneity both in the first-stage and in the instrumental variable estimates across our sample. We also find substantial variation in the characteristics of the complier populations.

Third, we use the dyads described above to run regressions of external prediction error against differences between the target and reference countries and to characterize empirically which differences significantly drive prediction error. We show that prediction error increases considerably in covariate differences between the reference and target contexts of interest.

Fourth, we use the timeline of our country-year samples to examine how the accumulation of evidence from additional samples affects out-of-sample predictive accuracy. We find that, using all available evidence from reference country-years, extrapolation error is low in the sense of being statistically indistinguishable from zero while also being bounded quite tightly.

Fifth, we compare the prediction error from our extrapolation technique to the prediction error from using OLS within the target site. Our interest here is to understand the extent to which non-experimental evidence, even if it recovers a treatment effect that is both potentially biased and different from the complier population, compares to externally extrapolated quasi-experimental evidence from an instrumental variable estimate from another site or sites. In the development economics literature, Pritchett and Sandefur (2013) have speculated that bias from a non-experimental estimate within the target site might be smaller than bias due to failure in external validity. In our application we find the opposite: extrapolation error tends to be lower than endogeneity bias from within-target OLS estimation.

The paper begins by outlining our identifying assumptions and empirical approach, and then examines each of these five questions in turn, followed by a series of extensions and robustness checks.

2. Methods

We define the conditions and methods for extrapolating from an instrumental variables estimate to the causal effect in a target population.[1] We focus on the case of using local average treatment effects estimated from a set of reference contexts to identify and estimate the LATE in a target context, where the LATE's are the complier average causal effects in the respective populations (Angrist et al. 1996). We focus on extrapolating to the LATE in a target context because of the nature of the data that we have for our benchmarking exercise. The IPUMS-I data allow us to estimate LATE's in different populations defined by country and year. As such, we can use these data to conduct benchmarked comparisons between actual, estimated LATE's 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.

Following Angrist and Fernandez-Val (2010), our setup supposes that a randomly sampled unit *i* has two potential outcomes, $Y_i(1)$ and $Y_i(0)$, that would obtain under assignment of a treatment D_i to the treated (=1) versus control (=0) condition, respectively. 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 are given by,

$$Y_i = \alpha + r_i D_i + \eta_i \tag{1}$$

[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.

where $\alpha = E[Y(0)]$, $\eta_i = Y_i(0) - \alpha$, and $\gamma_i = Y_i(1) - Y_i(0)$ $\alpha = E[Y(0)]$is the unit-level causal effect of D_i. We also define an instrument, $Z_i = 0,1$, that affects treatment assignment. Thus, we have potential treatment assignments, $D_i(1)$ and $D_i(0)$, corresponding to the treatment values that would obtain for a unit under $Z_i = 1$ versus $Z_i = 0$, respectively. In our application, the instrument is an indicator for whether the sexes of the first two children are the same. The observed treatment is given by,

$$D_i = \gamma + p_i Z_i + \nu_i \tag{2}$$

where $\gamma = E[D(0)]$, $\upsilon_i = D_i(0)- \gamma$, and $p_i = D_i(1) - D_i(0)$.

Units are characterized by covariates, X_i, which include both unit- and population-level variables. These covariates play a central role in extrapolation. We assume throughout that conditions required for identifying the conditional LATE hold (Angrist and Fernandez-Val 2010, p. 7). These include:

C1(a) *Conditional independence and exclusion*: $(Y(1),Y(0),D(1),D(0)) \perp\!\!\!\perp Z|X$.

C1(b) *Valid conditional first stage*: $E[p|X] \neq 0$ *and* $0 < P[Z=1|X] < 1$.

C1(c) *Conditional monotonicity*: $P[D(1) \geq D(0)|X] = 1$ *or* $P[D(1) \leq D(0)|X] = 1$.

We also define an indicator, W, for whether a population is the target to which we want to extrapolate, $W=1$, or whether it is part of the reference set for which we have LATE estimates, $W=0$.

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

$$\Delta_z(x,1) = E[r|D(1) > D(0), X = x, W = 1], \tag{3}$$

and in a reference set population as

$$\Delta_z(x,0) = E[r|D(1) > D(0), X = x, W = 0]. \tag{4}$$

Given covariate-specific LATE's for a target population *(W = 1)*, the marginal LATE for the target context is given by (Froelich 2007):

$$\Delta_z(1) = \int \Delta_z(x,1) dF(x|D(1) > D(0), W = 1). \tag{5}$$

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, following Hotz et al. (2005).

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

C2 *Unconfounded location: (Y(1),Y(0))* $\perp\!\!\!\perp$ *W|D(1) > D(0),X=x, and*

C3 *Covariate overlap: 0 < P[W = 0|D(1) > D(0),X = x] <1.*

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

Proof: Under C1-C3, we have

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

for all x in the support of X in the target population, in which case

$$\Delta_z(1) = \int \Delta_z(x,0) dF(x|D(1) > D(0), W = 1). \tag{7}$$

■

We adopt an estimation approach based on interacted regressions.

Proposition 2 (Complier-centered interaction estimation): *Suppose conditions C1-C3 hold and that*

$$Y_i = \beta_0^Y + \beta_1^Y Z_i + \sum_{k=1}^{K} (X_{ik}^c \phi_k^Y + Z_i X_{ik}^c \lambda_k^Y) + \epsilon_i^Y, \tag{8}$$

and

$$D_i = \beta_0^D + \beta_1^D Z_i + \sum_{k=1}^{K} (X_{ik}^c \phi_k^D + Z_i X_{ik}^c \lambda_k^D) + \epsilon_i^D, \tag{9}$$

with X_{ik}^c being the value of the covariate X_{ik} centered on the sample complier mean in the target context ($W = 1$), and $E(Z_i \epsilon_i^R) = E(X_{ik} \epsilon_i^R) = 0$ for $R = Y, D$. Let $\tilde{Z}_i$ be the residuals from a linear regression of the sample Z values onto the matrix of sample $(X_1^c, \ldots, X_K^c, ZX_1^c, \ldots, ZX_K^c)$ values. For a sample of N_0 units with $W_i = 0$,

$$plim \frac{\sum_{i=1}^{N_0} \frac{(Y_i - \bar{Y})\tilde{Z}_i}{\tilde{Z}_i^2}}{\sum_{i=1}^{N_0} \frac{(D_i - \bar{D})\tilde{Z}_i}{\tilde{Z}_i^2}} = \Delta_z(1). \tag{10}$$

Proof: By standard results for centered regression with heterogeneous effects (e.g., Imbens and Wooldridge 2009, pp. 28-30), we have

$$plim \sum_{i=1}^{N_w} \frac{(R_i - \bar{R})\tilde{Z}_i}{\tilde{Z}_i^2} = \int \frac{Cov(R, \tilde{Z} | X = x)}{Var(\tilde{Z} | X = x)} dF(x | D(1) > D(0), W = w). \tag{11}$$

for $R = Y, D$. 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 consist of indicators for an exhaustive set of strata, estimation via a centered interaction regression is algebraically equivalent to the type of 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 ($D(1) > D(0)$). By Theorem 3.1 of Abadie (2003), one can accomplish this task via "kappa weighting." Specifically, for our target population with $W = 1$, we have

$$E[X|D(1) > D(0), W = 1] = \frac{E[\kappa(X,1)X]}{E[\kappa(X,1)]}, \tag{12}$$

where

$$\kappa(x,1) = 1 - \frac{D(1-Z)}{1 - E[Z|X = x, W = 1]} - \frac{(1-D)Z}{E[Z|X = x, W = 1]}. \tag{13}$$

We use 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.

3. Illustrative simulation

We use a simulation to demonstrate the properties of IV extrapolation under assumptions C1-C3. To construct a naturalistic 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 223 observations) and the United States in 1990 (yielding 3,343 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 used these covariates to generate potential outcomes under treatment and control for women in each simulation replicate sample.[2] We generated 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.

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 assumptions C1a and C1c based on arguments of Angrist and Evans (1998). To generate the extrapolations, we use the complier-centered interactions 2SLS model defined above. We conduct 1,000 simulation runs.

Figures 1 and 2 display results from the simulation exercise. Figure 1 shows results for simulations where Cuba was the target and the US was the reference sample. We see in Figure 1 that the distribution of extrapolations is centered on the true LATE. The extrapolation

[2] The potential outcomes were agnostically generated by a vector of coefficients drawn from a normal distribution. For a detailed description of the simulation data-generating process, please see Appendix B.

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 (3,343 observations) is much larger than the target population sample (223 observations).

[FIGURE 1 ABOUT HERE]

Figure 2 focuses only on the extrapolated estimate distributions as we reduce the size of the reference population. Again, the target distribution is depicted in light gray in the background. As we reduce the size of the extrapolation population, we see the distribution widen until the smallest reference size of 167 observations is even less precise than the target distribution, simulated using 223 observations. This outcome illustrates the theoretical possibility of “super-external validity,” whereby extrapolations from an existing evidence base may provide more accurate estimates of the LATE in a target context than would be the case if one estimated the LATE using only data from that context.

[FIGURE 2 ABOUT HERE]

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 LATE’s 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.

4. A world of LATE's: same-sex, more kids, and mothers' labor supply

4.1 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 incremental 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 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 to take the Angrist-Evans strategy to the world. The IPUMS-I data provided harmonized coding that in principle yields measures of the above variables that are comparable across countries and years. Data are available for a maximum of 139 country-years, although accounting for missing data our sample becomes smaller for some specifications. Individual covariates include the mother's age, her age at 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.

[TABLE 1 ABOUT HERE]

As discussed in the introduction, while extrapolating from reference to target country makes use only of micro covariates, 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 included GDP per capita, women's 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.

In our application 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 empirically whether it affects the IV extrapolation from reference to target. In Section 8.4, we also show that dropping potentially sex-selective countries does not significantly affect the results. A second concern is violation of the exclusion restriction, 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 instrumental validity for US data). We address this in a similar fashion, namely by examining the extent to which GDP per capita affects extrapolation error and in Section 8.5 using the Huber-Mellace (2015) test to detect and drop country-years unlikely to satisfy instrument validity.

4.2 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 3, 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 geographic regions. Figure 3, panel b, plots the density of first-stage effects. We note that all but a handful of first stages 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.

[FIGURE 3 ABOUT HERE]

Figure 4 recreates the analysis of Figure 3 but replaces the first-stage results with the IV estimates of the effect of an additional child on the mother's work status. The striking difference between Figures 3 and 4 is that the IV estimates are both negative and positive. While we can see in Figure 4, 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, estimates from high labor force participation economies such as North America, Western Europe, and Eastern Europe tend to be negative, while estimates from less developed regions are more positive.

[FIGURE 4 ABOUT HERE]

In Figure 5, we examine how the population of individuals who comply with the instrument differs from the overall population. 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.

[FIGURE 5 ABOUT HERE]

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 emphasis 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 reduce external validity.

5. Dyadic 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. As noted in the introduction, 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* LATE's in the reference country. Since, as in our US-Cuba example, for any two country-years the extrapolation differs depending on which is the target country and which 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 *i* LATE estimate minus the extrapolated treatment effect from the reference country-year *j*), its standard error, and covariate differences between reference and target (which for simplicity we assume these can be summarized simply as $D_{ij} = X_i - X_j$).

As in Dehejia, Pop-Eleches, and Samii (2015) we use this setup to estimate the external validity (or *X*) function:

$$E_{ij} = \beta D_{ij} + \epsilon_{ij},$$

where we weight the regression by the inverse of the variance of the extrapolation error. In the spirit of the Heckman, Ichimura, Smith, and Todd (1998) bias function, which characterizes selection bias as a function of covariates, our interest is to characterize 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, D_{ij} includes country-year level macro covariates as well, including GDP per capita, labor force participation, and total fertility rate.

