FROM LOCAL TO GLOBAL: EXTERNAL VALIDITY IN A FERTILITY NATURAL EXPERIMENT

Rajeev Dehejia, Cristian Pop-Eleches, and Cyrus Samii*

May 2015

Abstract

Experimental evidence on a range of interventions in developing countries is accumulating rapidly. Is it possible to extrapolate from an experimental evidence base to other locations of policy interest (from "reference" to "target" sites)? And which factors determine the accuracy of such an extrapolation? We investigate applying the Angrist and Evans (1998) natural experiment (the effect of boy-boy or girl-girl as the first two children on incremental fertility and mothers' labor force participation) to data from International IPUMS on 166 country-year censuses. We define the external validity function with extrapolation error depending on covariate differences between reference and target locations, and find that smaller differences in geography, education, calendar year, and mothers' labor force participation lead to lower extrapolation error. As experimental evidence accumulates, out-of-sample extrapolation error does not systematically approach zero if the available evidence base is naïvely extrapolated, but does if the external validity function is used to select the most appropriate reference context for a given target (although absolute error remains meaningful relative to the magnitude of the treatment effect). We also investigate where to locate experiments and the decision problem associated with extrapolating from existing evidence rather than running a new experiment at a target site.

*Dehejia, Wagner Graduate School of Public Service, 295 Lafayette Street, New York, NY 10012 (Email: rajeev@dehejia.net). Pop-Eleches, School of International and Public Affairs, Columbia University, 420 W 118th Street, New York, NY 10027, (Email: cp2124@columbia.edu). Samii, Department of Politics, New York University, 19 West 4th Street, New York, NY 10012 (Email: cds2083@nyu.edu). The authors thank: Morris Chow for excellent research assistance; Hunt Allcott, Joshua Angrist, Peter Aronow, Gary Chamberlain, Drew Dimmery, and Raimundo Undurraga for valuable comments and suggestions; and seminar participants at NEUDC 2014, NYU, Yale, the 2014 Stata Texas Empirical Microeconomics Conference, and the Stanford 2015 SITE conference for helpful feedback.

1. Introduction

The use of randomized controlled trials in economics is widespread. Indeed, in the field of development economics, it can be called a global phenomenon, with hundreds of experiments being run around the world. The initial emphasis on experiments was driven by the ability of randomized controlled trials to generate internally valid results, and by the concomitant failure of non-experimental methods to deliver a similar promise (e.g., Lalonde 1986). While a clean and careful experiment is never to be taken for granted, with the increased expertise and experience within economics in implementing large-scale field experiments, internally valid results can reasonably be viewed as an attainable benchmark.

At the same time, the global scale of field experiments points to the less-emphasized but central concern of external validity. In evaluating the external validity of a set of experiments, one poses the question, "to what population, settings, and variables can this effect be generalized?" (Campbell 1957). In other words, external validity can be measured in terms of the error in prediction of treatment effects for new populations beyond those covered in the evidence base. With a single or handful of experiments, external validity is a matter of assumption. But with a large number of experiments it is reasonable to hope that researchers are accumulating knowledge, i.e., not just learning about the specific time and place in which an experiment was run but learning enough to predict what would happen if a similar intervention were implemented in another time or place. One could judge the success of an experimental research program in terms of the diversity of settings in which one can reliably predict the treatment effect, possibly obviating the need for further experimentation with that particular treatment. This is the

issue we address in this paper. More specifically, assuming consistent experimental or internally valid evidence is available across a variety of settings, is it possible to predict the treatment effect in a new setting? Is it possible to understand how differences between actual and predicted treatment effects vary with differences between the setting of interest and the settings in which experimental evidence is available? And if so which differences are more important: context-level (e.g., macro or institutional) variables or individual-level variables? How might we judge whether an existing evidence base is adequate for informing new policies, thereby making another experiment with a given treatment unnecessary?

Our approach in this paper is to use a natural experiment for which data is, in fact, available for a wide variety of settings. In particular, we use the Angrist and Evans (1998) sex-composition variable (same sex of the first two children) as a natural experiment for incremental fertility (having a third child) and for mother's labor supply and in the context of the Integrated Public Use Microdata Series - International (IPUMS-I) data. Cruces and Galiani (2005) and Ebenstein (2009) have studied how the effects in this natural experiment generalize to Argentina and to Mexico and Taiwan, respectively. Our analysis extends this to all available IPUMS-I samples around the world going back to 1960, allowing for a very rich examination of both micro- and macro-level sources of heterogeneity.

We discuss the strengths and weaknesses of this data in greater detail Section 4. But, briefly, it is important to acknowledge that *Same-Sex* is not a perfect experiment. At the same time, we would argue that, since even the best field experiments have their flaws, it is a not an unreasonable exercise in external validity. Furthermore, to the extent

that fertility choices could be viewed as especially culture and context specific, we believe we are setting a high bar for the exercise: if we are able to find a degree of external validity for a fertility natural experiment, then there is hope that it might be possible for other experiments as well.

The paper is both a methodological thought experiment and an empirical investigation. As a thought experiment, we consider the rather fanciful situation of having replications of a randomized experiment across a wide variety of contexts that we can use to inform an extrapolation to an external setting. This is an idealized setting in certain respects, given the large number of experiments and also the homogeneity in treatments and outcomes. At the same time, and in a manner that brings us back down to earth, we only have a limited amount of information that we can use to characterize effect heterogeneity. Given such data, what statistical tools would we use to construct extrapolations to new contexts? Our task in this paper is one of assessing the external validity potential of an evidence base in *extrapolating* to a new context. An evidence base consists of a set of experimental studies and its limitations are defined by the variety of contexts that it covers and, crucially, the measured covariates that it includes. A complementary exercise, which we do not undertake in this paper, might be to use an evidence base to try to explain effect heterogeneity for the sake of theory development.

We operate under the premise that working through this problem in the setting granted by our "global" experiments should prove useful for informing how experimental research programs should proceed. Substantively, we are interested to understand how fertility decisions and associated labor outcomes vary beyond the well-studied United

States context (Angrist and Evans 1998). If one were to design fertility or labor policies outside the US, what kind of evidence should one use and how?

The topic of external validity has been gathering increasing attention in the economics literature. Empirical assessments of external validity in economics include recent work by Allcott (2014), Pritchet and Sandefur (2013), and Vivalt (2014). Allcott (2014) tackles the question of site selection, and in particular whether the sites that select into an experiment can limit the ability to draw externally valid conclusions from internally valid experiments. He finds evidence that sites that opt in early into an experiment may be those more likely to benefit. We sidestep this question in our analysis. While we do not directly address site selection into the IPUMS-I data, we will examine how external validity evolves over time, where our emphasis is on the accumulation of evidence from experimental data points.

Using two examples from the education literature (class size effects and the gains from private schooling), Pritchett and Sandefur (2013) argue that estimates from observational (that is, non-experimental) studies within a context are superior to extrapolated experimental results from other contexts. They also argue that economy-wide or institutional characteristics often trump the importance of individual characteristics when attempting to extrapolate. We view our efforts as complementary. We do not directly examine the bias tradeoff that is central to Pritchett and Sandefur (essentially the tradeoff between bias from failure of internal validity and bias from the failure of external validity). But with a large number of (natural) experiments in our data set (166 compared to the dozen or so studies they use in their analysis) we are able to take an empirical approach to some of their central questions.

Vivalt (2014) uses a random effects meta-analysis to study sources of effect heterogeneity and extrapolation error for sets of development program impact evaluations. She finds evidence of program effects varying by the implementing actor, with government programs tending to fare worse than non-governmental organization programs. She also finds that with a small set of study-level characteristics (namely, implementer, region, intervention type, and outcome type), meta-regressions have only modest predictive power. In our analysis, we consider a somewhat larger number of covariates both at the micro- and macro-levels and we do so in a set of experiments that is more homogenous in terms of treatments and outcomes. This allows us to isolate issues of extrapolation per se from questions of outcome and treatment comparability.

Our results show that there is considerable treatment effect heterogeneity in the effect of sex composition on fertility and labor supply across country-years, but that some of this variation can be meaningfully explained both by individual and context covariates. We define and estimate the "external validity function," which characterizes how results from existing experiments may yield misleading predictions for a target setting, and show that in our application the estimated external validity function implies that increasing the covariate distance between experimental sites and a site at which one is trying to predict the treatment effect leads to increased prediction error. We examine the multivariate relationship between prediction error and individual and context covariates, and argue that both are potentially useful in reducing prediction error from external comparisons, although also discuss a case in which only context variables seem to matter. We also investigate the out of sample performance of prediction error models in selecting the best

comparison country for a target country of interest, and find that with enough experimental data points the quality of extrapolation increases considerably.

Finally, we present two applications of our approach. In the first, we use our estimated external validity function to determine the best location of an experiment. Specifically, choosing among our 166 country-year sites, we ask which location would minimize mean squared prediction error for the other sites? And given that choice of first site, which is the second-best site to add to the experimental sites, and so on. The thought experiment is to ask, given our full set of sites, what are the attributes of a good location for an experiment? In the second application, we ask when a policy decision maker should choose to run an experiment in a target context rather than use extrapolated estimates of the treatment effect from other sites.

The paper is organized as follows. In Section 2, we provide a brief review of the related literature, while in Section 3 we outline a simple analytic framework for our empirical analysis. In Section 4, we discuss our data and the sex composition natural experiment. In Section 5 we present a graphical analysis of treatment effect heterogeneity, and in Section 6 we perform the analogous hypothesis tests to reject homogenous treatment effects. In Section 7, we present non-parametric estimates of the external validity function for selected covariates of interest. In Section 8, we use multivariate regressions to examine the relative importance of individual and context-level predictors in determining the external validity of experimental evidence. In Section 9, we present evidence on the in-sample fit and out-of-sample predictive accuracy of the model, and in particular examine how external validity evolves with the accumulation of

evidence. Section 10 presents our two applications, the choice of experimental site and of whether or not to run an experiment to inform a policy decision. Section 11 concludes.

2. Related methodological literature

Our analysis follows on the call by Imbens (2010) to scrutinize empirically (rather than speculatively) questions of external validity. Focused consideration of external validity goes back at least to Campbell (1957), whose approach is taken up in the text by Shadish et al. (2002). This classical literature omits a formal statement of how external validity may be achieved, and so debates in the classical literature often confuse semantic issues with conceptual ones. More recently, Hotz et al. (2005), Stuart et al. (2011), and Hartman et al. (2013) use the potential outcomes framework to characterize conditions necessary for extrapolation from a reference population for which experiments are available to a target population. These conditions are analogous to those required for identifying causal effects under "strong ignorability." The difference is that the relevant conditional independence assumptions pertain to inclusion in the reference versus target population rather than in the treatment versus control group. We review these conditions in the next section. These authors apply various approaches to extrapolation, including matching, inverse probability weighting, and regression (see also Cole and Stuart 2010, on inverse probability weighting, and Imai and Ratkovic 2013, and Green and Kern 2012, on response surface modeling). Crump et al. (2008) develop non-parametric methods, including sieve estimators, for characterizing effect heterogeneity. Our analysis combines these various methods. Angrist (2004), Angrist and Fernandez-Val (2010), and Aronow and Sovey (2013) consider extrapolation from local average treatment effects identified

by instrumental variables to a target population. This is an issue we avoid, as we focus only on reduced form or intention-to-treat effects.

While conditions for identifying extrapolated effects are straightforward to express, the implications warrant some deeper reflection. Heckman and Vytlacil (2007) and Pearl and Bareinboim (2014) discuss how these identifying assumptions require that structural relations between background characteristics and treatment effects are invariant as we move from the reference to the target population. Pearl and Bareinboim give a useful example of a case where there is only one variable that moderates effects, but it is measured via proxy. If the distribution of the proxy measure differs across the reference and target contexts, then without additional invariance assumptions, we cannot know the consequences for identification. The difference could be because the underlying moderator distribution differs or because the relationship between the moderator and the proxy differs. By the same token, just because the proxy variables have the same distribution across contexts does not imply that contexts are comparable unless all relevant structural relations are invariant across context. Bareinboim and Pearl (2013) demonstrate how a set of reference experiments, each on its own inadequate for identifying an effect in a target setting, might be combined to identify the target effect.

Our analysis is also related to the meta-analysis literature (Glass, 1976; Hedges and Olkin, 1985; Sutton and Higgins, 2008). Applications in economics include Card et al. (2010), Dehejia (2003), and Stanley (2001), as well as meta-analytic reviews that appear in the *Journal of Economic Surveys*. What the meta-analysis literature lacks, however, is a general (i.e., non-parametric) characterization of the conditions required for extrapolating from reference to target contexts. Classical approaches to meta-analysis

use meta-regression to determine correlates of effect heterogeneity---so called "moderator" analysis. The classical literature tends to leave unclear the purpose of such moderator analysis with some discussions suggesting that it is merely descriptive, with no claim of identifying an effect in a target population, and others suggesting the much more ambitious goal of trying to establish a full generative model of the conditional effect distribution (Greenland 1994; Rubin 1992). The work on non-parametric identification of extrapolated effects, which we use as the foundation of our analysis, is much clearer about the role of moderator analyses.[1]

3. Analytical framework

We are interested in using the results of existing randomized experiments to inform our expectations of what might happen in a new, external context. This is an issue of external validity. Following Hotz et al. (2005) suppose, formally, that we are interested in the effects of a treatment, $T = 0,1$. Define potential outcomes associated with this treatment as $Y(1)$ and $Y(0)$. Let $D=0,1$ denote locations from which we have experimental data ($D=0$, the "reference" context) and to which we want to extrapolate ($D=1$, the "target" context), respectively. By virtue of random assignment, experiments in the $D = 0$ contexts are such that

$$\text{(C0)} \quad T \perp\!\!\!\perp (Y(0), Y(1)) | D = 0.$$

[1] We sidestep altogether a few themes that are central to the classical meta-analysis literature, including the construction of standardized effect size metrics and evaluating publication bias or other "file drawer" problems. See Slavin (1984, 1986) for a trenchant critique of how these methods have been applied in practice. Related to the issue of publication bias are the issues of site selection and nonrandom study recruitment, which bias the distribution of effects that are available to synthesize relative to the underlying potential distribution of effects in a population (Olsen et al. 2012).

Define W as covariates necessary to satisfy an unconfounded location condition, that is

(C1) $D \perp\!\!\!\perp (Y(0), Y(1)) | W.$

We also define a common support condition,

(C2) $\delta < \Pr(D = 0 | W = w) < 1 - \delta,$

for $\delta > 0$ and for all w in the support of W over units in the D=1 target population.

Conditions C1 and C2 imply that data on effects in context D=0 are sufficient to identify effects in context D=1 (Hotz et al. 2005, Lemma 1).[2] That is, observed outcomes are given by

(1) $Y = TY(1) + (1 - T)\ Y(0),$

and by C0-C2 we have

(2a) $E[Y(1) - Y(0) | D=1] = E_{W|D=1}[E[Y(1) - Y(0) | D=0, W]]$

(2b) $= E_{W|D=1}[E[Y|T=1, D=0, W] - E[Y|T=0, D=0, W]]$

[2] In cases where random assignment is conditional (e.g., in situations resembling stratified random assignment or where assignment probabilities vary with some covariates), the situation is nearly identical —the only difference being that we need to incorporate the relevant covariates into the analysis.

where $E_{W|D=d}[.]$ denotes marginalizing over the *W* distribution in context *D=d*. To extrapolate from one context to a new context, we compute the treatment effects for each value of *W* that appears in the new context and then marginalize over the distribution of *W* in the new context. Consistent extrapolation requires that a third condition holds:

(C3) There exists a consistent estimator for the covariate-specific effects as defined in 2b.

In typical applied settings, one assumes that conditions C0-C3 hold. What our data allow is a *test* of some of these conditions. We assume that gender composition of the first two children provides a valid natural experiment for having a third child, satisfying C0. As we discuss in the next section, there are a variety of reasons why this assumption might not hold. We think of the leading reasons for such a violation (e.g., sex selection or women's labor force participation) as being context-specific covariates (elements of *W*). We use highly flexible estimation methods such that condition C3 is satisfied. Our covariate set is sufficiently parsimonious that C2 is also typically satisfied. As such, we are primarily concerned with testing C1, unconfounded location.