Results are presented in Tables 2 and 3. We begin in Table 2 by examining the univariate relationship between covariate differences and extrapolation error. We find that the differences in all covariates save for mother' age at first child's birth, the labor force participation rate, and temporal and geographic distances, are significant predictors in the expected direction: greater reference-target differences are associated with increased extrapolation error. The magnitude of the bias is considerable for each covariate. Since the significant covariates are in logs, the coefficients can be directly compared in terms of percent changes; this suggests that a ten percent increase in the difference between reference and target in mother's education, spouse's education, mother's age at the time of the survey, per capita GDP, the gender ratio, and total fertility rate is associated with a 0.017, 0.02, 0.009, 0.013, 0.006, or 0.01 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 (substantively an increase of roughly 4,650km) corresponds to a 5 percent increase in extrapolation error although this is not significant at conventional levels. Similarly, the effect of time is noisily measured but, again, pointing in a positive direction.

[TABLE 2 ABOUT HERE]

When we include all covariates simultaneously, we find in Table 3 that differences in the spouse's education, differences in mother's age at the time of the survey, and differences in the gender ratio remain significant in the full sample. The coefficients on spouse's education and mother's age at the time of the survey are robust to restricting the analysis to country-years where first-stage t-statistics are greater than 2, 5, and 10 although the estimate for mother's age becomes less precisely estimated. As in Table 2, the magnitudes are considerable for spouse's education. A ten percent increase in the difference in spouse's education corresponds to a 0.018 increase in the absolute difference between the target and extrapolated estimates (again relative to an average world LATE of -0.129). It is worth noting that these coefficients decline both in significance and in magnitude as we restrict the sample to stronger first stage targets. This suggests that the accuracy of our extrapolation technique is less susceptible to differences in the covariate profile when our target estimate is more precisely estimated. However, certain covariates remain significant, particularly differences in spouse's education.

[TABLE 3 ABOUT HERE]

Overall the results underline an intuitive but important result: when the target and reference countries are close together in the covariate space, extrapolation error tends to be smaller. As will be demonstrated in Section 7, this intuition motivates our use of minimized Mahalanobis distance to predict the optimal dyad for extrapolation.

6. Accumulation of evidence

While the dyadic regressions in Tables 2 and 3 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. We address both issues in Figures 6 to 11.

In Figures 6 to 9, we depict the extrapolation error for specific target country-years (Greece-1971, Colombia-1985, Belarus-1999, Colombia-2005), and show how the prediction error changes as additional reference country-years become available at each point in time. Note that in any given year, we use all of the available reference country-years available up to that point in time (excluding the target country in earlier periods) to predict the target country-year. Our exclusion of the target country data in earlier periods constitutes a more difficult test of our extrapolation technique by removing the (presumably) most similar reference data from the accumulated pool of observations.

[FIGURES 6 – 9 ABOUT HERE]

Several patterns become evident. First, as more evidence becomes available through an increased reference set, prediction error typically decreases. Second, although there are counter-examples, the pattern that extrapolation error is not statistically significant when using the maximal reference set does usually hold. Third, there are instances in which the target LATE itself is not precisely estimated, and in these cases extrapolation error, although not statistically significant, can be large in terms of magnitude. Fourth, for target countries later in the sample for which a larger reference set is available, prediction error tends to converge to zero.

Figures 10 and 11 average the extrapolation error across all target countries with the dots color-coded by the number of observations in the accumulated reference data and the vertical bars representing the standard deviation of the averaged estimates. Figure 10 averages with respect to years relative to the target country-year (so for example, 1970 is -4 with respect to Ecuador 1974). Figure 11 averages by calendar year. In Figure 10, predictions by t=0 combine some country-years early in the sample with few reference countries and those later in the sample with more country-years available. Conversely, Figure 11 presents an unbalanced panel, with country-years rotating out as targets for years after their own year (so for example US 1980 does not enter the average as a target country beyond 1980, but remains in use as a reference country). Both figures confirm the pattern that bias tends to be smaller for country-years later in the sample although Figure 10 exhibits greater variance in the estimates, owing to the inclusion of early target-years for which larger reference sizes are unavailable.

[FIGURES 10 & 11 ABOUT HERE]

Overall, Figures 6 to 11 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 validity of the IV strategy, this in turn serves as a test of the validity of our key identifying assumption of uncounfounded location.

7. Extrapolation vs. interpolation

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. In other words, is there any reason to believe that errors associated with extrapolation are systematically larger or smaller than biases associated with endogeneity? 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. In Section 8.1, we present results comparing an extrapolated ATE to target country-year OLS estimates.

The results are presented in Figure 12, where the x-axis depicts the mean of the dyadic extrapolated LATE's for a given target country-year less the within-country estimated target country-year LATE and the y-axis depicts within country-year OLS less the estimated LATE.

Both values are represented in absolute terms, reflecting that we are agnostic as to whether the extrapolated estimate is larger or smaller than the target. In addition, we divide the absolute error by the sum of the squared standard errors of each estimate. Doing so helps account for the precision of the point estimates of both the target and the reference. For example, a 5 unit difference between target and extrapolated estimates where the target standard error is 5 should not be as concerning for our technique as a 5 unit difference where the target standard error is 0.1. By dividing the absolute error by the sum of the target and reference variation, we account for the first two moments of the results, giving us a more comprehensive understanding of the efficacy of our technique.

[FIGURE 12 ABOUT HERE]

Points are coded in different shades of grey by the size of the reference sample for the target country-year. One pattern that again emerges is that, while extrapolation error can be large when the size of the reference sample is small, much of the data lies close to zero. In this range, OLS and extrapolated errors are roughly equivalent. However, as we move further away from the best cases, there is evidence in favor of the extrapolated results over OLS. Note, however, that the mean extrapolated results are much noisier than the OLS estimates, as evidenced by the wider horizontal confidence intervals as compared to the vertical lines, which are almost hidden at the scales presented.

Figures 13 and 14 use the same framework to compare the OLS estimates against the dyadic prediction error where the reference country-year is chosen to minimize geographical distance or Mahalanobis covariate distance with the target country-year. As demonstrated above, differences in the covariate profile significantly predict absolute error 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. Meanwhile, minimizing geographic distance is included as a second-best heuristic to follow if additional comparison data are unavailable. In both figures, there is clear evidence in favor of the extrapolated estimation technique over OLS. For small-sample reference sets, the extrapolated IV consistently outperforms OLS; for larger sample sizes this also appears to be the case for minimized geographic distance.

[FIGURES 13 & 14 ABOUT HERE]

Having demonstrated that extrapolated IV performs well relative to OLS using either a raw average for each target country-year or choosing the best reference country-year based on minimized Mahalanobis or geographic distance, we finally turn to a similar comparison using the cumulative results from Section 6. Given the convergence trends summarized in Section 6, 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. We calculate absolute error in the same fashion as described above and plot the cumulative error on the x-axis. On the y-axis, we re-plot the OLS errors as well as the minimized Mahalanobis and geographical distance results.

[FIGURE 15 ABOUT HERE]

As depicted in Figure 15, there is strong evidence in favor of the cumulative approach over OLS estimates and suggestive evidence in favor of cumulative results over the best dyadic extrapolated estimates.

As a final best practice, we turn to combining the minimum Mahalanobis distance technique from the dyadic results with the cumulative data, leveraging both the ability to reduce distance in the covariate profile as well as the larger reference sample sizes afforded by the cumulative results. Specifically, from each year prior to the target's year, we select the year whose reference country-years minimize Mahalanobis covariate distance with the target. As depicted in Figure 16, this approach even more strongly favors extrapolation over within-country-year OLS estimation. Almost all prediction errors are above the 45 degree line, indicating a better fit for the cumulative minimized Mahalanobis distance approach.

[FIGURE 16 ABOUT HERE]

Our results suggest that while extrapolation error remains a concern, at least for this application, the endogeneity bias of within country-year OLS is generally larger.

8. Robustness checks and extensions

8.1 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 our extrapolation procedure to estimate the target ATE from both the target data and the reference data. 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 17.

[FIGURE 17 ABOUT HERE]

Figure 17 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 10. Twenty or more years prior to the target country-year, the extrapolation tends to be noisy. 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 18 we revisit our extrapolated reference IV to target OLS comparison for this case.

[FIGURE 18 ABOUT HERE]

In particular, Figure 18 compares extrapolated reference IV average treatment effects to OLS, where the nearest geographical country is used as the reference. The results are similar to Figure 13. Most of the points lie about the 45-degree line, with a significant concentration of points at very low values of extrapolation error for IV estimates on the x-axis.

Thus our conclusions regarding the reference LATE to target LATE extrapolation also carry over to reference ATE to target ATE extrapolation.

8.2 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 19 presents cumulative extrapolation error results. The results are again similar to Figure 10. When using reference countries available twenty or more years prior to the target, estimates are noisy and tend to bounce around from year to year. But moving closer in time to the target, estimates home in on, and are not statistically significantly different from, zero extrapolation error.

[FIGURE 19 ABOUT HERE]

8.3 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 20 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 IV LATE's 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.

[FIGURE 20 ABOUT HERE]

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. This is depicted in Figure 21. 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 4, 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.

[FIGURE 21 ABOUT HERE]

[TABLE 4 ABOUT HERE]

8.4 Dropping sex-selectors

The most direct challenge to the validity of the IV assumptions in our application is the well-known 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. 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, and Nepal and Vietnam). Figure 22 presents the cumulative extrapolation error results corresponding to Figure 11. The results are qualitatively and quantitatively similar. After an initial "burn in" period where average extrapolation error bounces around from year to year, it homes in on, and is not statistically significantly different from, zero.

[FIGURE 22 ABOUT HERE]

8.5 Dropping country-years with invalid IV's

As an extension of Section 8.4, 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 8.4, 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*:

$$f(y,D=1|Z=1) \geq f(y,D=1|Z=0),\ f(y,D=0|Z=0) \geq f(y,D=0|Z=1),$$

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 constitute smoking-gun evidence that: *Z* 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.[3] 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 5 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 24, represented by light gray circles. The results are not meaningfully different from those presented above.

[TABLE 5 ABOUT HERE]

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

A final check is to 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 stratify over these covariates and 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 23 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 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.

[FIGURE 23 ABOUT HERE]

Again, our conclusions are largely robust to the omission of these country-years. Figure 24 overlays the cumulative running counter analysis from the main results with the same results calculated after dropping the invalid country-years. The convergence is still striking. The robustness-check results outperform the main analysis in the earliest counters. Although for the most stringent robustness exercise of dropping country-years that fail the stratified IV validity test, we note that it takes longer for the extrapolation results to converge to approximately zero error. This is not surprising given the reduced sample size.

[FIGURE 24 ABOUT HERE]

9. Conclusion

In this paper we have investigated the degree to which LATE's from one context can be extrapolated to another. 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 that given sufficient data, even with a small set of covariates, average extrapolation error is close to zero when extrapolating LATE's from one country-year to another. 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 covariate differences. 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, general conclusions from this stream of evidence but also 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.

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.

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.

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.

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

[TABLE A-1 ABOUT HERE]

Appendix B: Simulations

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 (C1(c)) discussed in Section 2. Combining the covariate profile with type yielded an extended covariate vector W_i which has length k and, when stacked on observations, yields matrix $W_{n\times k}$.

To generate the potential outcomes, we multiplied $W_{n\times k}$ by a k-length vector of coefficients B to yield y_1. B was drawn from a multivariate normal distribution. 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 W_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 $Z_i = 1$ conditional on X_i using the following specification (as above, let $X_{n\times k'}$ be a stacked matrix of k' covariates and n observations):

$$\Pr[Z = 1|X] = \frac{1}{1 + e^{-(XA)}}, \tag{B.1}$$

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 W.

The simulation was run 1,000 times, using the centered-interactions technique specified above to extrapolate the target estimate from the reference population. The code (in both Stata and R) is available upon request.