To carry out such tests, we define an "external validity function," $\epsilon(W)$, which takes expression (2b) and subtracts from it the target effect estimate in (2a):

$$(3)\quad \epsilon(W) = E_{W|D=1}\{E[Y|T=1, D=0, W] - E[Y|T=0, D=0, W]\} - E[Y(1) - Y(0)|D=1] = \hat{\tau}(W) - \tau,$$

where the first term on the right-hand side of (3) is the extrapolated effect from the D=0 context(s) and the second is the true treatment effect in the D=1 context. The external validity function is analogous to the bias function defined by Heckman et al. (1998), with the latter defined as the difference between the unobserved conditional control mean for treated units and the conditional mean for untreated units. In applied settings $\epsilon(W)$ is only a hypothetical quantity since one does not know the true effect, τ. The empirical exercise that we carry out is one where we actually *have* estimates of τ for the various country-years in the data-set. We can then use these as benchmarks to assess extrapolations from other country-year contexts.

Under C0-C3, $\epsilon(W) = 0$. But in our empirical analysis, estimates of this quantity will tend to be non-zero. Such "prediction error" is due to the combination of random variation and bias resulting from failures of C0-C3. Random variation arises in our $\hat{\tau}(W)$ estimates as a result of sampling of units within the reference contexts and then treatment assignment in those contexts. Random variation arises in our estimates of τ as a result of sampling of units in the target contexts and treatment assignment.[3]

Much of our empirical analysis below focuses on the prediction error captured by estimates of $\epsilon(W)$.[4] We conduct both dyadic and cumulative analyses. In the dyadic analysis, we pair each country-year in our sample to each other country-year, creating dyads consisting of target country and comparison country. Using a flexible linear

[3] Our data satisfy random sampling of units, condition C0 for the reference contexts, and a similar random assignment condition in our target context. We work with linear least squares estimators for $\hat{\tau}(W)$ and τ. Thus, conditional on W, by standard arguments, our estimates of $\hat{\tau}(W)$ and τ are statistically independent and asymptotically normal (e.g., Freedman 2008), in which case our estimate of $\epsilon(W)$ is also asymptotically normal.

[4] Our use of the term "prediction error" connotes a presumption that our models of conditional treatment effects will be, inevitably, approximate due to imperfect knowledge of the conditional effect distribution as well as limitations in the covariate set that is available.

regression in the comparison country, we predict the treatment effect in the target country (the first expression on the right-hand side of (3)), which we then compare to the treatment effect from the (natural) experiment in the target country (the second expression on the right-hand side of (3)). Note that dyadic predictions between a given target and reference country-year pair are not symmetric, since prediction error will be different depending on which country-year is used as the reference and which as the target. For example, predicting the treatment effect in Cuba using US data is a very different exercise than the reverse. Each dyad gives us a vector of covariate differences between target and comparison country, and an estimated prediction error from the comparison. The cumulative analysis examines how prediction error changes as we proceed forward in time, using all data from countries in previous years to predict the treatment effect in a country in a given year.

Our analysis of prediction error begins by presenting local linear regressions of prediction error against a single dimension of the covariate difference between target and comparison country-years. We refer to this as the unconditional prediction error estimates; since we do not control for other covariate differences, prediction error associated with differences in a single variable (for example education) could be driven by any other correlated variable (for example GDP per capita). We then use these dyadic differences in a multivariate regression of prediction error on the covariate differences between the target and comparison country. We refer to these as conditional prediction error estimates, since we are able to examine the effect of the target-comparison difference in a covariate of interest on prediction error, conditional on all other target-comparison covariate differences. We also use this multivariate regression approximation

of the external validity function to select the error-minimizing comparison for a target country of interest when we conduct our out-of-sample model checks.[5]

4. A global natural experiment

There are two main challenges for assessing directly methods for extrapolating causal effects. First is to find a randomized intervention or a naturally occurring experiment that has been implemented in a wide range of settings around the world. The second is to find data that is readily available and comparable across the different settings.

For the first challenge, we propose to use sibling sex composition to understand its impact on fertility and labor supply decisions. The starting point of our paper is Angrist and Evans (1998), who show, using census data from 1980 and 1990 in the US, that families have on average a preference to have at least one child of each sex. Since gender is arguably randomly assigned, they propose to use the sibling sex composition of the first two children as an exogenous source of variation to estimate the causal impact of fertility on labor supply decision of the mother.

For the second challenge, we make use of recently available data from the Integrated Public Use Microdata Series-International (IPUMS-I). This project is a major effort to collect and preserve census data from around the world. One important dimension of IPUMS-I is their attempt to harmonize the data and variables in order to make them comparable both across time and space. For our particular application, we

[5] An alternative approach to characterize the external validity function is to use matching, for example to select a given country-year as the target and to use the remaining country-years as the reference setting, then to match individuals in the former to the latter (using for example propensity score matching or direct matching) and then to cycle through all possible country-years as targets. To explore the functional form of the extrapolation error associated with non-zero differences in covariates, we can impose prior constraints on the matches in the reference setting (e.g., constrain matches to given a level of education or some other covariate). Results using this method are similar to those presented below, but much more computationally intensive.

were able to use 169 country-year samples (with 66 unique countries) with information on fertility outcomes and 166 samples (and 61 unique countries) with information on labor supply decisions (although our sample size decreases to 142 and 128 country-years respectively when we merge in additional country-level covariates).

The use of the Angrist-Evans same-sex experiment on a global scale brings additional challenges, which were not faced in the original paper. In particular, sex selection for the first two births, which does not appear to be a significant factor in the United States (Angrist and Evans 1998), could be a factor in countries where son-preference is a stronger factor than the US. We view sex selectivity as one of the context covariates, W, that could be controlled for when comparing experimental results to a new context of interest, or if not appropriately controlled for could undermine external validity. In our results below we pursue three approaches: not controlling for differences in sex selectivity and examining whether external validity still holds; directly examining its effect on the external validity; and excluding countries in which selection is known to be widely practiced.

Another challenge is that, if the cost of children depends on sibling sex composition, then *Same-Sex* would violate the exclusion restriction that formed the basis of Angrist and Evans's original instrumental variables approach, affecting fertility not only through the taste for a gender balance but also through the cost of additional children (e.g., with two same sex children hand-me-downs lower the cost of a third child and thus could affect not only fertility but also labor supply). Butikofer (2011) examines this effect for a range of developed and developing countries, and argues that this is a concern for the latter group. As a result, in this analysis, we use *Same-Sex* as a reduced-

form natural experiment on incremental fertility and on labor supply, and do not present instrumental variables estimates.

For our empirical analysis, we implement essentially the same sample restrictions, data definitions, and regression specifications as those proposed in Angrist and Evans (1998).[6] Since the census data that we use does not contain retrospective birth histories, we match children to mothers as proposed by Angrist and Evans (1998), using the harmonized relationship codes available through IPUMS-I, and we also restrict our analysis to married women aged 21-35 whose oldest child was less than 18 at the time of the census. In our analysis we define the variable *Same-Sex* to be equal to 1 using the sex of the oldest two children.

As outcomes we use an indicator for the mother having more than 2 children (*Had more children*) and for the mother working (*Economically active*). These two outcomes correspond to the two reduced-form specifications of Angrist and Evans. While there is a natural link between *Same-sex* and *Had more children*, the link is less intuitive for *Economically active*. In the context of instrumental variables, the link is presumably through incremental fertility (and is assumed exclusively to be so). In our application, since no exclusion restriction is assumed, the effect can include not only incremental fertility but also, for example, the income and time effects of having two children of the same sex. As such, identification of the reduced-form effect of *Same-sex* on *Economically active* relies only on the validity of the experiment within each country-year and on our identifying assumptions outlined in Section 3. As we will see below, the

[6] The data and programs used in Angrist and Evans (1998) are available at: http://economics.mit.edu/faculty/angrist/data1/data/angev98

contrast between the two reduced form experiments is useful in thinking through issues of external validity.

Next we discuss the choice of individual (micro) and context (macro) variables to be included in our analysis. In the absence of a well-defined theory for our specific context, the choice of individual level variables to explain effect heterogeneity is based on some theoretical notions based on related models and work (Angrist and Evans 1998; Ebenstein 2009). We use the education level of both the mother and the spouse, the age of the mother as well as the age at first marriage for the mother as our main individual level variables. For the case of context variables, the choice seems somewhat more difficult. One obvious candidate is a measure of female labor force participation as a broad measure of employment opportunities for women in a given country (Blau and Kahn, 2001). Since the choice of covariates is limited and imperfect, and the goal of our exercise is to achieve extrapolation and prediction, we include a number of macro variables that do not necessarily play a direct causal role in explaining fertility and labor supply decisions but rather have been shown to be important in explaining broad patterns of socio-economic outcomes across countries. The main variable is log GDP per capita, as a broad indicator of development, but also the legal origin of a country (La Porta et al., 1998) and a measure of geography measured by latitude and longitude (Gallup, Mellinger and Sachs, 1998). To the extent that these variables pick up variation that we cannot measure directly it is useful to include them in the regression models.

Descriptive statistics for our 169 samples are provided in Table 1. On average 60% of women have more than 2 children (*Had more children),* which is our main fertility outcome. Furthermore, 49% of women in our sample report being *Economically*

active, which is our main labor market outcome. Summary statistics for a number of additional individual level variables as well as country level indicators are also presented in Table 1 and they include the education of the woman and her spouse, age, age at first marriage, and log GDP per capita.

For our main empirical specification for each country-year sample, we examine the treatment effect of the *Same-Sex* indicator on two outcome variables (*Had more children* and *Economically active*), and control for age of mother, own education, and spouse's education, subject to the sample restrictions discussed above. The country-year treatment effects are summarized in Appendix Table 1.

5. Graphically characterizing heterogeneity

To motivate our analysis, we start by providing a graphical characterization of the heterogeneity of the treatment effects in our data. Figure 1 is a funnel plot, which is a scatter plot of the treatment effect of *Same-Sex* on *Had more children* in our sample of 142 country-year samples against the standard error of the treatment effect. The region within the dotted lines in the figure should contain 95% of the points in the absence of treatment-effect heterogeneity. Figure 1 clearly shows that there is substantial heterogeneity for this treatment effect that goes beyond what one would expect to see were it a homogenous treatment effect with mean-zero random variation. A similar, but less stark, picture arises in Figure 2, which presents the funnel plot of *Same-Sex* on *Economically active* in the 128 samples that have census information on this labor market outcome.

The next set of figures further investigates the heterogeneity of treatment effects. In particular, to the extent that Figures 1 and 2 document cross-country-year heterogeneity in the treatment effect, is any of it driven by heterogeneity in observable covariates? In Figures 3 and 4 we plot the size of the treatment effect of *Same-Sex* on *Had more children* (Figure 3) and *Economically active* (Figure 4) on the y-axis against the proportion of women with a completed secondary education based on data from 142 census samples (on the x-axis). Figure 3 shows a positive linear relationship that suggests that the treatment effect is larger in countries with a higher proportion of educated mothers. The same figure also displays geographic heterogeneity by color-coding each point based on geographic region, which suggests small (or zero) effects in countries of Sub-Saharan Africa. The corresponding effects for *Economically active* in Figure 4 are suggestive of a negative relationship between the treatment effect size and the level of education in a country, without a strong geographical pattern.

Finally, in Figures 5 and 6 we repeat the analysis from the previous two figures but instead we describe the heterogeneity with respect to log GDP per capita in a country. Figure 5 shows a striking linear pattern, suggesting the treatment effects of *Same-Sex* on *Had more children* increase with income per capita. Since the proportion of women with a secondary education and the log of GDP per capita are clearly correlated, it implies that Figures 3-6 are not informative of the relative importance of one covariate over another. Nonetheless, these graphs as well as the funnel plots presented earlier all provide suggestive evidence showing that there is substantive heterogeneity for both of our treatment effects.

6. Homogeneity tests

The next step in our analysis is to quantify the heterogeneity described in the previous graphs. We start by presenting, in Table 2, the results of Q-tests for effect homogeneity, which quantify what is depicted in Figures 1 and 2 in terms of the heterogeneity in the observed effect sizes against what one would obtain as a result of sampling error if there were a homogenous effect. The resulting test statistics, which are tested against the Chi-square distribution, are extremely large (and the resulting p-values are essentially zero) and confirm statistically the visual impression of treatment effect heterogeneity for both treatment effects from Figures 1 and 2. The results are similar when the unit of observation is the country-year-education group.

Given that there is heterogeneity, for the second test we investigate if the effects are distributed in a manner that resemble a normal distribution. For this we have implemented an inverse-variance weighted Shapiro-Francia (wSF) test for normality of effect estimates. This test modifies the Shapiro-Francia test for normality (Royston 1993) by taking into account the fact that the country-year treatment effects are estimated with different levels of precision. Our modification involves using an inverse-variance weighted correlation coefficient as the test statistic rather than the simple sample correlation coefficient. The test statistic is the squared correlation between the sample order statistics and the expected values of normal distribution order statistics. In our specific example, where the outcome is *Had more children*, we take the order sample values for our 142 country-year observations and look at the squared correlation between the ordered statistics from our sample and the expected ordered percentiles of the standard normal distribution. The results in Table 2 confirm that for both of our outcome

variables we can reject the hypothesis of normality (i.e. we can reject that the correlation is 1). This result is not surprising in light of the visual evidence presented in Figures 1 and 2, which suggested that the distribution of our country-year effects is over-dispersed from what a normal distribution would look like. These findings are suggestive of over-dispersion being driven by variation in covariates that are prognostic of the magnitude of the treatment effects.

The rejection of homogeneity suggests the need to use available covariates to extrapolate to new contexts. In our example, the set of covariates is limited. At the micro level we have only the basic demographic characteristics included in the standardized IPUMS data. The set of country-year covariates is larger, although for reasons discussed above we have little reason to believe that a more extensive set of country-year characteristics beyond basic development indicators and more specific indicators like female labor force participation would add much in terms of explanatory value. We expect that such limits to available covariates would be typical of experimental evidence bases. With a limited set of covariates, using flexible and fairly agnostic methods for estimation best satisfies our goal of extrapolation with minimal prediction error. In the applications that follow, we use flexibly specified regressions, although matching yields similar results[7].

7. Characterizing heterogeneity: unconditional external validity functions

In this section we empirically characterize the unconditional external validity function. As discussed in Section 3, within each possible pair of target-comparison country-year

[7]The number of available experiments limits the richness of the extrapolation model. In such cases, prediction may benefit from regularized methods such as the lasso, which we investigate in an appendix.

samples, we run separate regressions for *Same-Sex*=0 and 1 in the comparison country; the regressions include our main individual-level covariates (education, education of spouse, age of mother, year of census, and age at first marriage) as well as their interactions. We then predict what would have happened in the treatment and non-treatment groups in the target country. For each target-comparison pair we observe the prediction error in the estimated treatment effect, where the prediction error is defined as the difference between the predicted treatment effect and the quasi-experimental treatment effect using the actual *Same-Sex* values within a target country. We perform this exercise for all the possible combinations of target-comparison country years, which produces close to 28,000 dyads of all possible pair-wise combinations of country and year. For each dyad, we record the prediction error and also the dyad-level differences in the covariates of interest, such as education, age, year of census, or GDP per capita.

In this section, we characterize how prediction error changes with education levels, log GDP per capita and geographical distance, each considered individually (i.e. unconditionally, so for example prediction error arising from differences in education could be driven by correlated differences in GDP per capita). In order to do this, we run a local linear regression of prediction error at the dyad level on the covariate difference between the target and comparison countries associated with the prediction error.

The unconditional external validity function estimates for education are presented in Figure 7. Three features are notable. Prediction error is approximately zero at zero education distance. Prediction error increases with increasing differences in education levels; for a one-point education difference (on a four point scale) bias increases by approximately 0.1 (relative to the world treatment effect of 0.04 in Figure 1). The figure

also plots the variance of the external validity function, which is relatively flat over the range of -1 to +1 educational differences, but increases considerably at greater differences.

Figure 8 shows a similar pattern when we explore how the prediction error changes with GDP per capita. The error at zero GDP per capita distance is close to zero, and increases to about 0.1 for GDP per capita differences of $20,000. Unlike Figure 7, we note an increase in the variance of the error even for small increases in the GDP difference. In Figure 9 we focus on women's labor force participation differences and again we observe that any deviations in labor force participation distance are associated with higher prediction error.