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)	1.92	0.64
N = 132 / 13,539	(0.56) [0.63]	(0.467)
Average Education (spouse)	2.05	0.59
N = 132 / 13,539	(0.51) [0.85]	(0.434)
Average Age (mother @ survey)	30.05	0.93
N = 139 / 15,205	(0.80) [3.49]	(0.689)
Average Age (mother @ 1st birth)	20.73	1.09
N = 139 / 15,205	(0.95) [3.00]	(0.861)
GDP per capita	9,806	10,464
N = 139 / 15,205	(9,591) [-]	(9,383)
Gender Ratio (male::female)	0.012	0.008
N = 139 / 15,205	(0.008) [0.30]	(0.008)
Total Fertility Rate (children per mother)	2.68	0.73
N = 139 / 15,205	(0.65) [-]	(0.541)
Labor Force Participation Rate	0.52	0.24
N = 125 / 15,205	(0.21) [-]	(0.17)
Year	1989	11.7
N = 139 / 15,205	(11.8) [-]	(10.2)
Geographical Distance (km)	-	7,942
N = - / 15,205	-	(4,656)
2SLS Variables		
Economically active mother (Y)	0.44	0.27
N = 125 / 15,205	(0.24) [0.44]	(0.20)
More Kids (D)	0.57	0.22
N = 139 / 15,205	(0.19) [0.46]	(0.16)
Number of Children (D)	3.05	0.59
N = 139 / 15,205	(0.52) [1.11]	(0.43)
Two children of same sex (Z)	0.51	0.008
N = 139 / 15,205	(0.008) [0.50]	(0.008)
Two Girls (Z)	0.24	0.009
N = 139 / 15,205	(0.009) [0.43]	(0.008)
Two Boys (Z)	0.26	0.009
N = 139 / 15,205	(0.009) [0.44]	(0.008)

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.

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

	I	II	III	IV	V	VI	VII	VIII	IX	X
Mother's Education (log)	.17***									
	(.03)									
Spouse's Education (log)		.20***								
		(.04)								
Mother's Age @ Survey (log)			.09***							
			(.03)							
Mother's Age @ First Birth (log)				.04						
				(.04)						
GDP pc (log)					.13***					
					(.03)					
Gender Ratio (log)						.06*				
						(.03)				
Labor Force Part. Rate (log)							.05			
							(.04)			
Total Fert. Rate (log)								.10***		
								(.04)		
Temporal Dist. (1SD = 10yrs)									.01	
									(.03)	
Geographic Dist. (1SD = 4,650km)										.05
										(.04)
Constant	-2.48***	-2.40***	-2.55***	-2.60***	-3.79***	-2.27***	-2.53***	-2.52***	-2.62***	-2.62***
	(.04)	(.05)	(.04)	(.04)	(.29)	(.20)	(.09)	(.06)	(.04)	(.04)
N	13539	13539	15205	15205	15205	15205	15205	15205	15205	15205
R^2	.03	.05	.01	.00	.02	.00	.00	.01	.00	.00

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: 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)	.02	.02	.01	.01
	(.05)	(.05)	(.05)	(.06)
Spouse's Education (log)	.18***	.18***	.18***	.18***
	(.06)	(.06)	(.06)	(.06)
Mother's Age @ Survey (log)	.06**	.06**	.06*	.05
	(.03)	(.03)	(.03)	(.03)
Mother's Age @ First Birth (log)	.02	.02	.02	.03
	(.03)	(.03)	(.03)	(.03)
GDP pc (log)	.03	.03	.03	.05*
	(.03)	(.03)	(.03)	(.03)
Gender Ratio (log)	.05*	.05	.05	.05
	(.03)	(.03)	(.03)	(.03)
Labor Force Part. Rate (log)	-.02	-.02	-.02	-.02
	(.03)	(.03)	(.03)	(.03)
Total Fert. Rate (log)	.04	.04	.04	.03
	(.03)	(.03)	(.03)	(.04)
Temporal Dist. (1SD = 10yrs)	-.01	-.01	-.01	-.03
	(.03)	(.03)	(.03)	(.04)
Geographic Dist. (1SD = 4,650km)	-.02	-.02	-.02	-.03
	(.03)	(.03)	(.03)	(.03)
Constant	-2.31***	-2.32***	-2.37***	-2.61***
	(.31)	(.31)	(.31)	(.33)
N	13539	10861	7832	4089
R^2	.06	.06	.06	.06

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 4: Shrinkage weights on measures of extrapolation error in dyadic data.

	Root Mean Squared Error (RMSE) *(1)*	*Mean Prediction Error* (2)
Weight on Prior	-0.142***	-0.031***
	(0.004)	(0.001)
Reference Size	0.04	-0.07***
(1SD = 73,160)	(0.11)	(0.023)
Extrapolated SE	3.56***	1.03***
(1SD = 6,394)	(0.11)	(0.023)
N	12,624	12,624
R^2	0.16	0.22

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.

Table 5: List of country-years with partial P-values less than 0.40.

Country	*Year*	*Partial P-value*	*St. Diff*$_0$	*St. Diff*$_1$
Egypt	1996	0.001	-0.063	-0.142
France	1990	0.011	-0.124	-0.112
France	1999	0.055	-0.121	-0.12
Uganda	2002	0.112	0.017	0.002
India	1987	0.122	-0.052	0.016
Portugal	2001	0.147	0.024	-0.203
Panama	1960	0.158	-0.157	0.046
Malaysia	1980	0.205	-0.058	0.024
Israel	1995	0.243	0.004	0.021
Malaysia	1991	0.245	-0.057	0.014
Chile	1970	0.28	0.018	-0.097
India	1993	0.329	-0.014	0.009
Greece	2001	0.365	-0.061	-0.206
India	1983	0.37	0.002	0.011
Mali	1998	0.37	-0.005	0.008
Rwanda	1991	0.375	0.019	-0.004
Guinea	1996	0.383	-0.014	0.007
Costa Rica	1973	0.389	-0.015	0.009

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 (2014).

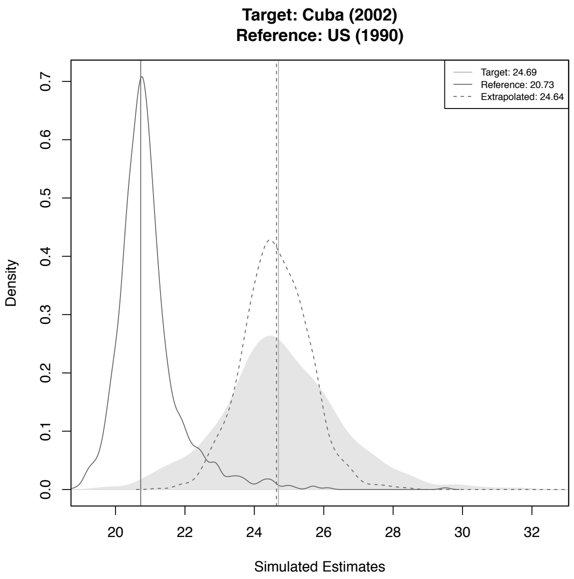

Figure 1: *Simulated results of recovering Cuba (2002) LATE using data from US (1990). 1,000 simulations run to generate distributions. The light-gray shaded polygon depicts the distribution of the simulated estimates using the Cuban (2002) data while the vertical light-gray line represents the mean estimate. The solid unshaded polygon depicts the unadjusted simulated estimates using the US (1990) data. The dashed distribution represents the extrapolated simulations using the same US (1990) data after calibrating to the complier covariate profile from the target country-year.*

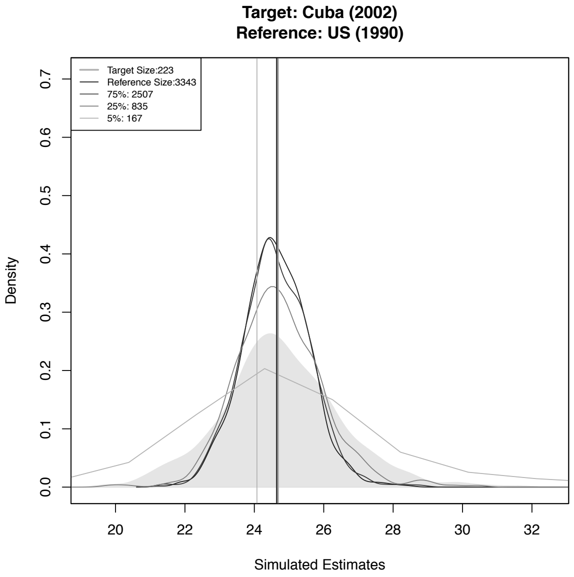

Figure 2: *Simulation results reducing size of reference population for US (1990) to extrapolate to Cuba (2002). The light-gray shaded polygon depicts the distribution of the simulated estimates using the Cuban (2002) data. The solid lines depict the extrapolated estimates calculated using the US (1990) data, dropping increasingly large fractions of the reference population at random.*

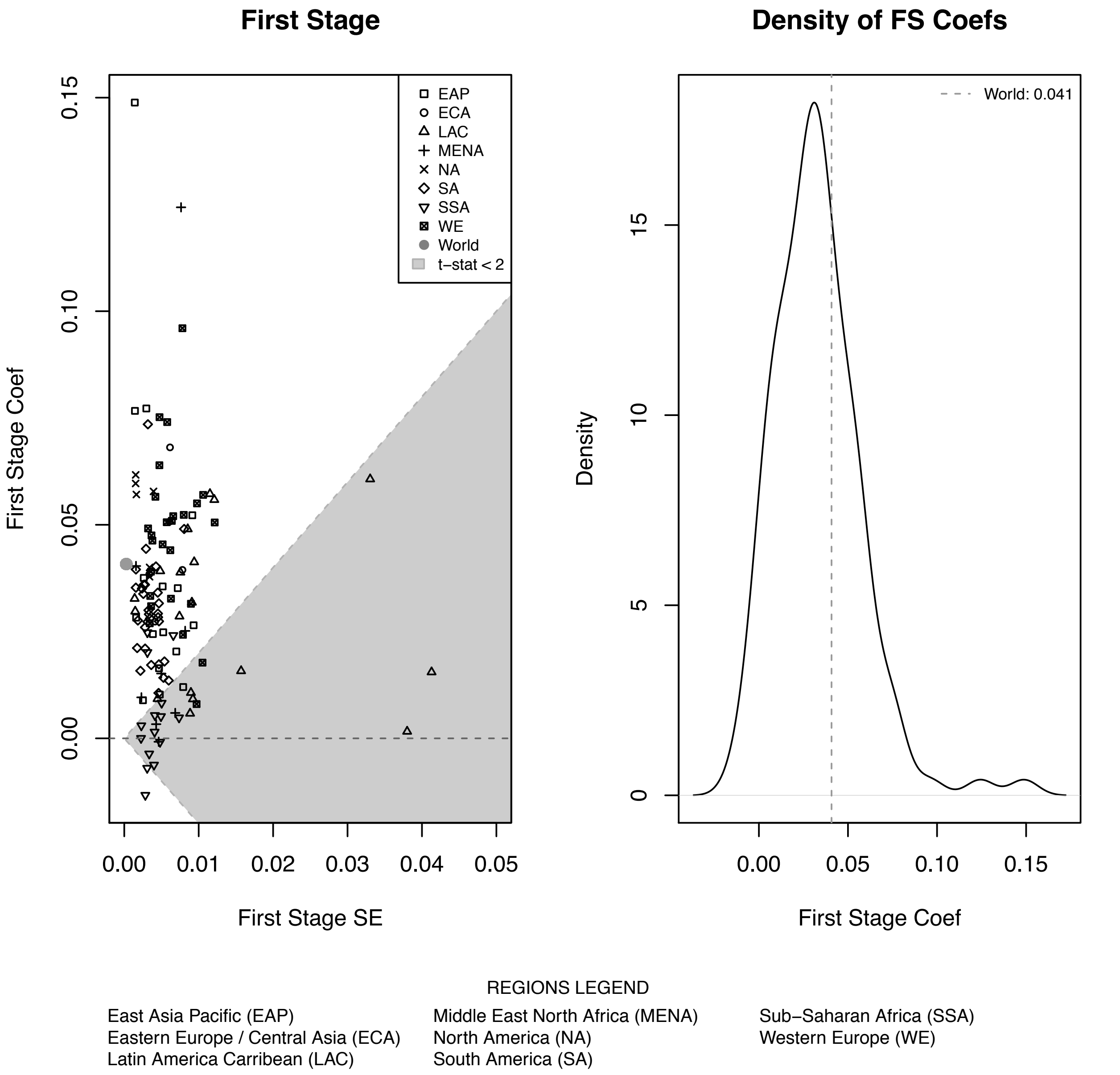

Figure 3: *Summary of first-stage results of morekids on samesex in full data. The left-panel is a scatter of all available country-years by region with the world coefficient indicated by a light gray circle. The light-gray cone covers the area in which the first-stage t-statistic is less than 2. The right-panel plots the density of the first-stage coefficients.*

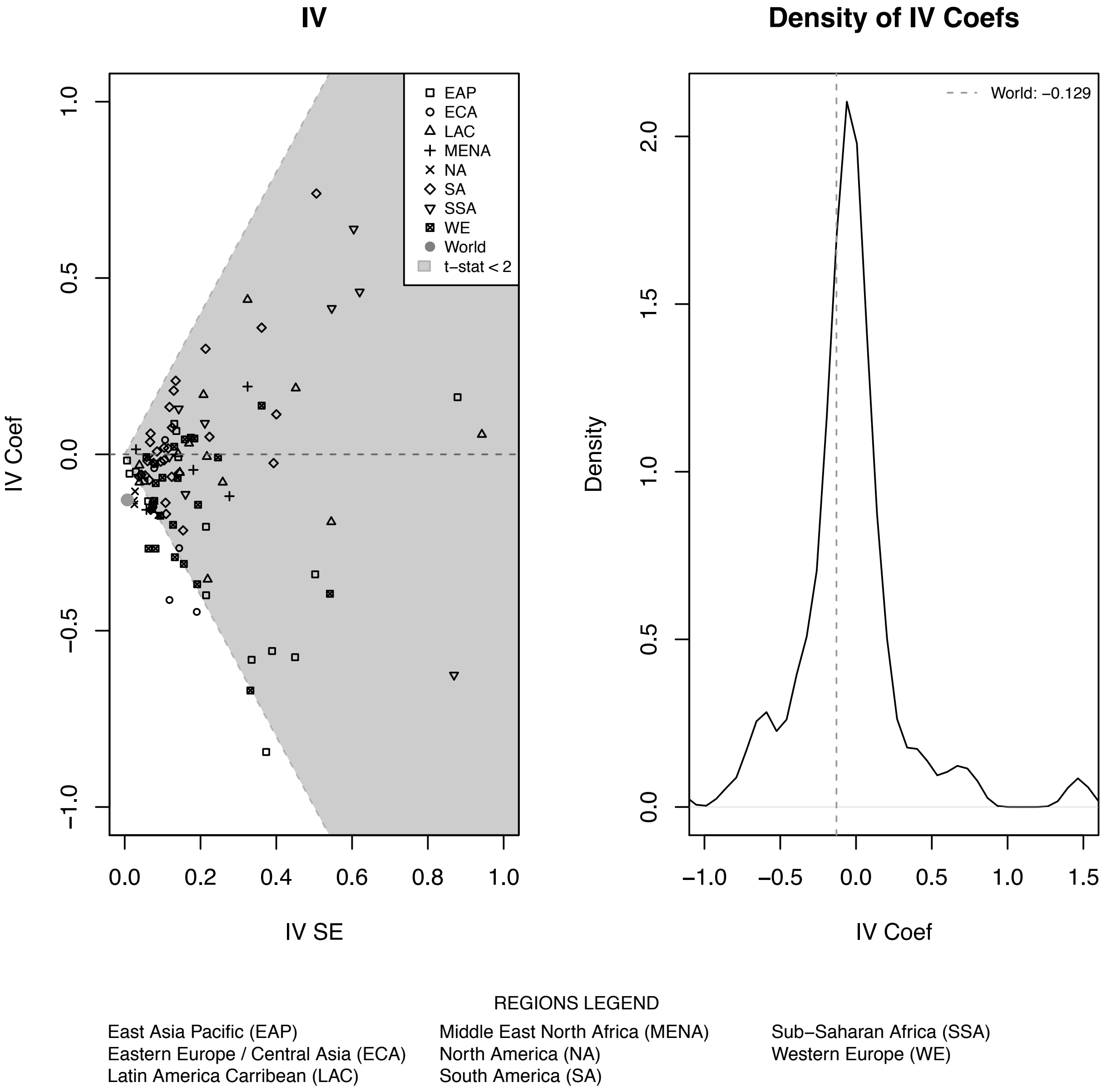

Figure 4: *Summary of 2SLS results of econactivem on morekids in full data. The left-panel is a scatter of all available country-years by region with the world coefficient indicated by a light gray circle. The light-gray cone covers the area in which the 2SLS t-statistic is less than 2. The right-panel plots the density of the 2SLS coefficients.*

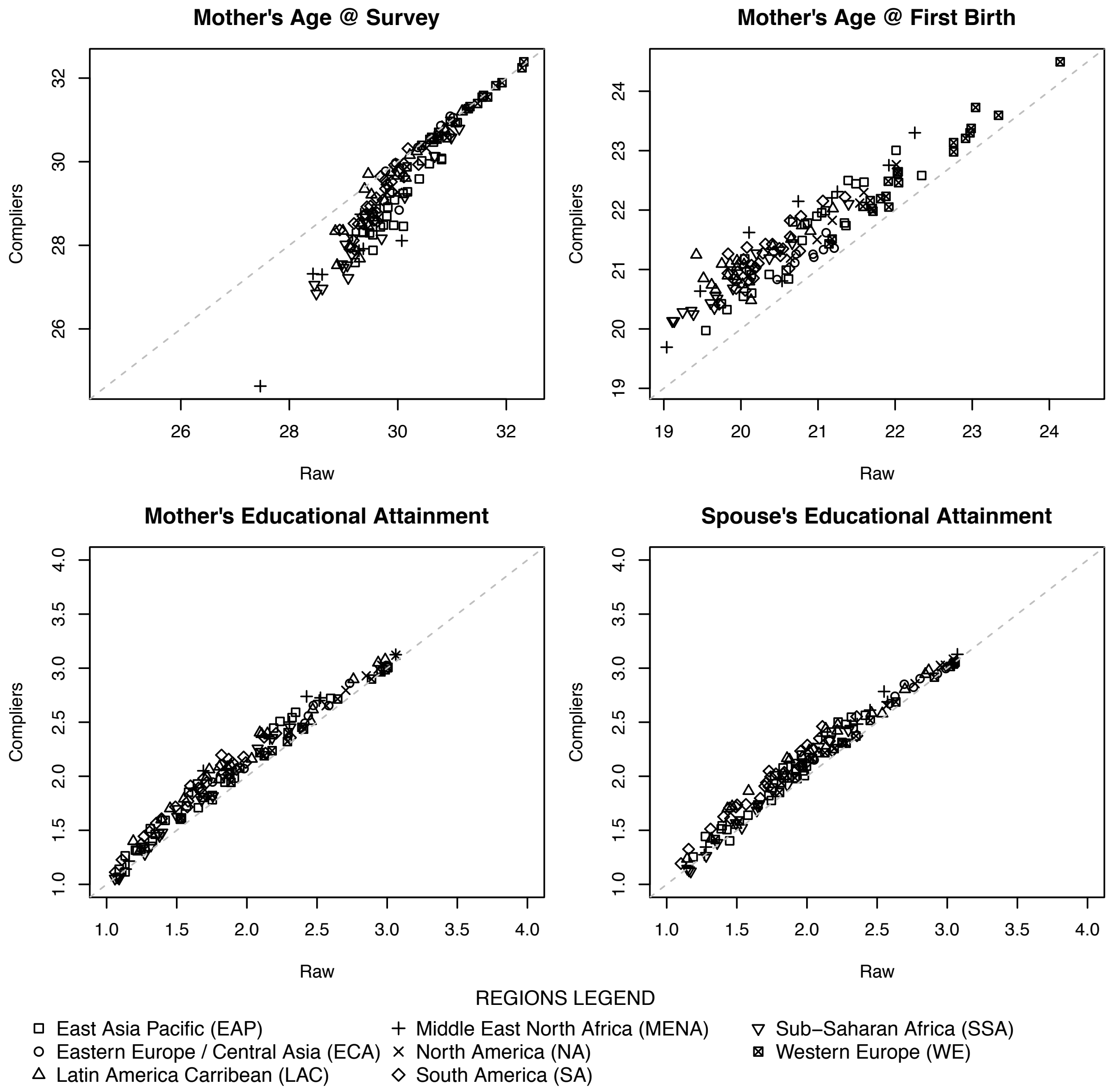

Figure 5: *Scatter of complier means (y-axis) versus raw means (x-axis) of mother's age at survey (top-left), mother's age at first birth (top-right), mother's educational attainment (bottom-left), and spouse's educational attainment (bottom-right), by region. Points falling above (below) the 45° line indicate compliers who are older (younger) or more (less) educated than the full population.*

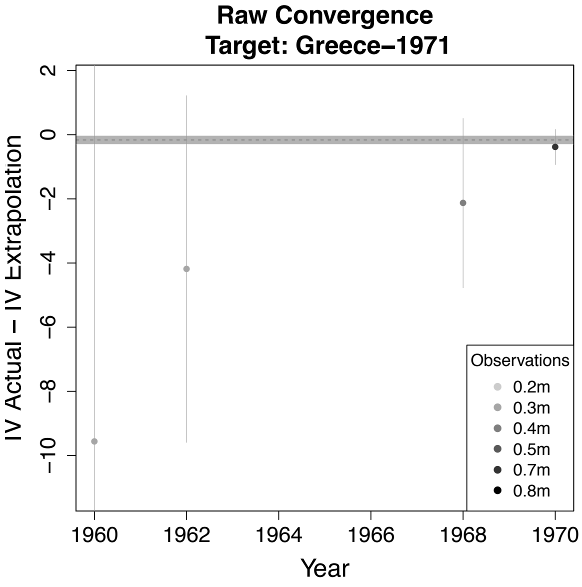

Figure 6: *Convergence of accumulated data on Greek (1971) IV estimate. The dashed line represents the target IV estimate while the shaded area charts the 2 standard error confidence intervals. The dots are shaded according to accumulated sample size while the vertical bars represent 2 standard errors.*

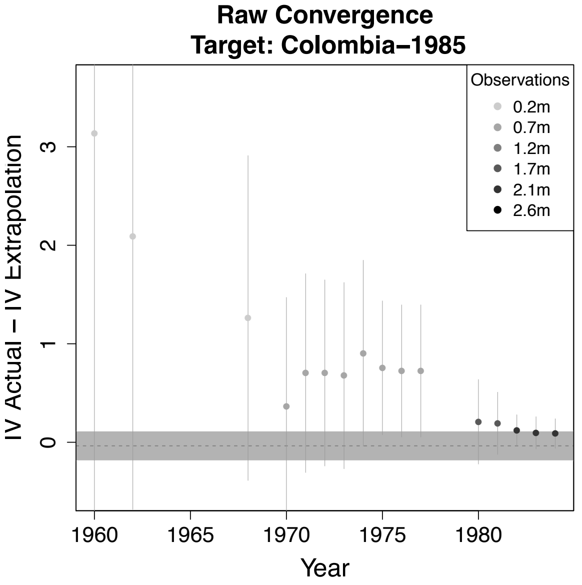

Figure 7: *Convergence of accumulated data on Colombian (1985) IV estimate. The dashed line represents the target IV estimate while the shaded area charts the 2 standard error confidence intervals. The dots are shaded according to accumulated sample size while the vertical bars represent 2 standard errors.*