In Figure 10, we present external validity function estimates with respect to geographic distance, measured as the standardized distance in kilometers between the centroid of a target and comparison country (where a one standard deviation difference is approximately 4800 km). Geographic distance is presumed to proxy for various cultural, climactic, or other geographically clustered sources of variation in fertility. Looking across all country-years, in Figure 10, panel a, we do not find a significant relationship between geographical distance and prediction error. Non-linear features of geographical distance, most notably oceans, complicate this relationship. To account for this, in Figure 10, panel b, we present differences within contiguous regions (North and South America, Europe, Asia, and Africa). Again, we do not find any statistically significant relationship for distances less than 10,000 km. The estimated external validity function is positively sloped, so for distances in excess of approximately 10,000 km, there is a statistically significant increase in extrapolation error.

8. Characterizing heterogeneity: conditional prediction error regressions

In this section we continue our characterization of heterogeneity using the dyadic prediction error regression approach outlined in Section 3 to evaluate how the prediction error changes with differences in a covariates of interest, controlling for other covariate differences. It is worth noting that our covariates of interest become country-year level averages, even if some of them, such as education or age, are constructed from census micro level variables.

The results from this exercise are presented in Tables 3 and 4, where we standardize covariate differences. In order to interpret the coefficients it is useful to note that the standard deviation of the education variable is close to 1, for age it is about 3.5 years, for census year it is 11 years, for log GDP per capita is about 10,000 dollars, and for distance it is about 4800 km.

In columns 1-8 of Table 3, we run the prediction error regressions with only one covariate at a time, giving us essentially the unconditional prediction error. One can observe that most of our covariates (measured as the absolute differences between country pairs in education, education of spouse, year of census, log GDP per capita and labor force participation) are statistically significant.

Maybe most interestingly, our prediction error regression framework allows us to include all covariates in the same regression in order to begin to run some meaningful "horse-races". Three main conclusions could be drawn from the results from these regressions, which are in presented in Columns 9 and 10 of Table 3. First, given our sample sizes, many of our variables are statistically significant, although we note that log

GDP per capita loses its significance once the other controls are included. Second, the size of the prediction error is generally large given an average treatment effect in the sample of 0.04. Third, the actual magnitude of the prediction error generated by differences in these respective variables does not seem to be very different between micro and macro variables. For example in column 9 of Table 3, a one standard deviation difference in female labor force participation leads to an increase in absolute prediction error of about 0.024 in the effect of *Same-Sex* on *Had more children*. This effect is similar compared for example to the education of the mother, where a one standard deviation difference in education (which is essentially one educational category), creates an absolute prediction error of 0.03 in the effect of interest.

The results in Table 4 for the effect of *Same-Sex* on *Economically active* show a similar picture. The age of the mother, log GDP per capita, geographic distance and labor force participation are important drivers of heterogeneity. Interestingly the education of the mother and spouse are not important explanatory variables.

In columns 9 and 10 of Tables 3 and 4 it is noteworthy that differences in the sex ratio imbalance (which is the difference in the proportion of two boys vs. two girls among the first two children and is our proxy for the importance of gender prediction error at the country level) is not a significant predictor of prediction error. This suggests that, even if sex selection in the gender of the first two children remains a concern, it nonetheless is not a driver of prediction error at the country-year level. In column 11 of Tables 3 and 4 we repeat the same analysis as in column 10 but drop sex-selecting countries from the analysis and note that the results are virtually unchanged when we perform this

robustness check (with one unexpected exception: the sex imbalance variable becomes significant in Table 4, column 11).

While the results in Tables 3 and 4 allow us to compare the simultaneous importance of a range of covariates difference on prediction error, they do not allow us to judge the importance of micro vs. country-level covariates. Since dyads are formed at the country level, micro-level covariates differences are aggregated to that level. In order to get at this issue, we perform the following exercise for each country-year sample. We take a given country-year as the target country, and the other 165 countries are treated as experimental sites. In the 165 experimental sites, we run a separate regression for the treated and the control observations, and we use these to predict the treatment and the control outcomes and the treatment effect in the target site. We consider four cases in terms of possible sets of regressors: (1) one without any covariates, which recovers the unadjusted estimates; (2) the individual micro covariates including age of the mother, a set of dummies on mother's educational attainment, a set of dummies on the education of the spouse, age at first marriage, as well as all the possible interactions of these individual-level variables; (3) macro covariates consisting of log GDP per capita, labor force participation, dummies for British and French legal origin, as well as a variables for the latitude and longitude of a country; and (4) the combined covariates that consist of the union of micro (group 2) and macro variables (group 3).

We use the difference between the actual treatment effect and the predicted treatment effect to generate the prediction error. This exercise generates 166 data points for each of the four covariate sets, which we plot for the case of *Had more children* in Figure 11 and for *Economically active* in Figure 12. The four groups are unadjusted

(blue), micro variables only (red), macro variables only (green), and micro and macro variables together (gold). In panel A of each figure, we plot the density estimates of these prediction errors, while in panel B we plot the CDFs of the absolute prediction error.

Looking at Figure 11, we observe that in the case of *Had more children*, both micro and macro variables are doing fairly well in terms of pushing prediction error towards zero. In panel B of the same figure where we plot the CDFs of the absolute prediction error, one can see that any set of covariates dominates the scenario of no covariates. At the same time, using both the micro and macro variables increase the mass of the distribution at lower levels of bias, although also pushes out the tails. The results in Figure 12, which use *Economically active* as the outcome variable of interest, provide a rather different picture. In this case, micro variables do not seem useful in terms of reducing the prediction error, a finding that is in line with the arguments provided in Pritchet and Sandefur (2013). But equally remarkable is how well macro variables do in terms of reducing prediction error. The implication of these results, at least for the case of explaining variation in *Economically active*, is that a set of easily available cross country variables has the potential to be useful in analyzing of external validity.

Overall, our results suggest that both micro and macro level covariates are important drivers of prediction error, although the relative importance depends on the outcome of interest that is studied.

9. The accumulation of evidence and prediction error

Our results so far imply that while our covariates have some predictive power in explaining treatment effect heterogeneity, the magnitude of the prediction error remains

considerable. As a result we continue our analysis with an attempt to understand if and how the accumulation of experiments over time improves our ability to extrapolate to new settings.

In Figure 13, we fit the model for the effect of *Same-Sex* on *Had more children* from Section 8 and Table 3 column 10 on the sample of country-year dyads available at each point in time, and then estimate the model's prediction error for those country-years dyads. As an example, we take all the census country-year dyads available by 1980; fit the model to these dyads; and then estimate the treatment effect prediction error for this sample. In Figure 13, we plot the resulting average prediction error values over time. The pattern clearly shows that as we add more data to the model over time, our predictions have smaller average prediction error and eventually get closer to zero. The second striking pattern on the graph can be seen looking at the two standard deviation bounds of the average prediction error over time: our predictions not only have smaller prediction error but they also become more precise. The corresponding analysis for the effect of *Same-Sex* on *Economically active* is presented in Figure 14 and shows broadly consistent patterns with those described in Figure 13.

The results so far imply that as we accumulate more experiments over time our ability to fit prediction error models from experiments to external environments improves. At the same time, Figures 13 and 14, depict in-sample predictive accuracy or model fit, and not an out-of-sample test of the accuracy of the model's predictions. This is examined in Figures 15 and 16.

For the target country-year samples in a given year on the x-axis (e.g., the U.S. in 1980), we measure the prediction error from estimating the treatment effect for those

country-years using four different groups of country-year comparisons selected from country-years available up to that point in time (e.g., continuing the example, country-year samples up to 1980, other than the U.S. in 1980). The four groups of comparison countries are: (1) all available country years (graphed as the red line); (2) the best comparison country-year as selected by our model (graphed as the blue line), where the prediction error model used to select the best comparison is fit using data from prior years (so in our example, estimated on data available prior to 1980); (3) the nearest country-year by geographical distance excluding own-country comparisons (graphed as the orange line); and (4) the nearest country-year by geographic distance, allowing own-country comparisons (graphed as the green line).

A number of interesting patterns arise from this exercise. First, the comparison of the first two groups of comparison countries (all available country-years in red versus the best comparison selected by the model in blue) confirms that when using our model we get much lower prediction error compared to using all the samples available. Second, the pattern over time of prediction error from using the best model-selected comparison country-year shows that the accumulation of more samples plays a modest but meaningful role in reducing the prediction error. Modest in the sense that the prediction error from the best comparison country-year suggested by the model hovers between 0.08 and -0.05, suggesting that the model is reasonably accurate in making predictions even with a limited number of available samples, at least for this particular setting. But also meaningful in the sense that the prediction error tightens considerably (ranging between 0.02 and -0.03) from 1985 onward.

Finally, and more speculatively, we are interested in how some simple rule-of-thumb selection criteria perform. We start with the fourth comparison group that contains the nearest country in geographic distance (and can include the country itself from a prior time period). The prediction error is initially negative and becomes smaller over time, suggesting that with the addition of more experiments, the rule of thumb starts to perform well, likely because the geographically nearest match tends to be quite similar. In contrast, and somewhat surprisingly, the third comparison group that contains the nearest country-year by geographic distance but excludes own-country comparisons performs well over the entire period and arguably as well as our model-based approach.

We would argue that this result illustrates the risks of rules of thumb compared to a model based approach. *A priori*, allowing own-country comparisons seems intuitive, but own country comparisons are usually at least 10 years apart and our model easily accommodates the optimal balance between these competing factors. This is underlined in looking at Figure 16, where the model outperforms both rules of thumb when the available comparison samples are sparse.[8]

10. Applications

In this section we consider two applications of the framework we have presented. While the natural experiment we have examined, the effect of *Same-sex* on fertility, clearly is not a intervention that could or would be implemented by a policy maker, as a thought experiment we treat it as such, and examine how our framework would be used to address two questions a policy maker could face: (1) where to locate an experiment to minimize

[8] In Figures 17 and 18 we repeat the analysis presented in Figures 15 and 16, but we drop the countries in our sample (China, India, Nepal, and Vietnam) that display sex selection at birth. The broad patterns are remarkably similar.

average prediction error over a set of target sites, and (2) when to rely on extrapolation from an existing experimental evidence base rather than running a new experiment in a target site of interest.

10.1 Where to locate an experiment

Imagine a policy researcher interested in the effect of an intervention around the world, but with limited resources to implement a randomized controlled trial. In this section we examine what the estimated external validity function implies for the best location of an experimental site. In particular, which site has the lowest mean squared prediction error for the treatment effect in other sites? And given the first choice, where would one locate a second experimental site?

At the country-year level, our estimates of the unconditional external validity function imply that a country with lowest average covariate distance to other country-years should be the best predictor. The question then is determining how appropriately to weight different covariates. With knowledge of the conditional external validity function in Tables 3 and 4 one would weight each covariate by its conditional importance for external validity, or more directly one could also weight each covariate by its conditional influence on the country-year treatment effect. In Figure 19, we consider the exercise of using each country-year to predict the other country-years in our sample, where the x-axis plots each country-year by the percentile of its composite covariate, i.e., the sum of covariates weighted by their conditional predictive relevance for the treatment effect, and where the y-axis plots the associated mean error from predicting the treatment effect for

other country-years. We see immediately that the lowest average prediction error is indeed at the median, which turns out to be the United States in 1980.

The challenge in thinking of this prescriptively is that a policy maker will presumably not know the conditional importance of each covariate for external validity without first running the full set of experiments. In Figure 20, we consider an alternative that does not rely on knowledge of the treatment effect; namely, we compute the average Mahalanobis distance between each country-year and the other country-years. The figure plots average prediction error against average distance of a country-year from other country-years. Again, it is evident that the country-year with the lowest average distance to other country-years offers the lowest prediction error of the treatment effect; the relationship is also monotonic.

Carrying the thought experiment further, in Figures 21 and 22 we consider adding a second country-year, conditional on the first choice. Again, the lowest prediction error is associated with country-years that are in the middle of the covariate distribution or that have the lowest average covariate distance to other country-years, which in this case turns out to be Chile in 1982.

As one carries the thought experiment yet further, however, it is less evident that one would continue to select sites in the middle of the covariate space. For example, one could imagine the value of adding sites in the tails of the covariate distribution to more precisely estimate non-linearities. One could also imagine the value of including large, heterogenous country-years that are useful in predicting the treatment effect in a broad range of countries.

10.2 To experiment or to extrapolate?

We consider a situation where a policy maker wants to make an evidence-based policy decision of whether or not to implement a program. The policy maker has a choice between using the existing evidence base versus generating new evidence by carrying out an experiment in the target context. That being the case, the choice is really between whether the existing evidence base can provide a reliable enough estimate of what would be found from the new experiment, thus making the new experiment unnecessary. One might imagine different ways to characterize the loss function governing this decision. We develop an approach based on the assumption that a new experiment is only worthwhile if the existing evidence base is sufficiently ambiguous about the potential effects of the treatment in the target context.

Let the target context be characterized by an $N_1 \times K$ coefficient matrix, W_1, that includes K subject- and context-level attributes for N_1 subjects.[9] We assume that the policy maker will decide that the existing evidence is sufficient to determine policy if a 95% prediction interval surrounding the conditional mean prediction at X_1 is entirely on one or another side of some critical threshold, c^*. We also assume that the experiment that the policy maker could run in the target context is adequately well powered that she would find it worthwhile to run the experiment if the existing evidence is ambiguous. For this analysis, we simply assert a decision based on c^* and a 95% prediction interval. Figure 23 illustrates the decision problem graphically. If the predictive interval resembles either of the solid-line distributions, then the evidence is certain enough to rule

[9] Alternative we could characterize the target context in terms of a joint distribution F_1 over the K dimensional covariate space. Thinking in terms of a finite number of subjects simplifies the exposition.

out the need for an experiment. If the interval resembles either of the dashed line distributions, then the existing evidence is too vague and a new experiment is warranted.

This is a reduced-form characterization of any number of more fully-fledged analyses. A fully Bayesian decision analysis under a Normal model could begin with the premise that the policy maker implements the program if the posterior distribution for the program effect provides a specified degree of certainty that the effect will be above some minimal desirable effect value. Then, c^* and the relevant prediction interval could be defined as a function of the minimum desirable effect value, the level of certainty required, posterior variance, and the moments of the predictive distribution. With c^* and the relevant prediction interval defined, the analysis would otherwise proceed as we describe here.

For a covariate matrix W recall that in expression (3) we defined a conditional effect estimate, $\hat{\tau}(W)$. We consider this relative to the c-th draw from the effect distribution, τ_c. Recall that for τ_c, the external validity function defines the prediction error as

$$\epsilon_c(W) = \hat{\tau}(W) - \tau_c.$$

At a given location in the covariate space, W, we have a distribution of effects, τ, and so a distribution of prediction errors, ϵ, given the effect estimate, $\hat{\tau}(W)$. Assume $Cov(\hat{\tau}(W), \epsilon(W)|W) = 0$. For $W = W_1$, we have,

$$Var[\tau|W = W_1] = Var[\hat{\tau}(W)|W = W_1] + Var[\epsilon(W)|W = W_1].$$

The first term on the right measures the estimation error given the existing evidence base, while the second measures the intrinsic variation in effects at $W = W_1$ regardless of the evidence base. We will work under an assumption that

$$\tau|W \sim N\big(\mu(W), \sigma^2(W)\big),$$

which is a substantive restriction that allows us to establish a reference distribution for constructing the prediction interval. This is a restriction that ought to be tested in any applied setting; whether it is satisfied will depend in part on how one defines W.[10] Then, the solution to the decision problem is to experiment if

$$\hat{\tau}(W_1) - t_{0.025}\sqrt{Var[\tau|W = W_1]} < c^* < \hat{\tau}(W_1) + t_{0.025}\sqrt{Var[\tau|W = W_1]}$$

and to accept the existing evidence otherwise, where $t_{.025}$ is the appropriate .025 quantile value for our approximation of the normalized conditional distribution of τ. These upper and lower bounds correspond to the upper and lower bounds of the 95% prediction interval for τ_c.