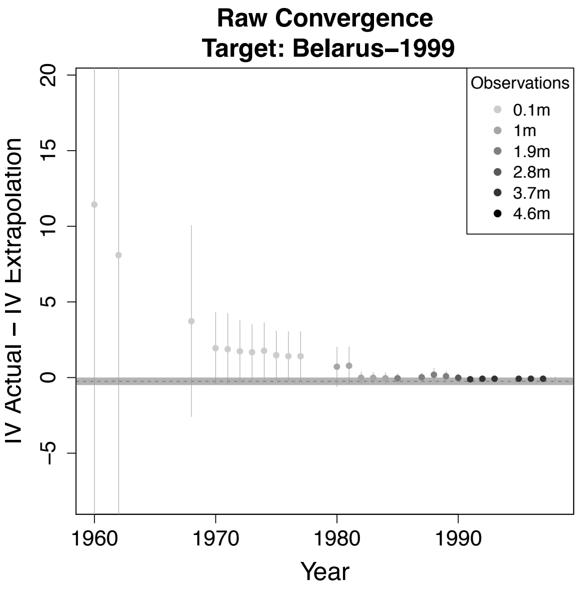

Figure 8: *Convergence of accumulated data on Belarusian (1999) IV estimate. The dashed line represents the target IV estimate while the shaded area charts the 2 standard error confidence intervals. The dots are shaded according to accumulated sample size while the vertical bars represent 2 standard errors.*

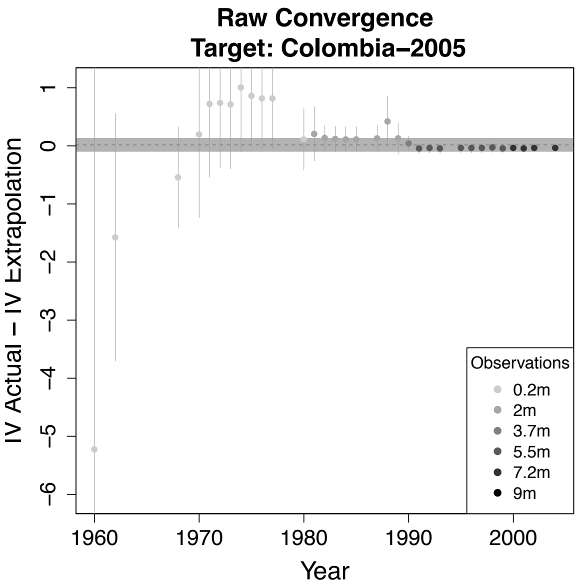

Figure 9: *Convergence of accumulated data on Colombian (2005) IV estimate. The dashed line represents the target IV estimate while the shaded area charts the 2 standard error confidence intervals. The dots are shaded according to accumulated sample size while the vertical bars represent 2 standard errors.*

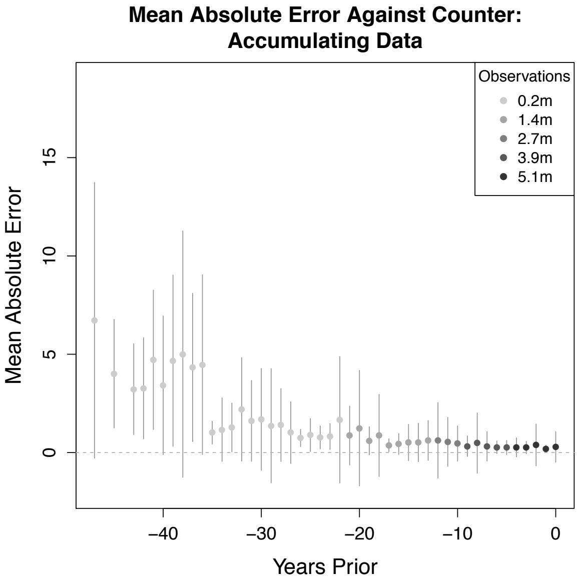

Figure 10: *Average cumulative predictions across all target country-years. Averaging with t=0 being year of observation. The dots are shaded according to accumulated sample size while the vertical bars represent 2 standard errors.*

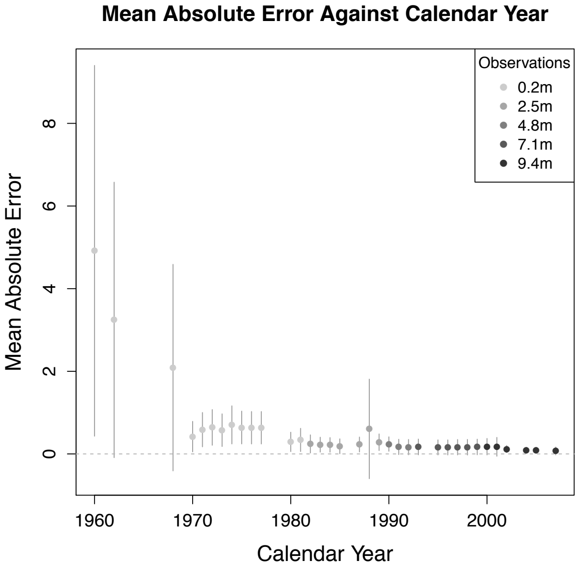

Figure 11: *Average cumulative predictions across all targets. Averaging by calendar year. The dots are shaded according to accumulated sample size while the vertical bars represent 2 standard errors.*

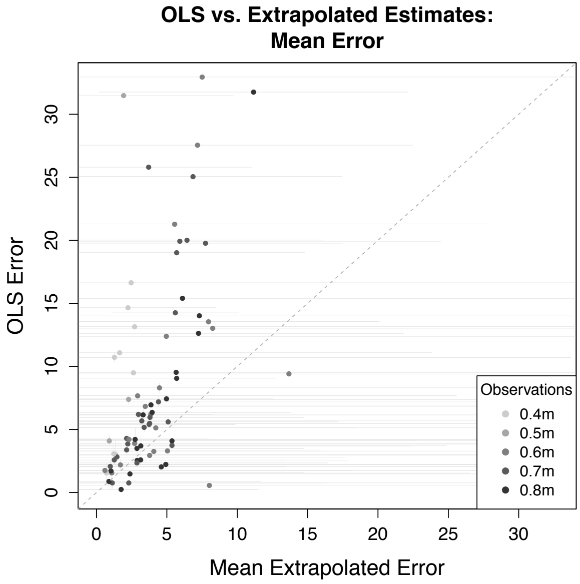

Figure 12: *Scatter plot of weighted absolute error of OLS estimate versus mean error averaged across all possible dyads for each target country-year. Weighted absolute error given by* $\frac{|\beta_{extrap.}-\beta_{target}|}{(se^2_{extrap.}+se^2_{target})}$. *Each dot represents a target country year. Two standard errors depicted by horizontal (for extrapolated error) and vertical (for OLS error) bars.*

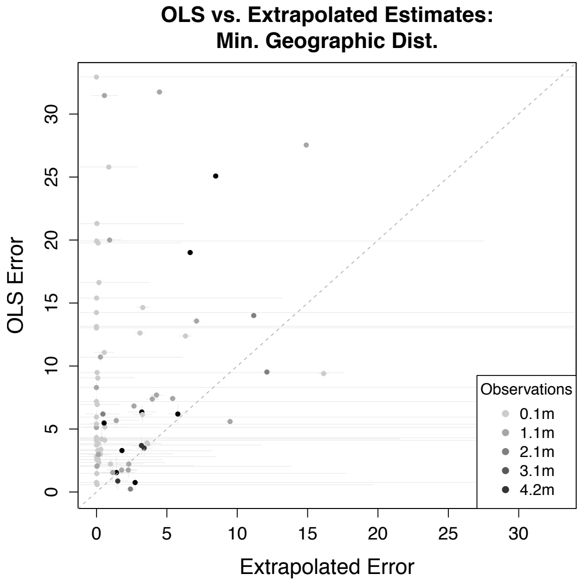

Figure 13: *Scatter plot of weighted absolute error of OLS estimate versus the extrapolated error associated with the dyad that minimizes geographical distance to the target country year. Weighted absolute error given by* $\frac{|\beta_{extrap.}-\beta_{target}|}{(se^2_{extrap.}+se^2_{target})}$. *Each dot represents a target country year. Two standard errors depicted by horizontal (for extrapolated error) and vertical (for OLS error) bars.*

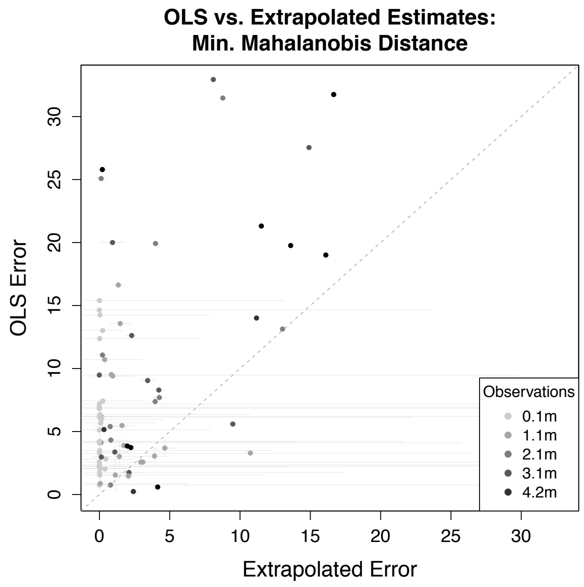

Figure 14: *Scatter plot of weighted absolute error of OLS estimate versus the extrapolated error associated with the dyad that minimizes the Mahalanobis distance to the target country year. Weighted absolute error given by* $\frac{|\beta_{extrap.}-\beta_{target}|}{(se^2_{extrap.}+se^2_{target})}$*. Mahalanobis distance 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.*

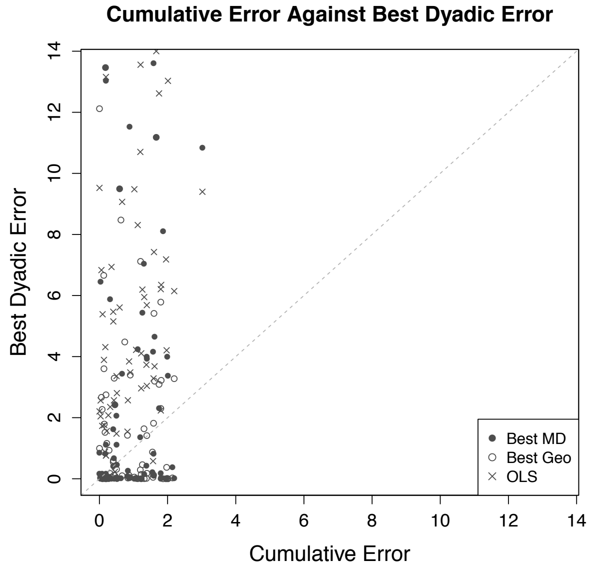

Figure 15: *Scatter plot of weighted absolute error of OLS (X's), minimized dyadic Mahalanobis distance (solid circles), and minimized geographic distance (hollow circles) versus most recent available cumulative extrapolated estimate. Weighted absolute error given by* $\frac{|\beta_{extrap.}-\beta_{target}|}{(se^2_{extrap.}+se^2_{target})}$*. Each dot represents a target country year.*

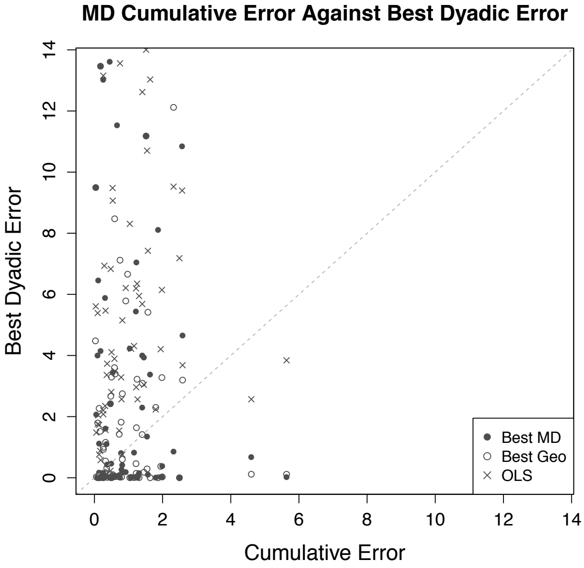

Figure 16: *Scatter plot of weighted absolute error of OLS (X's), minimized dyadic Mahalanobis distance (solid circles), and minimized geographic distance (hollow circles) versus minimized Mahalanobis distance for cumulative extrapolated estimate, where accumulated reference covariates are population weighted in household sample data. Weighted absolute error given by* $\frac{|\beta_{extrap.}-\beta_{target}|}{(se^2_{extrap.}+se^2_{target})}$. *Each dot represents a target country year.*

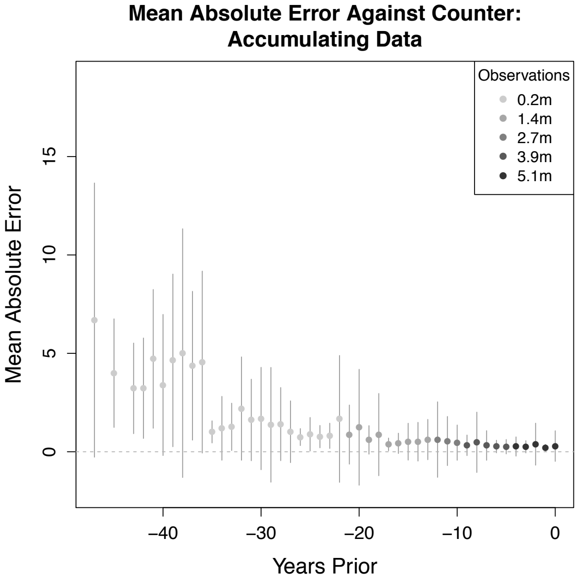

Figure 17: *Average cumulative predictions across all target country-years. Averaging with t=0 being year of observation. The dots are shaded according to accumulated sample size while the vertical bars represent 2 standard errors. Extrapolation technique reweighted to target ATE.*

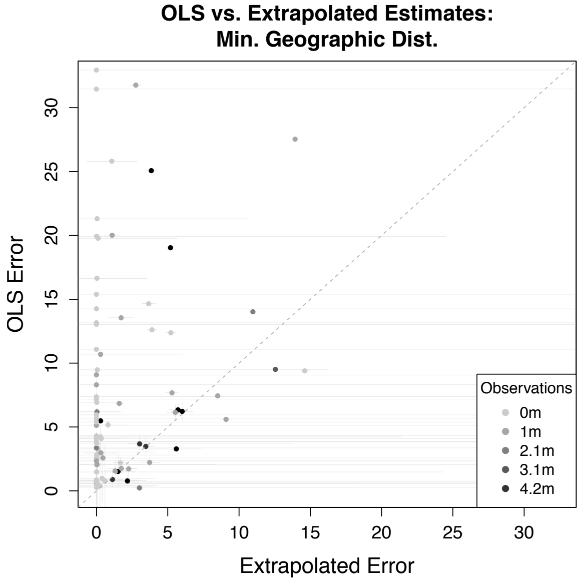

Figure 18: *Scatter plot of weighted absolute error of OLS versus minimized geographical distance dyad pair. Weighted absolute error given by* $\frac{|\beta_{extrap.}-\beta_{target}|}{(se^2_{extrap.}+se^2_{target})}$. *Each dot represents a target country year. Two standard errors depicted by horizontal (for extrapolated error) and vertical (for OLS error) bars. Extrapolation technique reweighted to target ATE.*

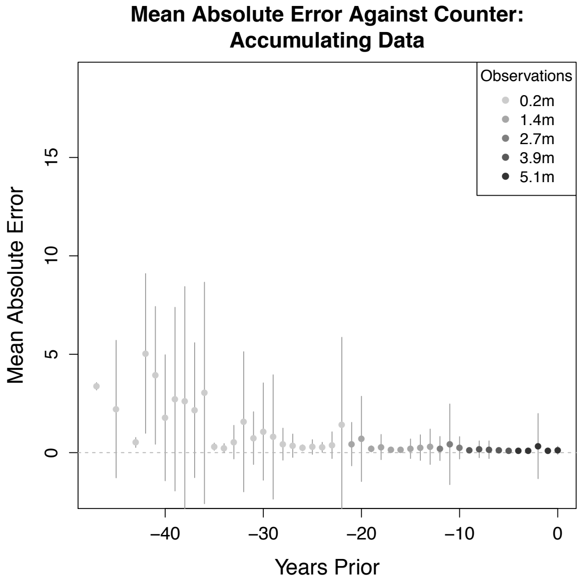

Figure 19: *Average cumulative predictions across all target country-years. Averaging with t=0 being year of observation. The dots are shaded according to accumulated sample size while the vertical bars represent 2 standard errors. Treatment measured by number of children.*

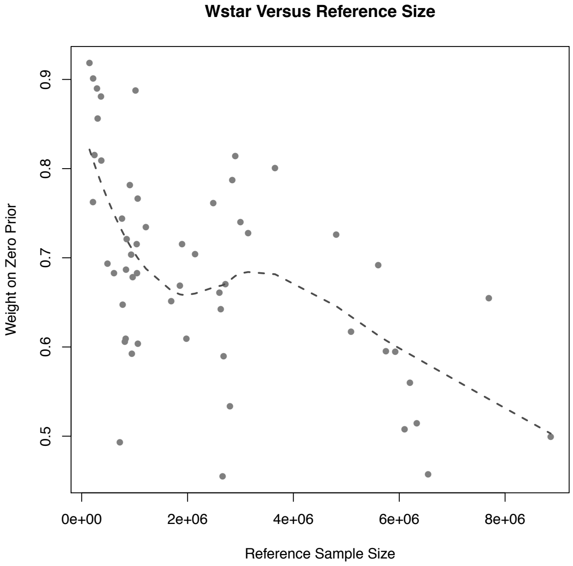

Figure 20: *Optimal weights on zero prior in full data on y-axis, calculated as binned averages over range of reference sample sizes. Dotted line represents loess smoother with α parameter of 0.9.*

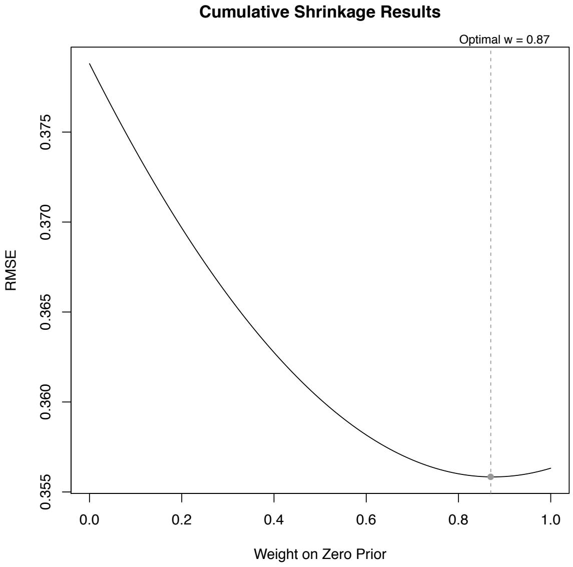

Figure 21: *Cumulative shrinkage results for all data, calculated on most recent cumulative reference population for each target country-year. Weight on zero prior (w) that minimizes RMSE for all target country-years across all targets measured on x-axis, such that higher values on x-axis reflect greater 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.*

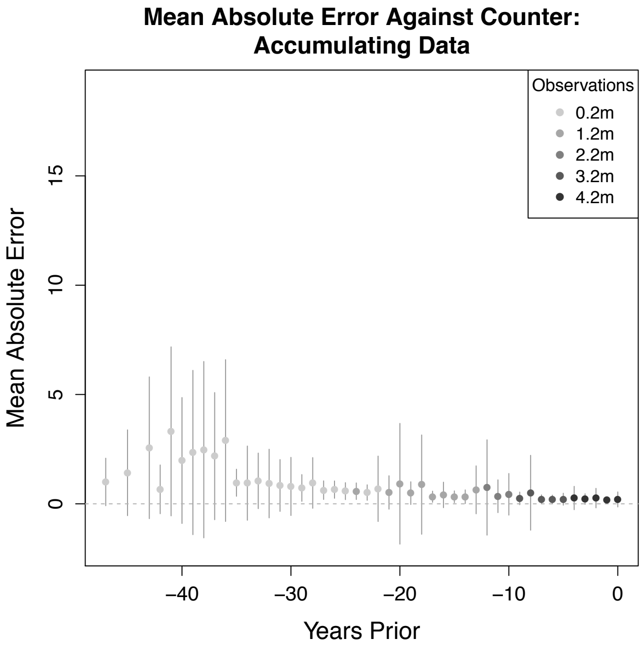

Figure 22: *Average cumulative predictions across all target country-years. Averaging with t=0 being year of observation. The dots are shaded according to accumulated sample size while the vertical bars represent 2 standard errors. Sex selectors dropped from sample include China, India, Nepal, and Vietnam.*

Figure 23 appears on the next page.

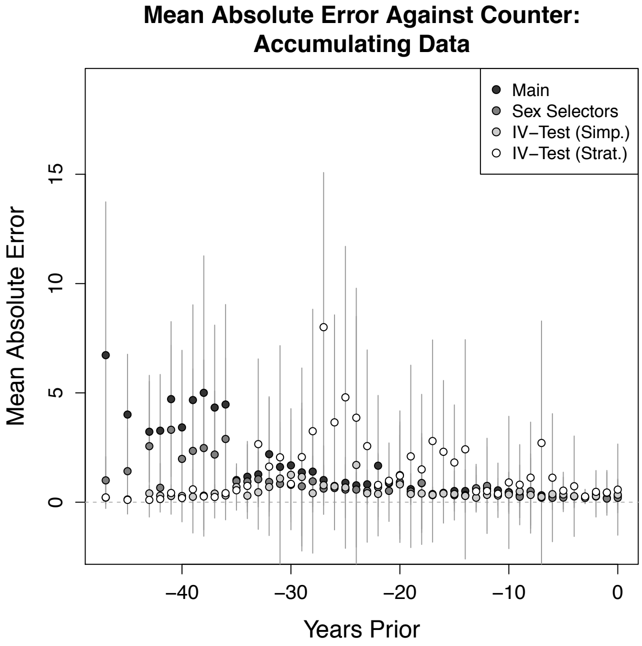

Figure 24: *Average cumulative predictions across all target country-years. Averaging with t=0 being year of observation. Dots represent different robustness checks as indicated in the legend.*