To implement these ideas, we use a flexible regression-based approach. We use OLS to fit a regression specified as follows,

$$y_{ic} = \alpha + \tilde{\tau}T_{ic} + W_{ic}^{1\prime}\ \beta + \epsilon_{ic},$$

[10] A more flexible approach would be to use a sieve-type estimator that uses normality as a starting point but then allows for departures from normality based on the data. However, for such an estimator to make a difference one would need more data on effect sizes than is typically available.

where i indexes individuals in experiments indexed by c, and the covariate vector W_{ic}^{1} contains a series of individual and context-level covariates centered on the means of the covariates as they appear in W_1 as well as interactions between these centered covariates and the treatment indicator, T_{ic}.[11] The OLS estimate for $\tilde{\tau}$ provides our estimate of $\hat{\tau}(W_1)$ and a cluster robust variance estimator (clustering by experiment) provides an asymptotically conservative approximation to $Var[\hat{\tau}(W)|W = W_1]$ under random selection of experimental contexts and random assignment within the experiments (Abadie et al., 2014; Lin, 2013). To estimate $Var[\epsilon(W)|W = W_1]$, we first generate $\hat{\tau}(W_c)$ estimates for all experiments in our reference sample, indexed by c. We also extract unbiased experimental estimates for each these contexts, $\hat{\tau}_c$ for all c. We then obtain

$$\hat{\epsilon}_c = \hat{\tau}_c - \hat{\tau}(W_c)$$

for all c. We use OLS to fit a regression specified as

$$\log(\hat{\epsilon}_c^2) = \theta + \overline{W}_c^{0\prime}\ \gamma + \nu_c$$

where $\overline{W}_c^0$ is contains a series of covariate means centered on the means for the target context defined by W_1.[12] We take $\exp(\hat{\theta})$ as our estimate for $Var[\epsilon(W)|W = W_1]$, with $\hat{\theta}$ being the OLS estimate for θ. This estimate will tend to be conservative

[11] With dummy variable covariates and a saturated specification, this centered interaction model yields an estimate of $\hat{\tau}(X_0)$ that is algebraically equivalent to a matching estimator (Angrist and Pischke, 2009).
[12] A richer specification could consider other moments of the covariates in X_0.

for $Var[\epsilon(W)|W = W_1]$, because $\exp(\hat{\theta})$ is generated off of the $\hat{\tau}$ values rather than the actual τ values.

Finally, we use the 0.025 critical value for the standard normal distribution in place of $t_{0.025}$; this will tend to understate the prediction dispersion in finite samples and therefore biases the analysis against running a new experiment. A more conservative approximation could be used to shift the bias in favor of a new experiment.

Table 5 shows the results of applying this approach to 79 of the country-year samples. These 79 samples were chosen because we were able to obtain country-year covariates data for them from the Penn World Tables, measuring population density, log real GDP/per capita, government spending share of real GDP/capita, ethnic fractionalization, female labor force participation rate, and year. These 79 samples were also recent enough such that there were enough pre-existing reference samples that allow for sufficient degrees of freedom to fit predictive models with the six country-year covariates (in addition to the micro-level covariates). Table 5 shows the estimated prediction intervals for the 79 cases as well as the actual in-sample estimates (from Appendix Table 1). We generated the prediction intervals using one-percent extracts from each of the country-year samples, rather than the entire sample, to facilitate computation.[13] The coverage rate for the nominal 95% prediction intervals is 93.7%, suggesting reasonable calibration.

Table 5 shows that intervals tend to shrink in size as the years move forward, but only when we have a rather large number of samples would we expect that decision makers begin to rely on the existing evidence base. For targets during the 1970s and

[13] For each prediction interval estimate, it is necessary to obtain not only an extrapolation estimate for the target country-year, but also to produce leave-one-out extrapolation for all reference country-years in order to properly model the conditional prediction error distribution.

1980s (22 target cases, drawing from a reference set of 19 samples starting in 1974 to 42 by 1989), the interval width was on average about 0.303. This is an order of magnitude larger than the underlying effect sizes. By the 1999, the number of reference cases reaches 72, and during the 1990s, the interval width averages 0.170. By the time we get to the 2000s, with between 80 to 100 reference cases available, the average interval width is 0.131. At the same time, interval widths do not tighten uniformly over time. Although the reference base grows richer as years move forward, it is also possible that a given country-year sample moves out to a sparsely covered area of the covariate distribution. The latter phenomenon would lead to wider prediction intervals, despite the availability of more reference cases.

11. Conclusion

This paper has examined whether, in the context of a specific natural experiment and a data context, it is possible to reach externally valid conclusions regarding a target country-year of interest using the available experimental evidence. We view this paper as having made six contributions to the literature. First, we provide and implement a simple framework to consider external validity. Second, we come up with a context in which it is possible, and meaningful, to ask and potentially to answer questions of external validity. While experiments are run globally, to our knowledge there is no one experiment that has been in run in as many countries, years, and geographical settings as the *Same-Sex* natural experiment. While it has challenges as a natural experiment, we view our exercise as a possibility result: is external validity – notwithstanding the challenges – possible? Third, using both parametric and non-parametric techniques, we present external validity

function results that directly answer the central question of external validity, namely the extent to which valid conclusions about a target context of interest can be drawn from the available experimental data. Fourth, we show that given the accumulation of sufficient experimental evidence it is possible to draw externally valid conclusions from our experimental evidence, but the ability to do so is meaningfully improved (over rule of thumb alternatives) by the modeling approach we adopt. Fifth, we show that prediction error can in general depend on both individual and context covariates, although in one of our applications (the effect of *Same-sex* on labor supply) only the latter reduced prediction error. Finally, we considered two applications for our approach, showing that experiments located near the middle of the covariate distribution tend to provide the most robust external predictions and that in some contexts it is possible that a policy maker may choose to extrapolate the treatment effect from an existing experimental evidence base rather than run a new experiment.

While our conclusions are cautiously optimistic, it is essential to underline the deductive nature of our exercise. Given the importance of the question and paucity of evidence, we believe even a single attempt to assess the external validity of experimental evidence is valuable, despite its flaws and limitations. A better understanding of our ability to learn from the rapidly accumulating experimental evidence base and to answer key policy and economic questions of interest will require further extensions and replications of the exercise we have begun here.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge (2014). "Finite Popultion Causal Standard Errors," NBER Working Paper 20325.

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

Angrist, Joshua (2004), "Treatment Effect Heterogeneity in Theory and Practice," *The Economic Journal*, Volume 114, C52-C83.

Angrist, Joshua, and William Evans (1998), Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review*, Volume 88, Number 3, pp. 450-477.

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

Aronow, Peter, and Allison Sovey (2013), "Beyond LATE: Estimation of the Average Treatment Effect with an Instrumental Variable," *Political Analysis*, Volume 21, pp. 492-506.

Bareinboim, Elias, and Judea Pearl (2013), "Meta-Transportability of Causal Effects: A Formal Approach," *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, pp. 135-143.

Blau, Francine D. and Lawrence M. Kahn, (2001) "Understanding International Differences in the Gender Pay Gap," NBER Working Paper 8200.

Butikofer, Aline (2011), "Sibling Sex Composition and Cost of Children," manuscript.

Campbell, Donald T. (1957), "Factors Relevant to the Validity od Experiments in Social Settings," *Psychological Review*, Volume 54, Number 4, pp. 297-312.

Card, David, Jochen Kluve, and Andrea Weber (2010), "Active Labor Market Policy Evaluations: A Meta-Analysis," NBER Working Paper 16173.

Cole, Stephen R., and Elizabeth Stuart (2010), "Generalizing Evidence from Randomized Clinical Trials to Target Populations: The ACTG 320 Trial," *American Journal of Epidemiology*, Volume 172, Number 1, pp. 107-115.

Cruces, Guillermo, and Sebastian Galiani (2005), "Fertility and Female Labor Supply in Latin America: New Causal Evidence," SSRN Working Paper 2359227.

Crump, Richard K., V. Joseph Hotz, Guido W. imbens, and Oscar A. Mitnik (2008), "Nonparametric Tests for Treatment Effect Heterogeneity," *The Review of Economics and Statistics*, Volume 90, Number 3, pp. 389-405.

Dehejia, Rajeev (2003), "Was There a Riverside Miracle? A Hierarchical Framework for Evaluating Programs with Grouped Data," *Journal of Business and Economic Statistics*, Volume 21, Number 1, pp. 1-11.

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

Efron, B., I. Johnstone, T. Hastie, and T. Tibshirani (2004), "Least Angle Regression," *Annals of Statistics*, Volume 32, Number 2, pp. 407-451.

Freedman, David A. (2008). "On Regression Adjustments in Experiments with Several Treatments," *The Annals of Applied Statistics*, Volume 2, Numbe 1, 176-196.

Gallup, John L., Mellinger, Andrew D., and Sachs, Jeffrey D. (1998), "Geography and Economic Development." NBER Working Paper 6849.

Glass, Gene V. (1976), "Primary, Secondary, and Meta-Analysis of Research," *Educational Researcher*, Volume 5, Number 10, pp. 3-8.

Green, Donald P., and Holger Kern (2012), "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees," *Public Opinion Quarterly*, Volume 76, Number 3, pp. 491-511.

Greenland, Sander (1994), "Invited Commentary: A Critical Look at Some Popular Meta-Analytic Methods," *American Journal of Epidemiology*, Volume 140, Number 3, pp. 290-296.

Hartman, Erin, Richard Grieve, Roland Ramashai, and Jasjeet S. Sekhon (2013), "From SATE to PATT: Combining Experimental with Observational Studies to Estimate Population Treatment Effects," manuscript, University of California, Berkeley.

Heckman, James J., and Edward J. Vytlacil (2007), "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments," In James J. Heckman and Edward E. Leamer, eds., *Handbook of Econometrics*, Volume 6B, pp. 4875-5143.

Heckman, James J., Hiehiko Ichimura, Jeffrey Smith, and Petra Todd (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica*, Volume 66, Number 5, pp. 1017-1098.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, New York, NY: Springer.

Hedges, Larry V. and Ingram Olkin (1985), *Statistical Methods for Meta-Analysis*, New York, NY: Academic Press.

Hotz, V. Joseph, Guido W. Imbens, and Julie H. Mortimer (2005), "Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations," *Journal of Econometrics*, Volume 125, Number 1, pp. 241-270.

Imai, Kosuke, and Marc Ratkovic (2013), "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation," *The Annals of Statistics*, Volume 7, Number 1, pp. 443-470.

Imbens, Guido W. (2010), "Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature*, Volume 48, Number 2, pp. 399.423.

La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert Vishny (1998), "Law and Finance," *Journal of Political Economy*, Volume 106, pp. 1113-1155.

Mundlak, Yair (1978), "On the Pooling of Time Series and Cross Section Data," *Econometrica*, Volume 46, Number 1, pp. 69-85.

Lin, Winston (2013) "Agnostic Notes on Regression Adjustment for Experimental Data: Reexamining Freedman's Critique," *The Annals of Applied Statistics*, Volume 7, Number 1, 295-318.

Olsen, Robert B., Larry L. Orr, Stephen H. Bell, and Elizabeth Stuart (2012), "External Validity in Policy Evaluations that Choose Sites Purposively," *Journal of Policy Analysis and Management*, Volume 32, Number 1, pp. 107-121.

Pearl, Judea, and Elias Bareinboim (2014), "External Validity: From do-Calculus to Transportability Across Populations," *Statistical Science*, forthcoming.

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 336.

Royston, Patrick (1993), "A Pocket-Calculator Algorithm for the Shapiro-Francia Test for Non-Normality: An Application to Medicine," *Statistics in Medicine*, Volume 12, Number 2, pp. 181-184.

Rubin, Donald B. (1992), "Meta-Analysis: Literature Synthesis or Effect-Size Surface Estimation?" *Journal of Educational and Behavioral Statistical*, Volume 17, Number 4, pp. 363-374.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Belmont, CA: Wadsworth.

Slavin, Robert E. (1984), "Meta-Analysis in Education: How Has It Been Used?" *Educational Researcher*, Volume 13, Number 8, pp. 6-15.

Slavin, Robert E. (1986), "Best Evidence Synthesis: An Alternative to Meta-Analytic and Traditional Reviews," *Educational Researcher*, Volume 15, Number 9, pp. 5-11.

Stanley, T.D. (2001), "Wheat from Chaff: Meta-Analysis as Quantitative Literature Review," *Journal of Economic Perspectives*, Volume 15, Number 3, pp. 131-150.

Stuart, Elizabeth A., Stephen R. Cole, Catherine P. Bradshaw, and Philip J. Leaf (2011), "The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials," *Journal of the Royal Statistical Society*, *Series A*, Volume 174, Part 2, pp. 369-386.

Sutton, Alexander J., and Julian P.T. Higgins (2008), "Recent Developments in Meta-Analysis," *Statistics in Medicine*, Volume 27, Number 5, pp. 625-650.

Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society*, Series B, Volume 58, No. 1, pp. 267-288.

Vivalt, Eva (2014), "How Much Can We Generalize from Impact Evaluation Results?", manuscript, New York University.

Appendix

A1. Characterizing heterogeneity via the interaction function: the Mundlak estimator and lasso regression

We characterize treatment effect heterogeneity with regressions that include individual (micro) level and country (macro) level predictors of heterogeneity. We use the Mundlak estimator, originally proposed by Mundlak (1978), and for our case, we include the country-year level means as well as the deviations from these means within country-year. The advantage of this approach is that the resulting coefficients on these micro level deviations from the means are the same as what one would obtain from a regression of the same micro level variables that includes country-year fixed effects. In addition, the coefficients on country-year level means account for the between country-year heterogeneity associated with those variables. Thus, it provides a direct way to evaluate the importance of micro versus macro level variation and their relative contribution to heterogeneity.

As an example, if one were to end up with significant coefficients on only the micro level variables but not on the country level means, it would suggest that within a country, this micro level variation matters. But once one has accounted for that variation, there is nothing more to be said about between country variation beyond that. On the other hand if one were to still get significance on any of the country level means, it would imply that there is a country level moderating effect of differences in the individual profiles of the population.

We explore the results in Table A2. The first set of variables, which are at an individual level of aggregation, capture the micro-level variation, such as age and

education of the mother and her spouse. Next, there are a number of variables, which are at the country-year level of aggregation, that capture between-country heterogeneity. Finally, the last group of variables, which are at the country level of aggregation, capture country-level variables that are time invariant, such as the level of ethnic fractionalization or continent indicators.

We implement a partial regression approach. More precisely we partial out all of the un-interacted covariates from terms that interact the treatment variable with the various covariates and from the outcome, and then we run the regression of those partial outcomes and partial interaction terms in order to obtain these moderator effects. As a result of this partial regression approach, the coefficients in Table A2 capture the interaction between the treatment effect and the variables, i.e., are showing us what variables explain treatment effect heterogeneity.

The size and significance of the coefficients in Table A2 allow us to describe the drivers of heterogeneity. One can observe that education of the mother and the spouse, the continent fixed effects as well as the level of GDP per capita are large and statistically significant. Some of the other variables, including the decade fixed effects or the age of the mother are more marginal in terms of their significance. In contrast, when we instead focus on the effect of *Same-Sex* on *Being economically active*, we generally find that our covariates are not able to explain the heterogeneity in treatment effects.