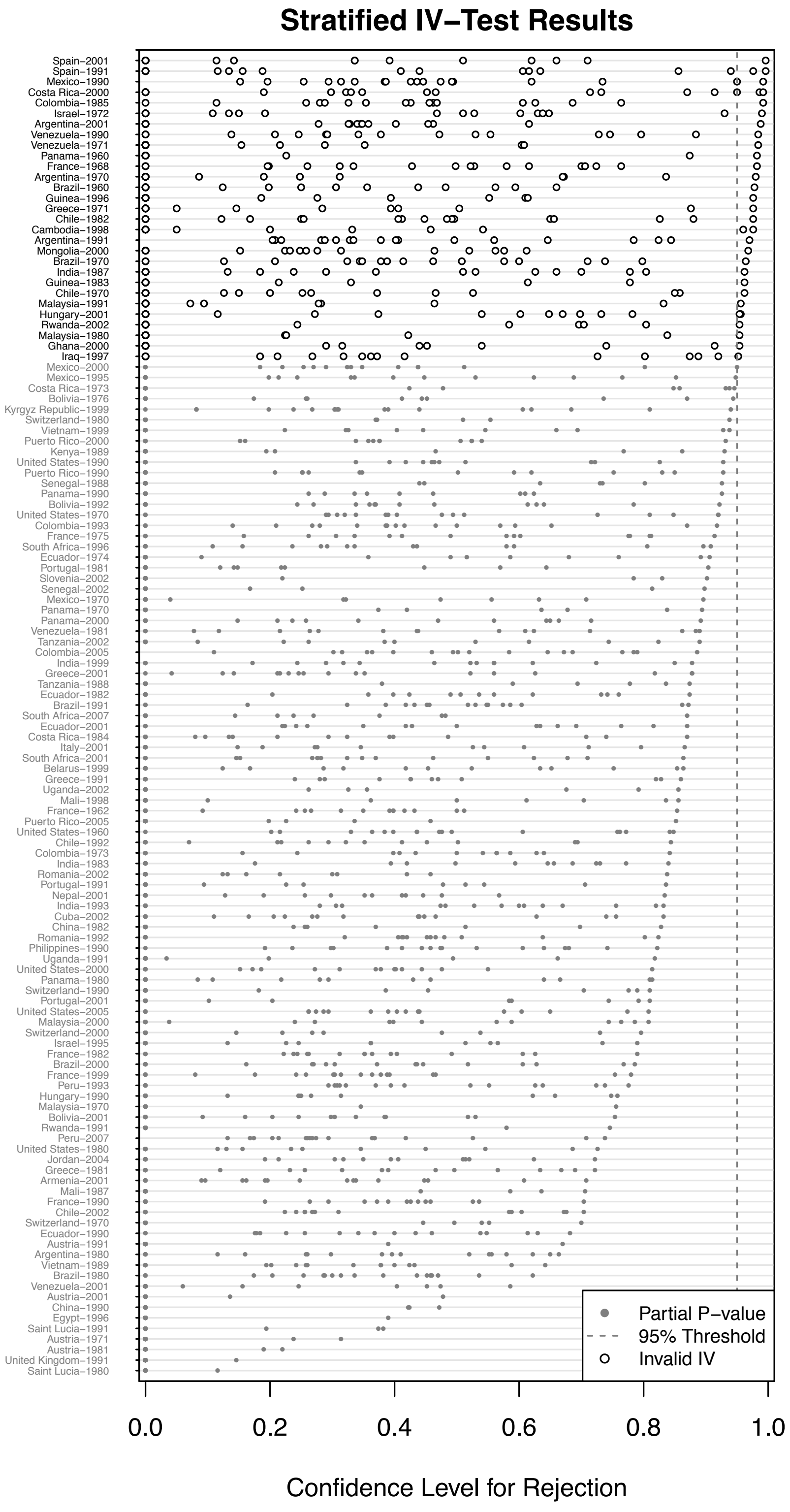

Figure 23: *Partial p-values for each of 18 possible strata, ranked by maximum confidence level for rejection of IV validity. Country-years with at least one strata failing the validity test at the 95% level of confidence (depicted by hollow-circles) are dropped.*

Table A-1: All country year statistics

Country	*Year*	*GDPpc*	*Sex Ratio*	*Educ.*	*Age*	*LFP*	*TFR*	*FS β*	*IV β*
Argentina	1970	7615	.011 (.342)	1.62 (.87)	34.8 (7.7)	.31	2.48	.043 (.005)	.023 (.093)
	1980	8487	.008 (.333)	1.65 (.63)	33.7 (7.9)	.28	2.67	.036 (.002)	-.020 (.047)
	1991	7423	.008 (.330)	2.04 (.80)	34.4 (7.5)	.51	2.59	.036 (.002)	-.118 (.046)
	2001	8552	.009 (.339)	2.30 (.75)	34.4 (7.6)	.58	2.52	.028 (.002)	-.136 (.068)
Armenia	2001	2837	.028 (.327)	3.12 (.55)	33.6 (6.4)	.76	2.24	.111 (.005)	-.080 (.045)
Austria	1971	16527	.013 (.397)	.00 (.00)	34.2 (8.2)	.60	2.08	.024 (.003)	-.323 (.144)
	1981	22437	.013 (.353)	.00 (.00)	34.8 (7.4)	.65	2.22	.042 (.004)	-.160 (.106)
	1991	27956	.015 (.357)	.00 (.00)	34.4 (6.6)	.67	2.07	.034 (.004)	-.394 (.138)
	2001	33839	.015 (.355)	.00 (.00)	35.9 (6.1)	.79	2.07	.038 (.004)	-.128 (.103)
Belarus	1999	5678	.014 (.380)	3.02 (.59)	36.1 (6.4)	.82	1.76	.022 (.002)	.057 (.106)
Bolivia	1976	3255	.006 (.330)	1.35 (1.14)	33.7 (8.6)	.22	3.06	.013 (.005)	.000 (.267)
	1992	2755	.008 (.320)	2.17 (2.03)	33.7 (8.2)	.52	3.14	.013 (.004)	.421 (.355)
	2001	3134	.014 (.338)	1.88 (1.14)	34.0 (8.4)	.49	2.84	.018 (.004)	-.058 (.201)
Brazil	1960	2469	.010 (.307)	1.11 (.66)	32.0 (7.4)	.14	3.83	.014 (.002)	.021 (.083)
	1970	3845	.009 (.304)	1.12 (.47)	32.5 (7.7)	.19	3.71	.020 (.001)	-.016 (.053)
	1980	6943	.009 (.316)	1.29 (.71)	32.7 (7.8)	.30	3.31	.027 (.001)	-.012 (.044)
	1991	6117	.010 (.328)	1.52 (.84)	33.3 (7.5)	.43	2.85	.031 (.001)	.015 (.035)
	2000	6834	.012 (.344)	1.66 (.87)	33.7 (7.6)	.58	2.37	.031 (.001)	.005 (.034)
Cambodia	1998	888	.007 (.312)	1.20 (.47)	34.2 (7.7)	.83	3.57	.018 (.003)	.055 (.122)
Chile	1970	4465	.005 (.319)	1.59 (.66)	33.5 (7.9)	.25	3.42	.020 (.004)	.103 (.131)
	1982	4308	.009 (.341)	1.85 (.72)	33.3 (7.5)	.27	2.67	.026 (.003)	.053 (.102)
	1992	6527	.010 (.351)	2.14 (.74)	33.7 (7.0)	.28	2.27	.036 (.003)	.079 (.071)
	2002	9664	.007 (.365)	2.35 (.73)	35.6 (7.0)	.43	2.09	.027 (.003)	.144 (.102)
China	1982	624	.025 (.313)	1.55 (.58)	33.8 (6.7)	.88	2.90	.066 (.001)	.009 (.011)
	1990	1157	.030 (.321)	1.65 (.59)	33.6 (6.6)	.90	2.26	.122 (.001)	.002 (.005)
Colombia	1973	4089	.006 (.310)	1.51 (1.25)	32.5 (7.8)	.31	3.77	.015 (.002)	.131 (.123)
	1985	4962	.009 (.332)	1.83 (1.23)	32.0 (7.4)	.45	2.91	.031 (.002)	-.047 (.066)
	1993	5785	.009 (.344)	1.96 (1.18)	32.9 (7.2)	.41	2.56	.034 (.002)	.036 (.052)
	2005	6491	.014 (.350)	2.13 (1.36)	34.1 (7.9)	.32	2.31	.031 (.002)	.224 (.054)
Costa Rica	1973	7067	.010 (.293)	1.45 (.69)	32.2 (7.6)	.22	4.12	.006 (.008)	-.840 (1.370)
	1984	7075	.010 (.321)	1.79 (.72)	31.5 (7.0)	.27	2.98	.049 (.007)	.044 (.112)
	2000	8870	.013 (.338)	2.11 (.80)	33.6 (7.1)	.38	2.44	.034 (.005)	.105 (.146)
Cuba	2002	7624	.014 (.442)	2.68 (.80)	33.7 (6.7)	.47	1.44	.016 (.001)	.124 (.168)
Ecuador	1974	4067	.010 (.310)	1.40 (1.02)	32.3 (8.0)	.17	3.74	.005 (.004)	1.340 (1.363)
	1982	5074	.010 (.319)	2.00 (1.88)	32.1 (7.8)	.25	3.39	.013 (.004)	.113 (.230)
	1990	4429	.012 (.326)	2.01 (1.31)	32.7 (7.6)	.33	3.03	.028 (.003)	-.004 (.109)
	2001	4824	.008 (.348)	2.04 (.92)	33.4 (7.8)	.36	2.49	.025 (.003)	-.058 (.118)
Egypt	1996	3233	.027 (.337)	.00 (.00)	31.6 (7.0)	.21	3.16	.049 (.001)	-.012 (.019)
France	1962	11116	.011 (.388)	1.56 (.65)	35.2 (7.8)	.47	2.21	.027 (.002)	-.201 (.067)
	1968	14312	.010 (.387)	1.74 (.69)	34.6 (7.7)	.55	2.24	.027 (.002)	-.216 (.069)
	1975	18472	.013 (.396)	1.91 (.81)	34.1 (8.0)	.71	2.13	.027 (.002)	-.142 (.068)
	1982	21910	.013 (.401)	1.99 (.87)	34.0 (7.5)	.79	1.93	.031 (.002)	-.189 (.057)
	1990	25766	.013 (.400)	2.24 (.94)	34.6 (6.7)	.85	1.88	.033 (.002)	-.169 (.053)
	1999	28716	.011 (.403)	2.48 (1.02)	36.1 (6.7)	.89	1.87	.030 (.001)	-.156 (.050)
Ghana	2000	1478	-.003 (.367)	1.44 (.64)	34.1 (8.3)	.85	2.67	-.003 (.003)	.514 (.735)
Greece	1971	13129	.015 (.393)	2.03 (1.53)	35.8 (7.7)	.25	1.95	.041 (.002)	-.206 (.067)
	1981	16555	.018 (.389)	2.14 (.73)	35.3 (8.1)	.33	1.86	.047 (.002)	-.055 (.055)
	1991	17768	.020 (.357)	2.43 (.72)	35.3 (7.0)	.44	2.00	.055 (.003)	-.105 (.064)
	2001	21887	.018 (.359)	2.79 (.77)	36.0 (6.1)	.50	2.01	.035 (.003)	.042 (.117)
Guinea	1983	740	.024 (.418)	1.08 (.50)	31.8 (8.9)	.52	2.21	.031 (.004)	-.127 (.140)
	1996	727	.022 (.342)	1.12 (.77)	31.7 (8.3)	.73	2.72	.013 (.004)	-.054 (.222)