As a next step, we fit a lasso regression (Tibshirani 1996; see also Hastie et al., 2009) to the same partial regression specification in order to rank these interaction terms in terms of the extent to which they explain heterogeneity. The lasso estimates a linear regression under an imposed constraint that the sum of the absolute value of the

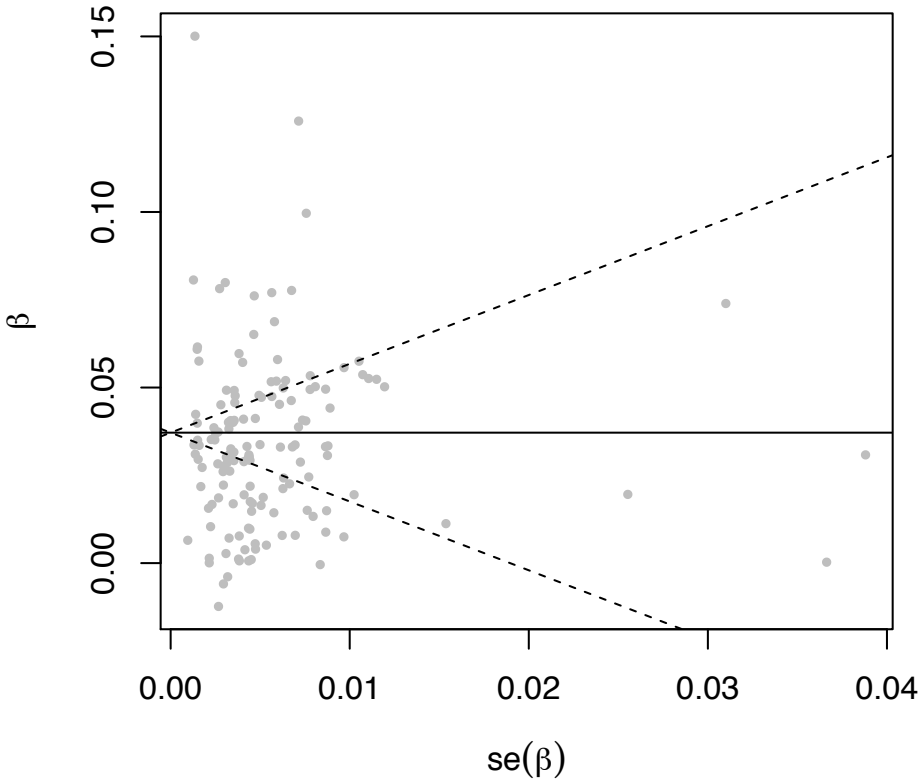
regression coefficients is less than or equal to a penalty parameter. Imposing this penalty on standardized variables creates a non-subjective criterion to determine the relative economic significance of coefficients, and in particular forces less significant coefficients to zero. Given our sample size, this is useful since most coefficients in a fully interacted model will tend to pass standard tests of statistical significance. We use the least-angle regression algorithm (Efron et al. 2004) to trace the solution path to fitting the lasso. This allows us to see which variables are successively included as we progressively relax the penalty.

The results in Tables A3 and A4 show a number of interesting patterns. As a start, if one wanted a model of just one variable to account for the treatment effect heterogeneity for the *Had more children* outcome, that variable would be log GDP per capita. Further, after having accounted for log GDP per capita, if one wanted to add one more variable it would be the country-year level mean of mothers' ages. Table A3 traces out the entire solution path; in other words, the procedure ranks the variables in terms of their contribution to the predictive accuracy when accounting for the heterogeneity. The third variable to enter is the country-year level average of mothers' education (an indicator for secondary education). Other variables to enter in successive steps include region dummies and then, finally at the individual level, spouse's education. Table A4 shows the same for the *Economically active* outcome. Again, we find that country-year and country-level variation tends initially to contribute most to predictive accuracy. In this case, the first few variables to enter include country-year mean education of spouses and mean mothers' age followed by region dummies. However, as anticipated by the regression results in Table A2, the inclusion of these variables, while doing the most to

boost predictive accuracy, only explain a negligible share of treatment effect heterogeneity as indicated by the extremely low R-squared values, which characterize variance explained in outcomes that have been residualized against the non-interaction terms.

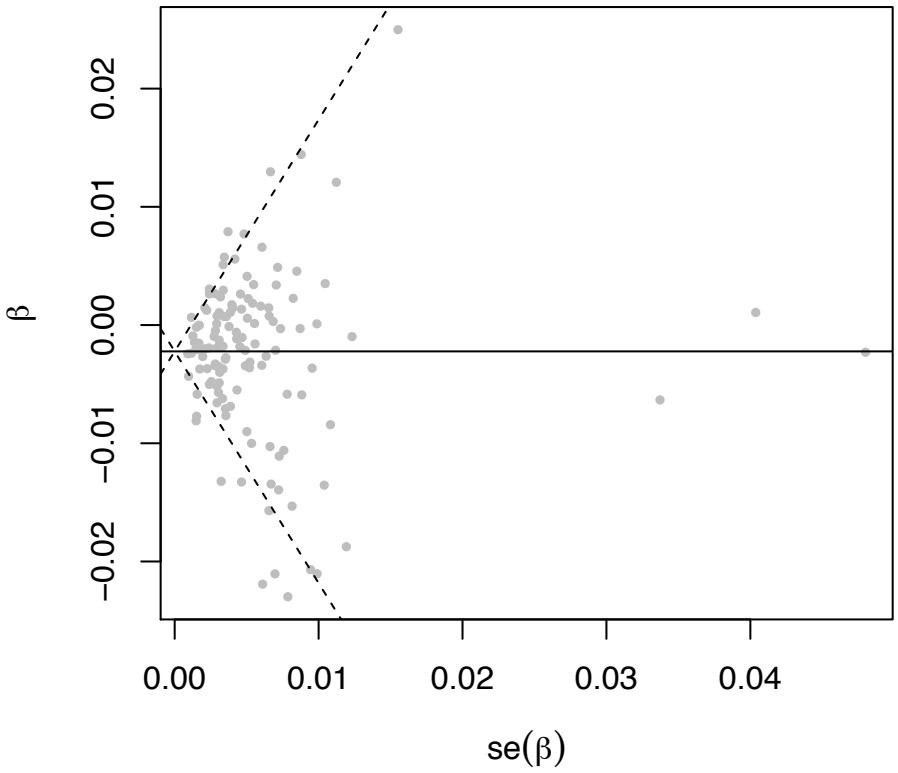
In conclusion, the analysis in this appendix suggests that at least in the context of our application, a number of macro level contextual factors, such as log GDP per capita, country-year and country-level mean education and age, and the region indicators as well as individual-level education and age variables are important drivers of treatment effect heterogeneity.

Figure 1: Funnel Plot of *Same-Sex* and *Having more children*



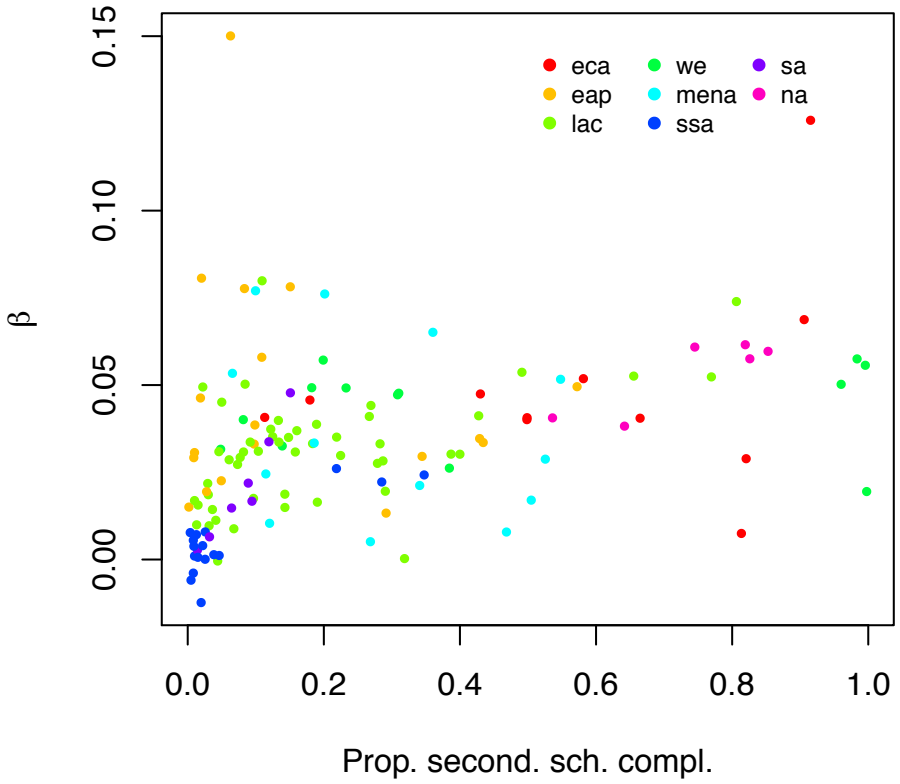
Notes: The funnel plot in this figure is based on data from 142 census samples. Source: Authors' calculations based on data from the *Integrated Public Use Microdata Series-International (IPUMS-I)*.

Figure 2: Funnel Plot of *Same-Sex* and *Being economically active*



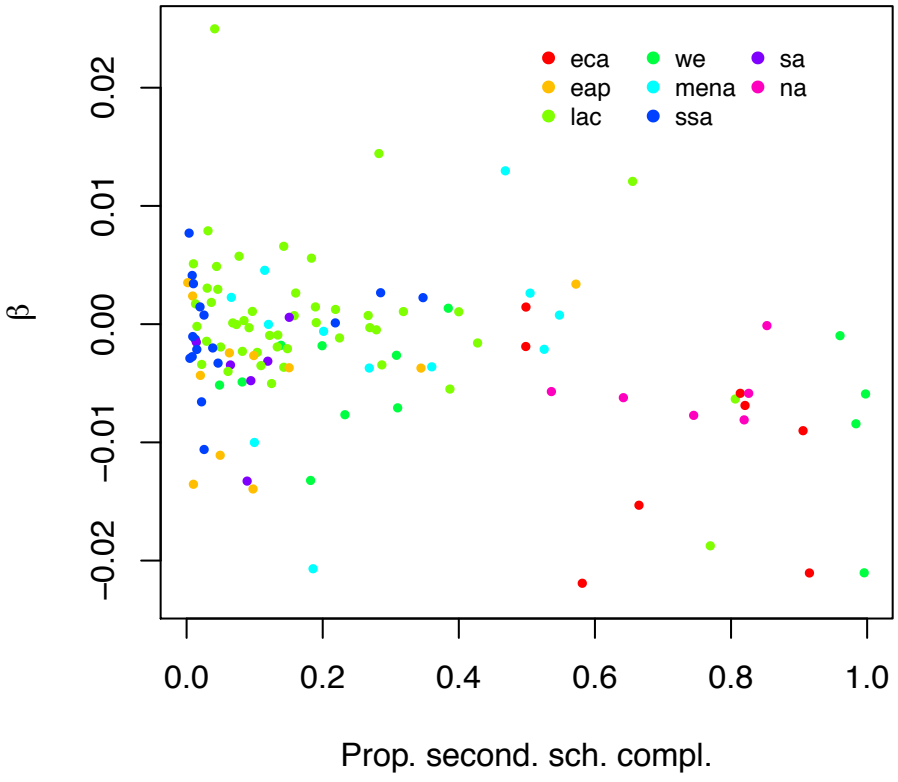
Notes: The funnel plot in this figure is based on data from 128 census samples. Source: Authors' calculations based on data from the *Integrated Public Use Microdata Series-International (IPUMS-I).*

Figure 3: Treatment effect heterogeneity of *Same-Sex on Having more children* by the proportion of women with a completed secondary education.



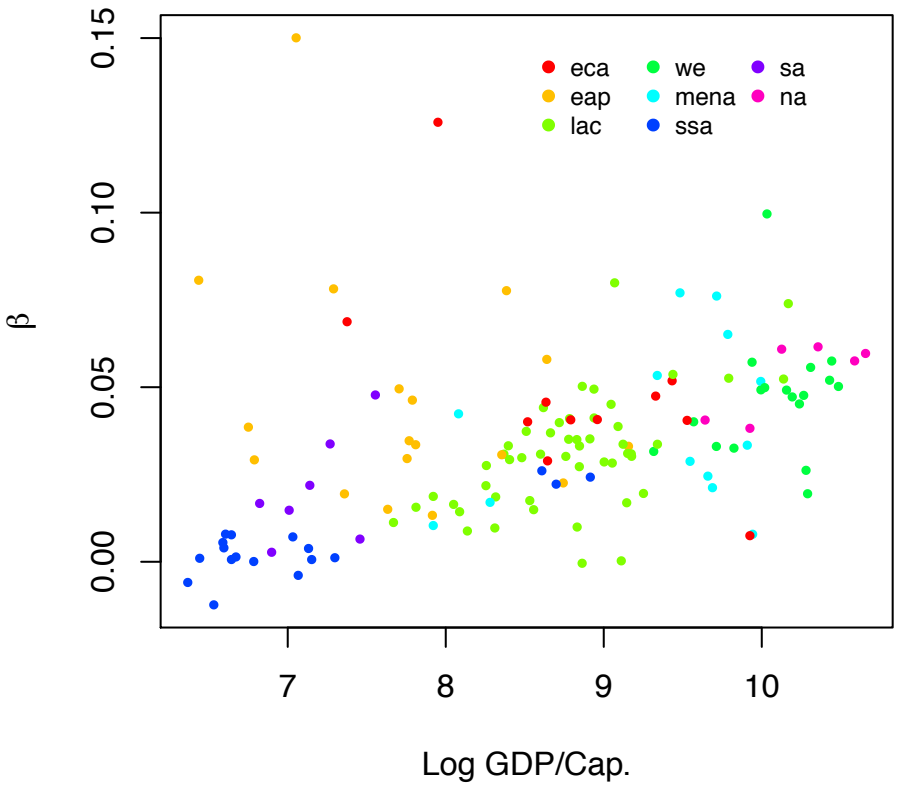
Notes: The graph plots the size of the treatment effect of *Same-Sex* on *Having more children* by the proportion of women with a completed secondary education based on data from 142 census samples. The graph also displays heterogeneity by geographic region. Source: Authors' calculations based on data from the *Integrated Public Use Microdata Series-International (IPUMS-I)*.

Figure 4: Treatment effect heterogeneity of *Same-Sex on Being economically active* by the proportion of women with a completed secondary education.



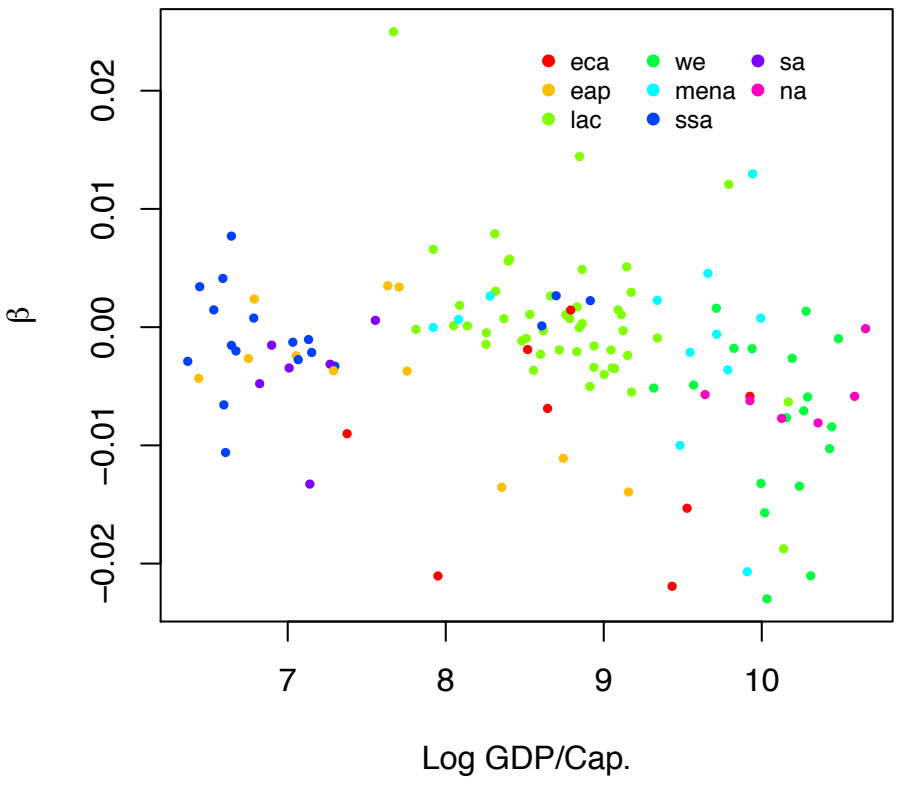
Notes: The graph plots the size of the treatment effect of *Same-Sex* on *Being economically active* by the proportion of women with a completed secondary education based on data from 142 census samples. The graph also displays heterogeneity by geographic region. Source: Authors' calculations based on data from the *Integrated Public Use Microdata Series-International (IPUMS-I)*.