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*	*Sex Ratio*	*Educ.*	*Age*	*LFP*	*TFR*	*FS β*	*IV β*
Hungary	1970	7779	.010 (.377)	1.98 (.53)	35.6 (7.7)	.00	1.88	.021 (.004)	.000 (.000)
	1980	11255	.014 (.368)	2.36 (.63)	33.0 (6.9)	.00	1.78	.038 (.004)	.000 (.000)
	1990	12489	.009 (.376)	2.59 (.60)	34.7 (6.5)	.75	1.76	.035 (.004)	-.283 (.121)
	2001	13732	.015 (.365)	2.82 (.73)	35.0 (6.6)	.62	1.88	.028 (.005)	-.240 (.190)
India	1983	1107	.039 (.352)	1.35 (.70)	30.8 (7.5)	.32	2.78	.039 (.004)	-.052 (.089)
	1987	1261	.043 (.353)	1.44 (.78)	30.7 (7.2)	.30	2.75	.045 (.004)	-.165 (.074)
	1993	1434	.047 (.361)	1.58 (.86)	30.8 (6.9)	.33	2.63	.052 (.004)	-.078 (.070)
	1999	1910	.043 (.358)	1.69 (.93)	31.1 (6.6)	.30	2.64	.050 (.004)	-.005 (.071)
Iraq	1997	2755	.010 (.284)	1.71 (.82)	31.2 (6.6)	.13	4.63	.007 (.002)	.015 (.222)
Israel	1972	13991	.020 (.330)	2.53 (1.32)	35.0 (8.3)	.32	3.07	.021 (.006)	-.025 (.240)
	1983	16123	.010 (.327)	2.90 (2.05)	33.3 (6.9)	.00	2.88	.020 (.005)	.000 (.000)
	1995	20790	.012 (.334)	2.64 (.95)	34.8 (6.9)	.59	2.78	.014 (.005)	.415 (.380)
Italy	2001	29146	.013 (.420)	2.57 (.67)	37.1 (6.9)	.66	1.72	.020 (.001)	-.035 (.090)
Jordan	2004	3947	.023 (.311)	2.61 (.84)	31.7 (6.2)	.27	4.12	.023 (.004)	-.031 (.160)
Kenya	1989	1172	.007 (.317)	1.48 (.72)	30.8 (7.9)	.76	3.55	.007 (.003)	-.387 (.408)
Kyrgyz Republic	1999	1597	.013 (.319)	3.02 (.51)	33.0 (6.7)	.78	2.89	.058 (.005)	-.009 (.062)
Malaysia	1970	2065	.008 (.306)	1.21 (.43)	32.9 (8.1)	.43	3.55	.013 (.007)	.418 (.625)
	1980	4250	.012 (.315)	1.51 (.56)	32.7 (7.4)	.53	3.28	.030 (.008)	-.034 (.257)
	1991	6272	.016 (.347)	1.87 (.84)	33.1 (7.1)	.47	2.91	.034 (.005)	-.355 (.149)
	2000	9474	.015 (.349)	2.07 (1.36)	34.5 (7.2)	.50	2.85	.032 (.005)	-.375 (.145)
Mali	1987	628	.013 (.341)	1.33 (1.42)	32.4 (8.7)	.50	3.03	.006 (.004)	.088 (.531)
	1998	768	.015 (.328)	1.28 (1.35)	32.0 (8.5)	.39	3.33	.011 (.003)	.235 (.311)
Mexico	1970	6848	.015 (.303)	1.23 (.48)	31.9 (8.0)	.17	3.94	.024 (.004)	.098 (.129)
	1990	9427	.009 (.314)	1.71 (.74)	32.5 (7.5)	.27	3.18	.028 (.001)	-.015 (.032)
	1995	9158	.008 (.360)	1.78 (.92)	31.7 (7.5)	.45	2.69	.022 (.005)	-.024 (.223)
	2000	11380	.010 (.327)	2.19 (1.64)	32.9 (7.5)	.35	2.74	.030 (.001)	.003 (.031)
Mongolia	1989	2740	.010 (.318)	2.16 (.88)	32.7 (7.8)	.00	3.38	.011 (.007)	.000 (.000)
	2000	2219	.007 (.382)	2.63 (.79)	31.9 (6.7)	.78	2.59	.028 (.005)	.073 (.161)
Nepal	2001	918	.032 (.313)	1.32 (.86)	32.3 (7.6)	.64	2.87	.033 (.002)	-.117 (.060)
Pakistan	1998	1732	.024 (.332)	1.28 (.66)	31.2 (7.8)	.00	3.66	.031 (.001)	.000 (.000)
Panama	1960	2142	.008 (.342)	1.40 (.75)	31.6 (8.0)	.38	2.98	.015 (.013)	.492 (.614)
	1970	3419	.017 (.307)	1.56 (.70)	31.8 (7.6)	.34	3.39	.028 (.008)	.044 (.277)
	1980	5200	.015 (.318)	1.84 (.89)	32.0 (7.4)	.41	3.20	.015 (.008)	-.031 (.470)
	1990	5531	.017 (.334)	2.19 (1.08)	32.3 (7.2)	.37	2.59	.048 (.007)	.101 (.143)
	2000	6950	.013 (.340)	2.27 (.91)	33.0 (7.2)	.45	2.42	.035 (.006)	.210 (.191)
Peru	1993	3855	.006 (.330)	1.97 (1.27)	33.4 (7.7)	.33	2.93	.026 (.002)	.126 (.079)
	2007	6374	.009 (.351)	2.22 (.97)	34.7 (7.7)	.41	2.39	.028 (.002)	-.006 (.073)
Philippines	1990	2334	.015 (.307)	2.34 (1.19)	32.8 (7.4)	.49	3.43	.028 (.001)	-.088 (.047)
	1995	2365	.017 (.342)	2.48 (1.04)	32.7 (7.5)	.00	3.12	.035 (.001)	.000 (.000)
	2000	2464	.020 (.356)	2.74 (1.59)	32.9 (7.4)	.00	2.99	.039 (.001)	.000 (.000)
Portugal	1981	11369	.010 (.405)	1.30 (.67)	34.3 (8.2)	.64	2.09	.028 (.003)	.179 (.153)
	1991	15661	.010 (.417)	1.62 (.86)	34.6 (7.4)	.76	1.80	.020 (.003)	.085 (.201)
	2001	20095	.011 (.432)	2.07 (.97)	35.0 (6.8)	.86	1.63	.015 (.003)	-.408 (.249)
Puerto Rico	1970	10418	.010 (.320)	2.06 (.89)	32.9 (7.9)	.00	3.05	.042 (.020)	.000 (.000)
	1980	12556	.010 (.327)	2.45 (.83)	33.1 (7.2)	.00	2.76	.033 (.008)	.000 (.000)
	1990	17870	.010 (.343)	2.78 (.75)	34.1 (7.3)	.50	2.26	.063 (.008)	-.018 (.126)
	2000	25284	.013 (.392)	3.03 (.67)	33.7 (7.7)	.56	1.89	.044 (.006)	-.233 (.160)
	2005	26054	.005 (.408)	3.12 (.68)	35.2 (8.0)	.68	1.77	.024 (.014)	-.791 (.808)

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*	*Sex Ratio*	*Educ.*	*Age*	*LFP*	*TFR*	*FS β*	*IV β*
Romania	1977	5622	.010 (.355)	1.70 (.83)	33.8 (7.4)	.00	2.19	.034 (.002)	.000 (.000)
	1992	5005	.012 (.352)	2.45 (.83)	34.0 (7.0)	.77	2.12	.035 (.002)	-.054 (.059)
	2002	6575	.012 (.365)	2.52 (.73)	34.4 (6.9)	.56	1.90	.031 (.002)	.064 (.094)
Rwanda	1991	768	-.004 (.311)	.00 (.00)	33.1 (7.4)	.97	3.98	-.001 (.004)	.603 (3.071)
	2002	732	-.004 (.332)	1.29 (.74)	33.4 (7.9)	.92	3.39	.004 (.004)	-.427 (.394)
Saint Lucia	1980	5432	.005 (.338)	1.27 (.93)	32.3 (8.6)	.53	3.90	.031 (.033)	-.166 (1.876)
	1991	9052	-.008 (.337)	1.85 (1.37)	31.4 (7.0)	.54	3.05	-.039 (.031)	-.240 (.856)
Senegal	1988	1251	-.008 (.344)	1.23 (.94)	30.4 (7.5)	.24	3.20	-.008 (.004)	.079 (.400)
	2002	1276	-.001 (.332)	1.18 (.46)	31.6 (7.7)	.33	3.25	-.009 (.004)	.333 (.387)
Slovenia	2002	20432	.015 (.368)	3.09 (1.24)	36.9 (6.0)	.92	1.94	.022 (.006)	.020 (.214)
South Africa	1996	5477	-.002 (.354)	2.21 (1.45)	33.7 (7.4)	.72	2.60	.017 (.002)	.185 (.114)
	2001	5996	-.003 (.357)	2.05 (.80)	34.5 (7.3)	.78	2.44	.018 (.002)	.181 (.105)
	2007	7442	.001 (.363)	2.30 (1.10)	34.9 (7.7)	.84	2.29	.023 (.004)	.083 (.142)
Spain	1991	20715	.014 (.344)	2.12 (.76)	36.2 (6.3)	.49	2.30	.049 (.003)	-.065 (.055)
	2001	26714	.013 (.417)	2.48 (.72)	36.5 (6.1)	.64	1.73	.026 (.001)	-.132 (.094)
Switzerland	1970	29439	.014 (.388)	3.24 (1.16)	35.0 (7.7)	.41	2.22	.021 (.005)	.156 (.245)
	1980	30010	.011 (.396)	3.20 (1.06)	35.8 (7.2)	.44	1.92	.033 (.004)	-.166 (.163)
	1990	34296	.016 (.397)	3.08 (.72)	35.6 (6.7)	.57	1.86	.036 (.004)	-.238 (.158)
	2000	35788	.010 (.351)	3.38 (1.49)	37.6 (6.0)	.64	2.19	.036 (.006)	-.054 (.171)
Tanzania	1988	687	.000 (.344)	1.35 (.55)	31.9 (8.3)	.88	3.13	.003 (.002)	.752 (.757)
	2002	790	.004 (.348)	1.68 (.57)	32.2 (8.1)	.76	2.91	.012 (.002)	-.071 (.119)
Thailand	1970	1570	.008 (.306)	1.35 (1.29)	33.0 (7.4)	.00	3.82	.012 (.004)	.000 (.000)
	1980	2413	.010 (.321)	1.14 (.63)	32.8 (7.7)	.00	3.28	.029 (.005)	.000 (.000)
	1990	4379	.009 (.351)	1.42 (.92)	33.8 (7.2)	.00	2.33	.057 (.005)	.000 (.000)
	2000	5651	.009 (.372)	1.77 (.98)	34.7 (6.7)	.00	1.97	.043 (.004)	.000 (.000)
Uganda	1991	582	.006 (.339)	1.28 (.54)	30.8 (7.9)	.68	3.09	.005 (.003)	-.439 (.501)
	2002	884	.001 (.322)	1.37 (.55)	31.2 (8.3)	.59	3.40	.002 (.002)	1.575 (2.143)
United Kingdom	1991	22766	.011 (.393)	.00 (.00)	34.6 (7.1)	.58	1.88	.050 (.004)	-.246 (.086)
United States	1960	15388	.007 (.336)	2.56 (.67)	34.7 (8.0)	.44	2.49	.033 (.002)	-.116 (.065)
	1970	20436	.009 (.340)	2.70 (.65)	35.1 (8.2)	.50	2.53	.032 (.002)	.016 (.072)
	1980	24985	.011 (.352)	2.86 (.63)	34.5 (7.6)	.63	2.17	.045 (.001)	-.106 (.023)
	1990	31452	.011 (.356)	3.01 (.60)	34.8 (6.8)	.68	2.06	.049 (.001)	-.101 (.019)
	2000	39643	.010 (.400)	3.11 (.61)	35.7 (7.6)	.73	1.86	.037 (.001)	-.099 (.020)
	2005	42482	.010 (.401)	3.21 (.61)	37.0 (7.8)	.74	1.85	.038 (.002)	-.033 (.045)
Venezuela	1971	9369	.009 (.306)	1.79 (1.82)	31.9 (7.9)	.25	3.60	.019 (.003)	.181 (.137)
	1981	9643	.012 (.355)	1.70 (.65)	30.9 (7.4)	.44	3.05	.028 (.003)	-.003 (.085)
	1990	8125	.013 (.322)	2.14 (1.87)	32.2 (7.3)	.36	2.99	.036 (.003)	-.048 (.064)
	2001	8681	.011 (.344)	2.00 (.73)	33.4 (7.4)	.41	2.40	.075 (.002)	-.034 (.029)
Vietnam	1989	855	.016 (.313)	1.71 (.82)	33.4 (7.3)	.87	3.23	.036 (.002)	-.018 (.039)

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.