Figure 5: Treatment effect heterogeneity of *Same-Sex on Having more children* by log GDP per capita



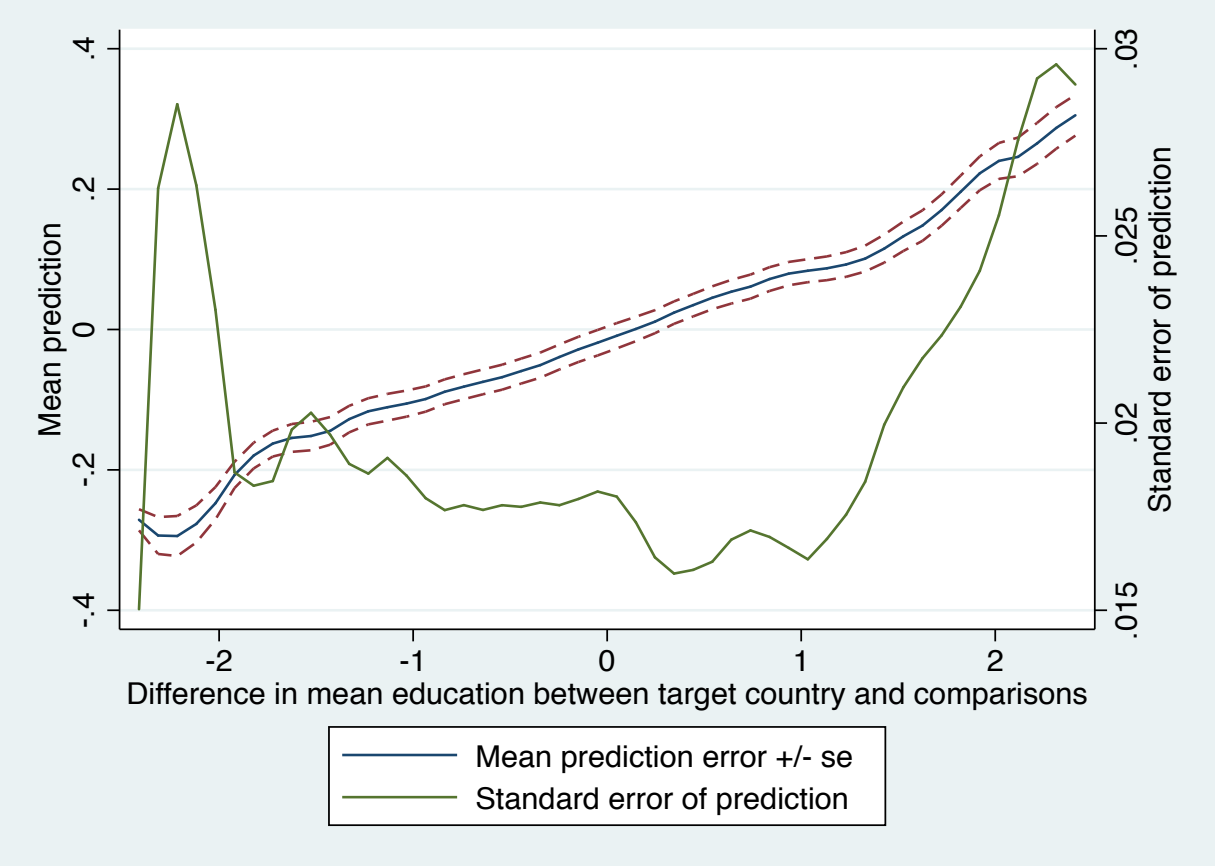
Notes: The graph plots the size of the treatment effect of *Same-Sex* on *Having more children* by log GDP per capita based on data from 142 census samples. The graph also displays heterogeneity by geographic region. Source: Authors' calculations based on data from the *Integrated Public Use Microdata Series-International (IPUMS-I)*.

Figure 6: Treatment effect heterogeneity of *Same-Sex on Being economically active* by log GDP per capita



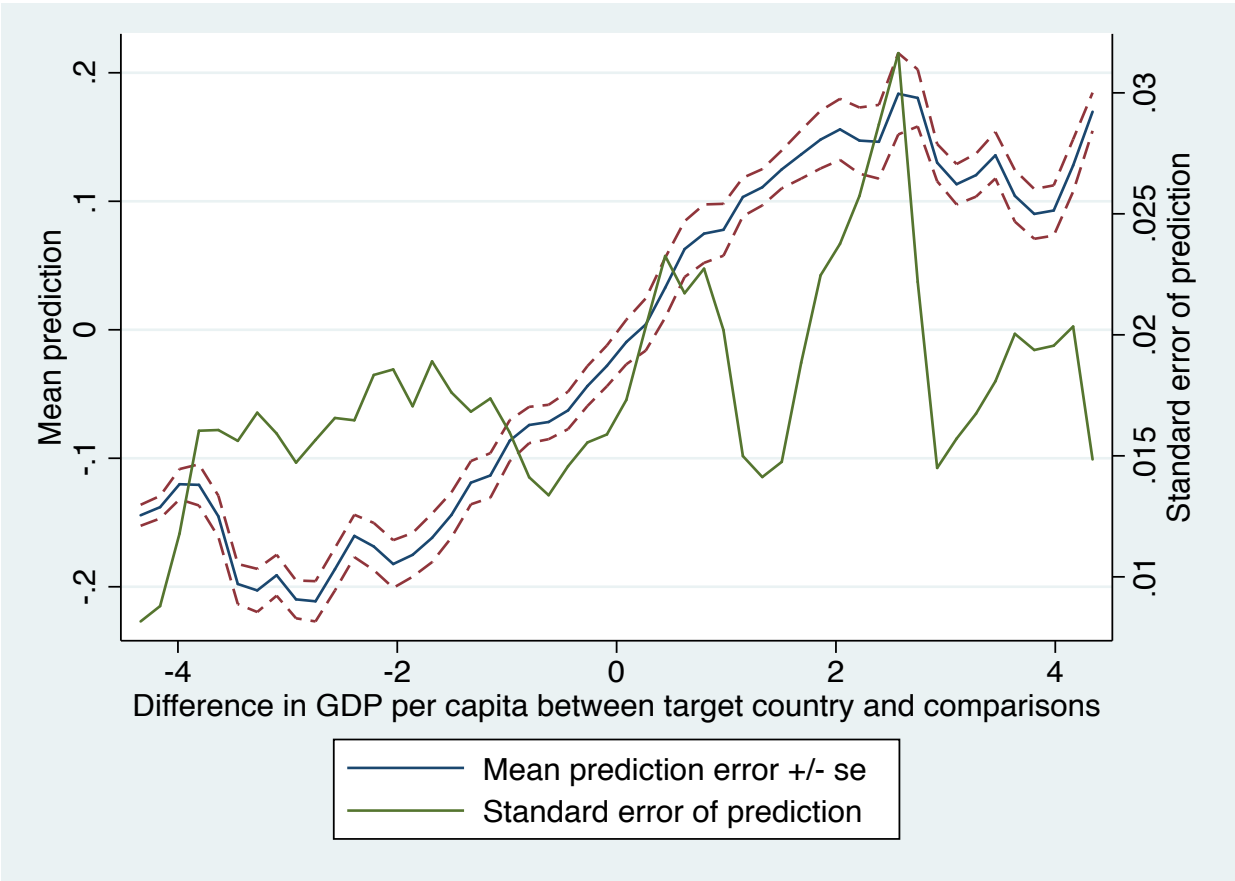
Notes: The graph plots the size of the treatment effect of *Same-Sex* on *Being economically active* by log GDP per capita based on data from 142 census samples. The graph also displays heterogeneity by geographic region. Source: Authors' calculations based on data from the *Integrated Public Use Microdata Series-International (IPUMS-I)*.

Figure 7: Unconditional external validity function: local linear regression of prediction error on standardized differences in education



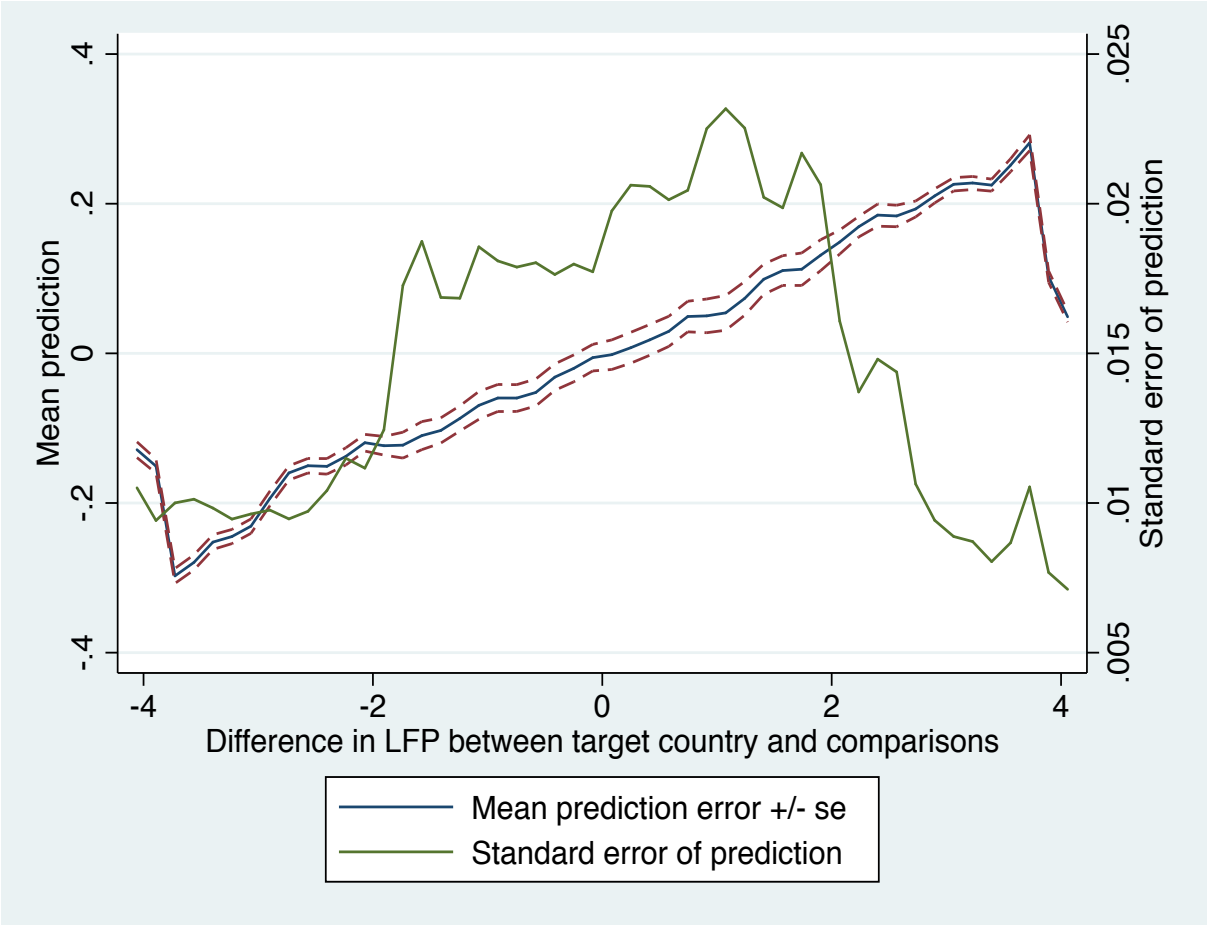
Notes: The graph plots the local polynomial regression of the dyadic prediction error and its standard against the standarized education difference between target and comparison country, where the education difference is standardized by its standard deviation (0.83). The variables are further described in Table 1. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

Figure 8: Unconditional external validity function: local linear regression of prediction error on standardized differences in log GDP per capita



Notes: The graph plots the local polynomial regression of the dyadic prediction error and its standard against the stanardized GDP difference between target and comparison country, where the GDP difference is standardized by its standard deviation ($9680). The variables are further described in Table 1. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

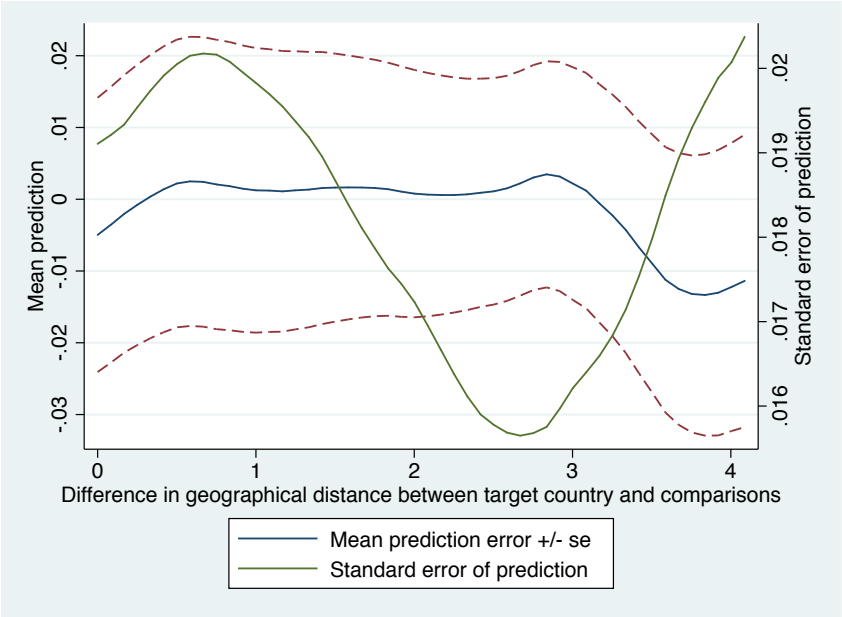
Figure 9: Unconditional external validity function: local linear regression of prediction error on standardized differences in women's labor force participation



Notes: The graph plots the local polynomial regression of the dyadic prediction error and its standard against the standardized labor force particiaption difference between target and comparison country, where the labor force participation difference is standardized by its standard deviation (0.21). The variables are further described in Table 1. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

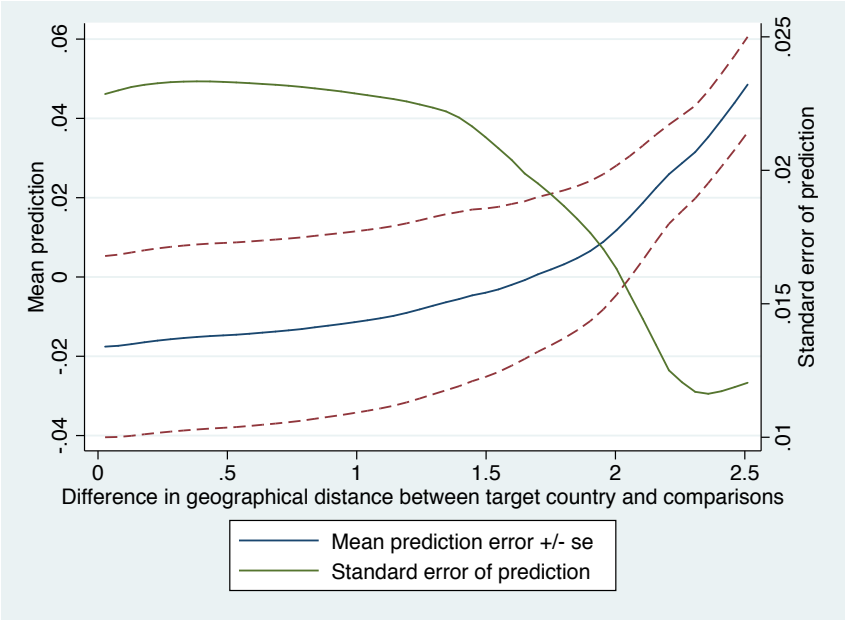
Figure 10a: Unconditional external validity function: local linear regression of prediction error on standardized geographical distance

Panel A: All country-year dyads



Notes: The graph plots the local polynomial regression of the dyadic prediction error and its standard against the standardized geographical distancebetween target and comparison country, where the geographical distance is standardized by its standard deviation (4800 km). The variables are further described in Table 1.
Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

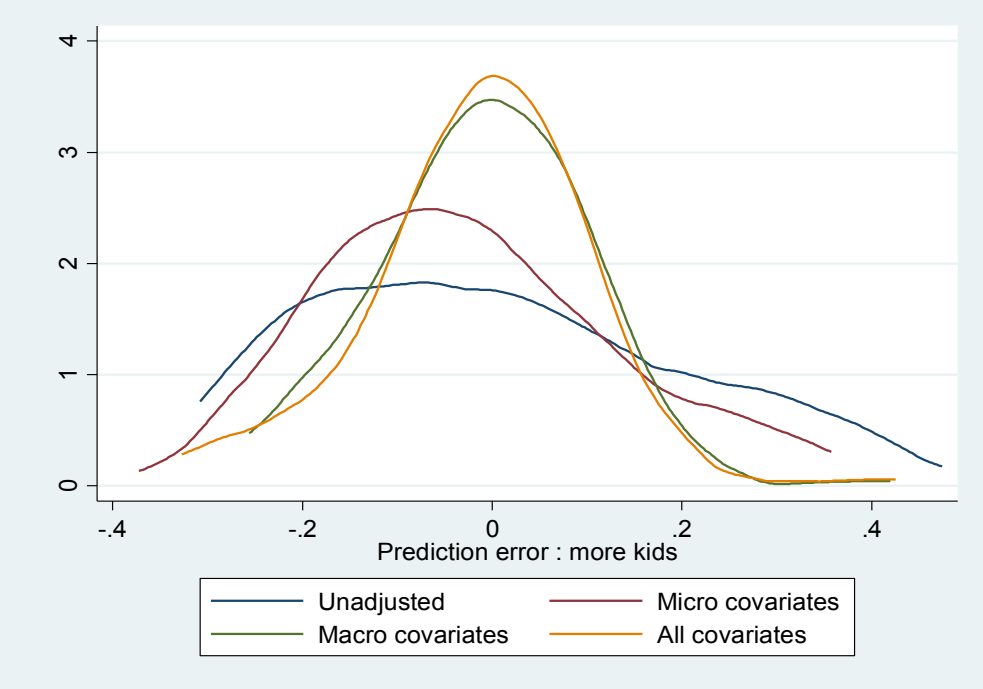
Panel B: All within-region country-year dyads



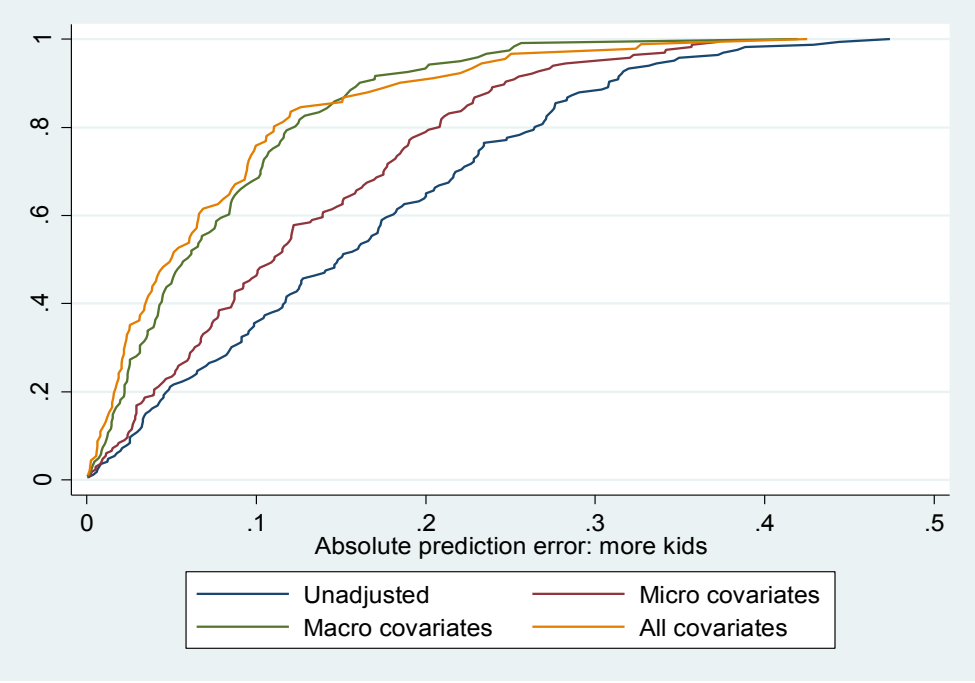
Notes: The graph plots the local polynomial regression of the dyadic prediction error and its standard against the standardized geographical distance between target and comparison country, for within-region dyads (where regions defined as North and South America, Europe, Asia, and Africa) and with the geographical distance is standardized by its standard deviation (4800 km). The variables are further described in Table 1.
Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

Figure 11: Individual versus macro covariates for *Having more children*

Panel A: Density estimate - prediction error



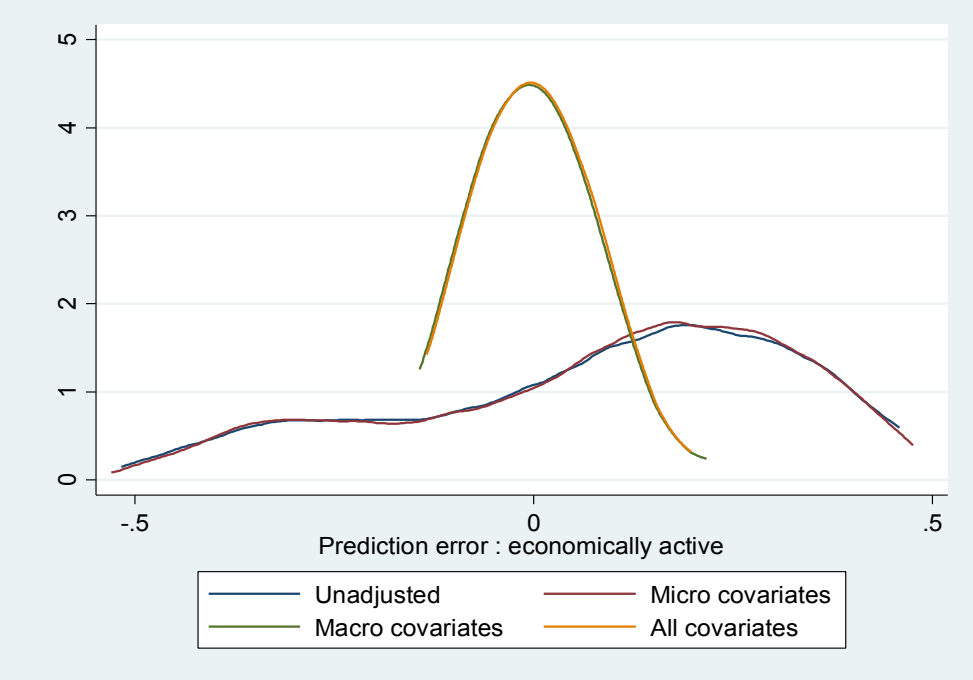
Panel B: CDF - absolute prediction error



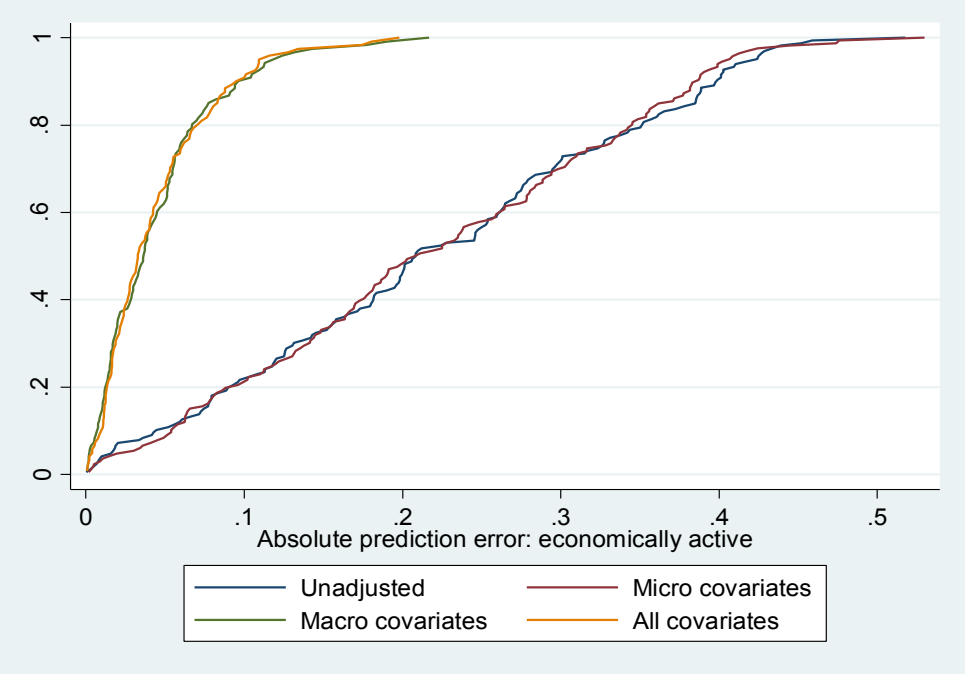
Notes: The graph plots the density estmates of the prediction error and CDF of the absolute prediction error based on the procedure described in Section 9 of the paper. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

Figure 12: Individual versus macro covariates for *Being economically active*

Panel A: Density estimate - prediction error

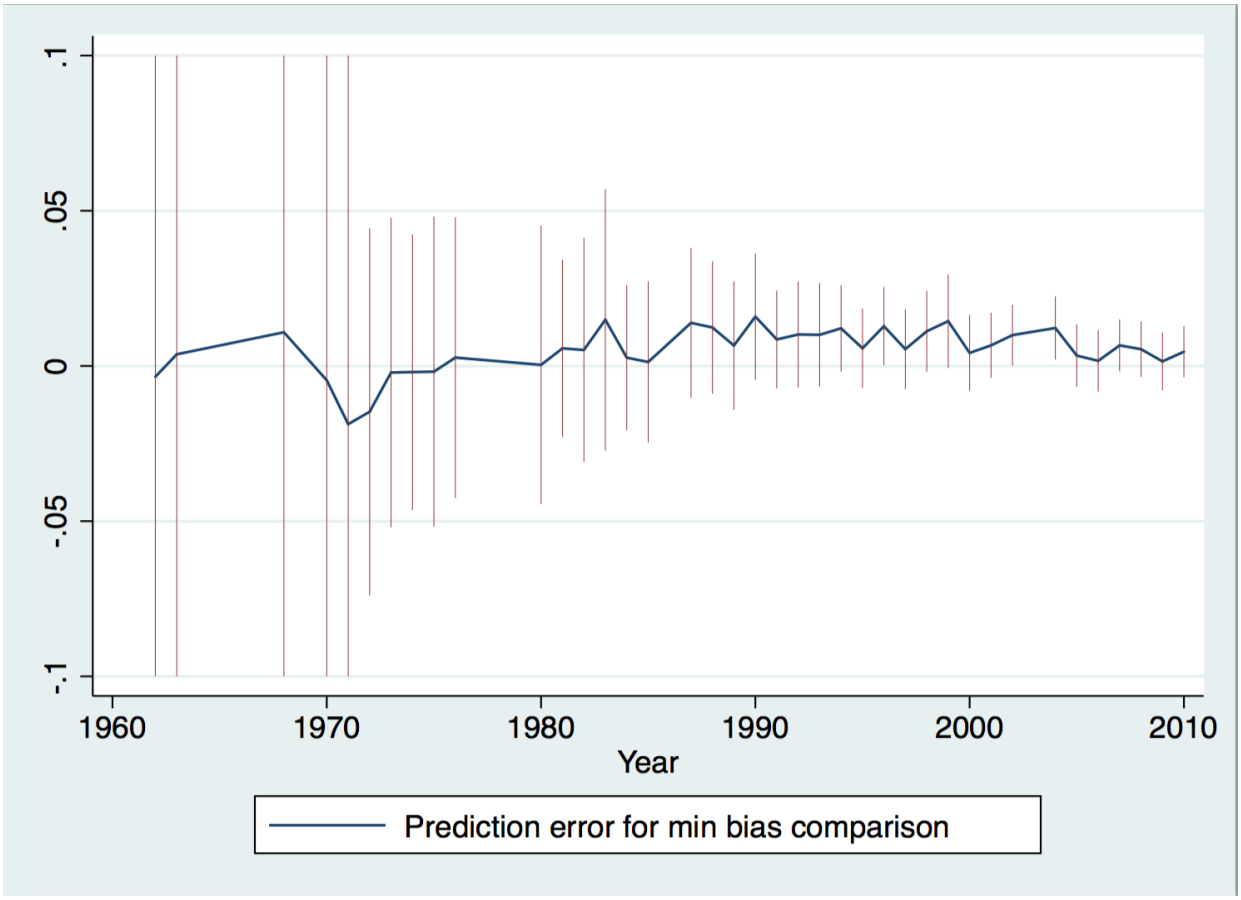


Panel B: CDF - absolute prediction error



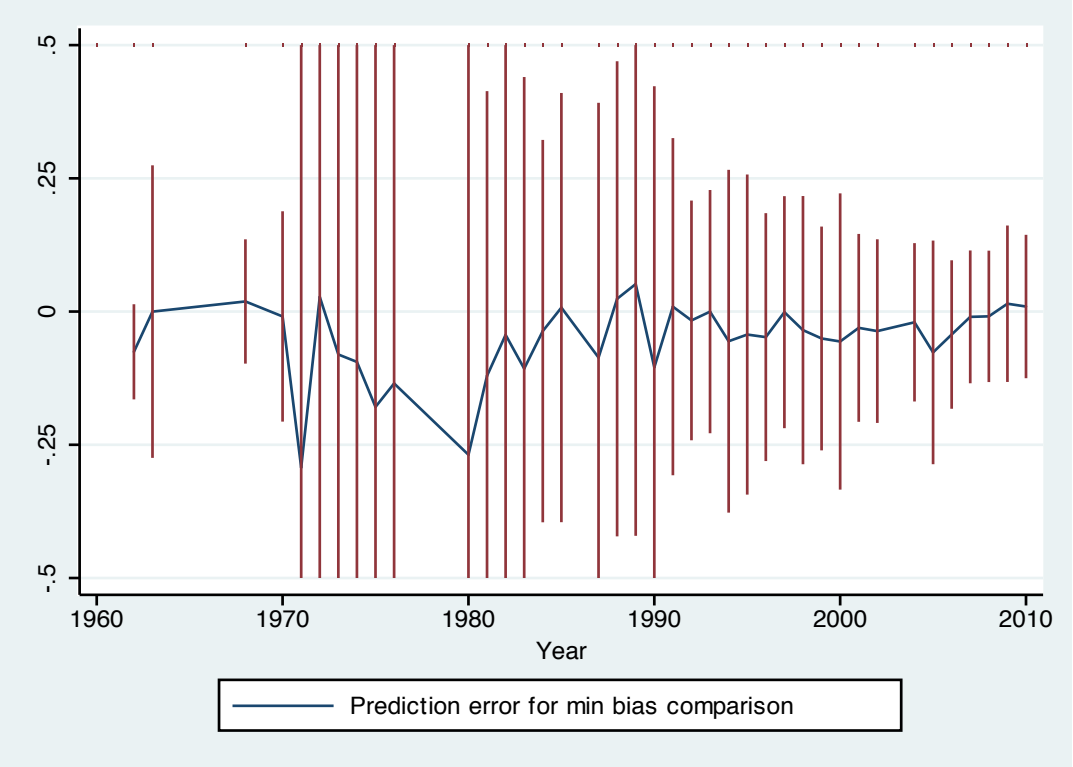
Notes: The graph plots the density estmates of the prediction error and CDF of the absoluteprediction error based on the procedure described in Section 9 of the paper. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

Figure 13: Prediction error over time of *Same-Sex* on *Having more children*



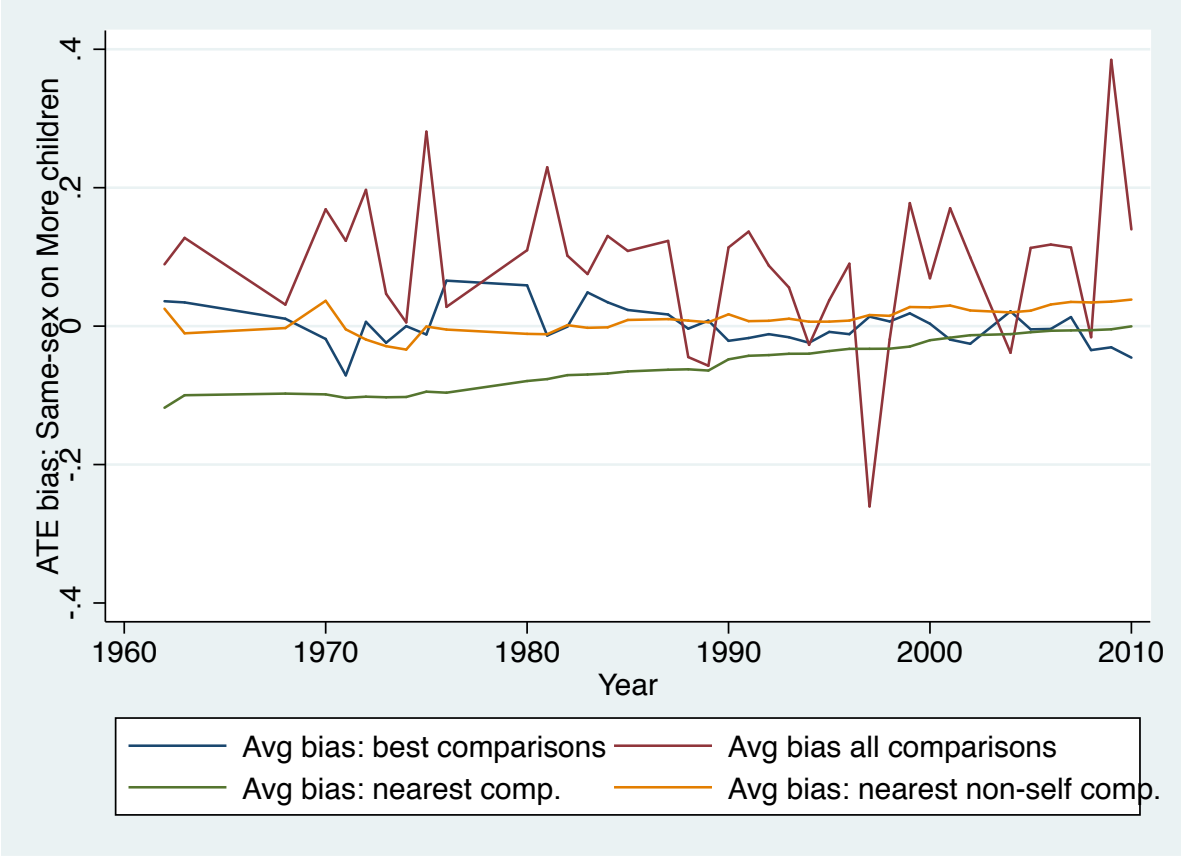
Notes: The graph plots the prediction error over time based on the procedure described in section 10 of the paper. The variable on the X-axis refers to the year when a census was taken. The variables are further described in Table 1. Source: Authors' calculations based on data from the *Integrated Public Use Microdata Series-International (IPUMS-I)*.

Figure 14: Prediction error over time of *Same-Sex* on *Being economically active*



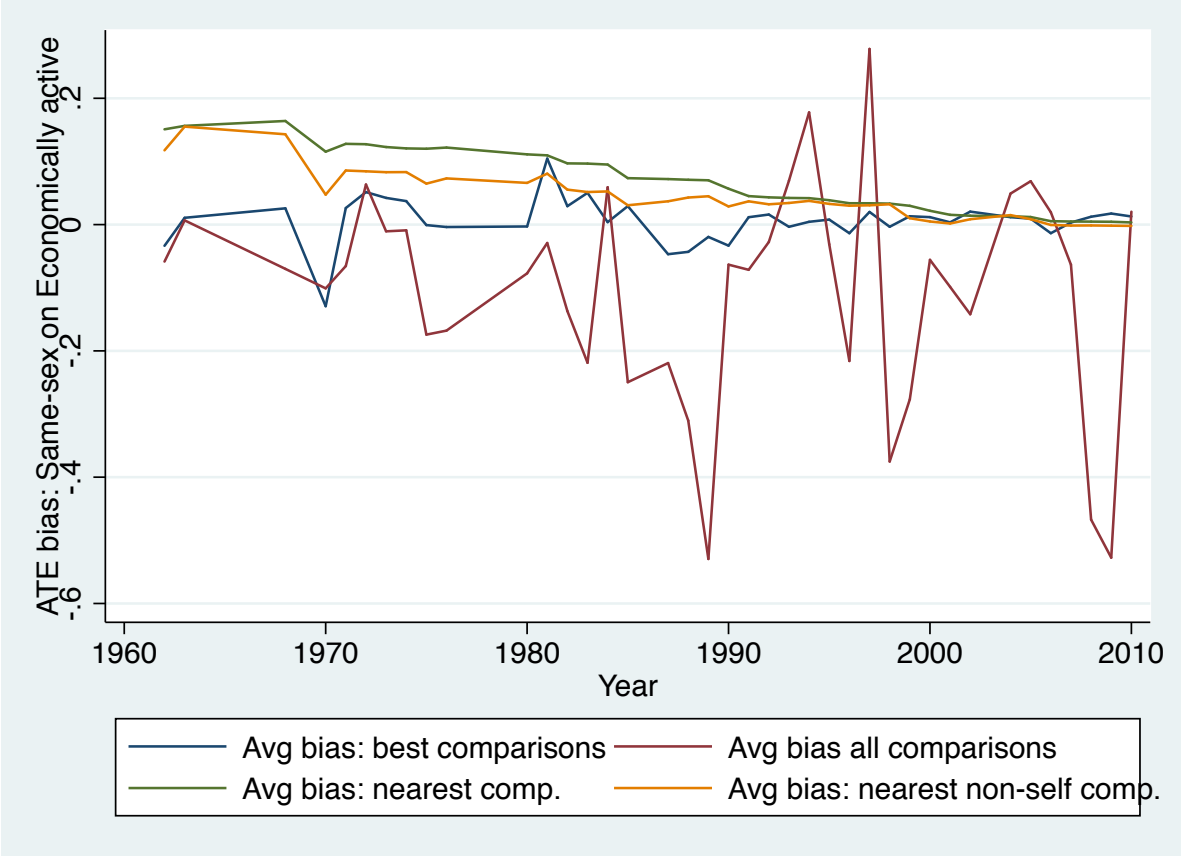
Notes: The graph plots the prediction error over time based on the procedure described in Section 10 of the paper. The variable on the X-axis refers to the year when a census was taken. The variables are further described in Table 1. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

Figure 15: Prediction error with different comparison groups of *Same-Sex* on *Having more children*



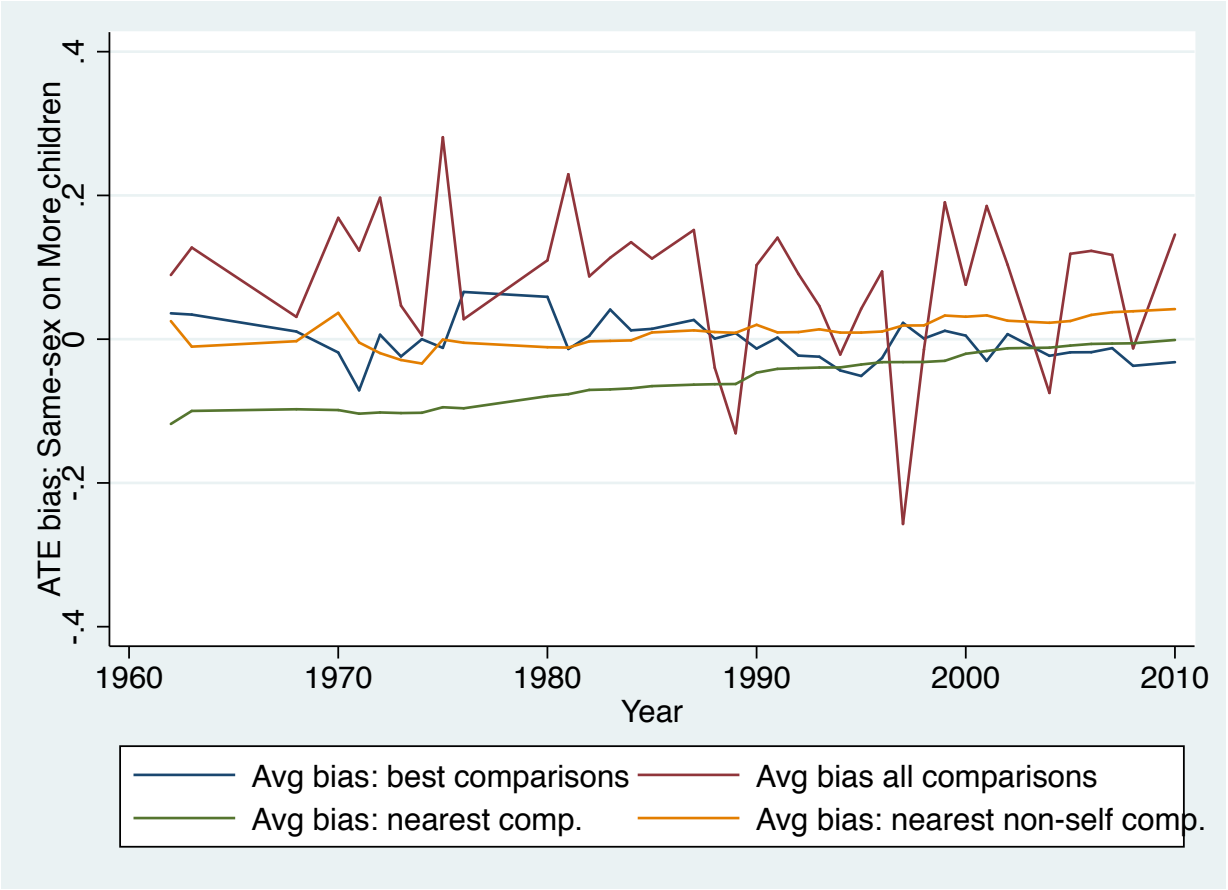
Notes: The graph plots the prediction error over time based on the procedure described in Section 10 of the paper. These four groups of comparison countries are: (1) all the available country years, (graphed as the red line), (2) the best comparison country-year as predicted by our model (graphed as the blue line), (3) the nearest country-year by distance excluding own-country comparisons (graphed orange line), and (4) the nearest country-year by distance, allowing own-country year comparisons. The variable on the X-axis refers to the year when a census was taken. The variables are further described in Table 1. Source: Authors' calculations based on data from the *Integrated Public Use Microdata Series-International (IPUMS-I)*.

Figure 16: Prediction error with different comparison groups of *Same-Sex* on *Being economically active*



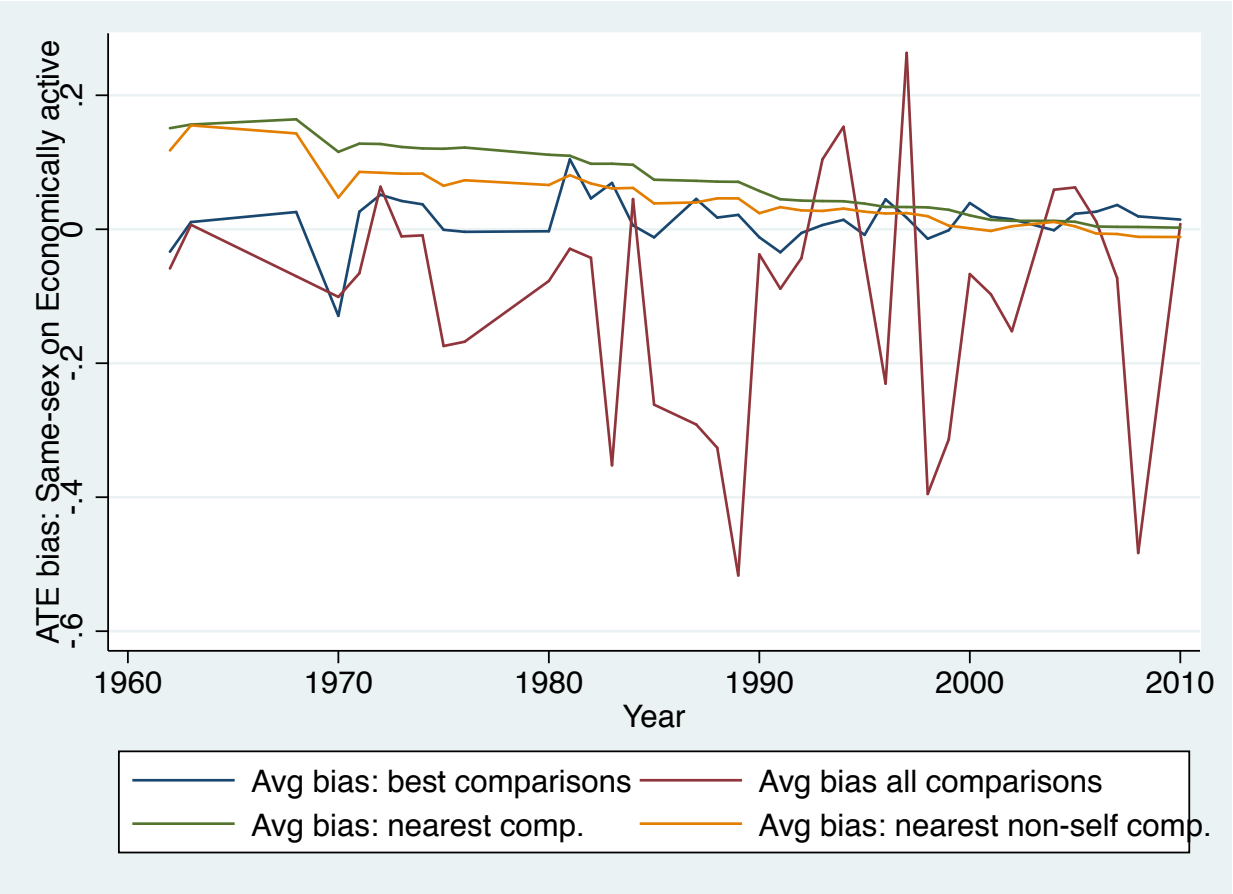
Notes: The graph plots the prediction error over time based on the procedure described in Section 10 of the paper. These four groups of comparison countries are: (1) all the available country years, (graphed as the red line), (2) the best comparison country-year as predicted by our model (graphed as the blue line), (3) the nearest country-year by distance excluding own-country comparisons (graphed orange line), and (4) the nearest country-year by distance, allowing own-country year comparisons. The variable on the X-axis refers to the year when a census was taken. The variables are further described in Table 1. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

Figure 17: Prediction error with different comparison groups of *Same-Sex* on *Having more children*, excluding sex-selecting countries



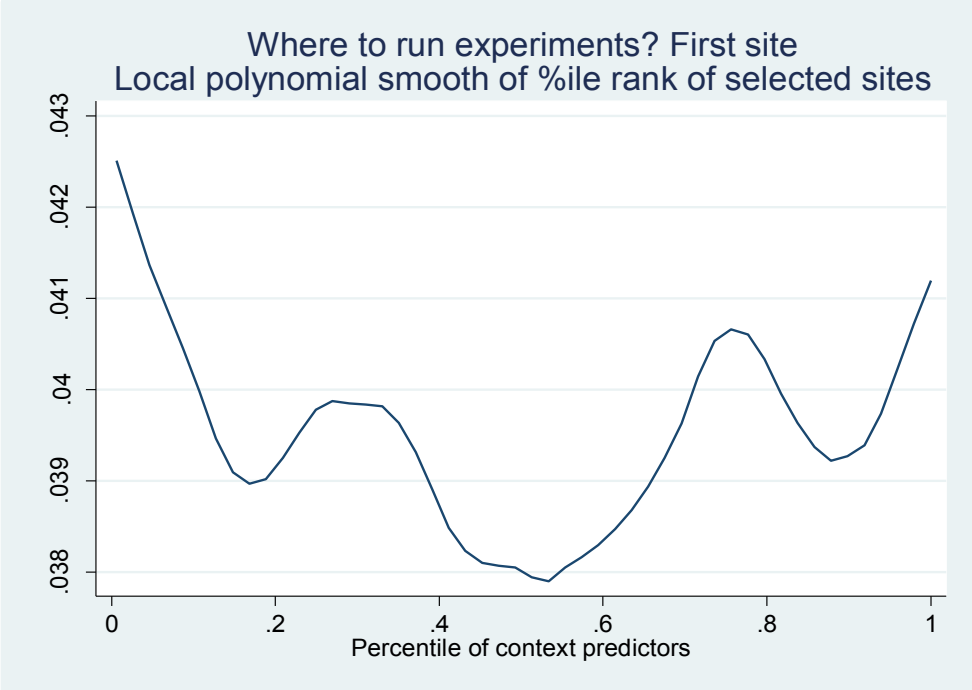
Notes: China, India, Nepal, and Vietnam are excluded from the analysis. The graph plots the prediction error over time based on the procedure described in Section 10 of the paper. These four groups of comparison countries are: (1) all the available country years, (graphed as the red line), (2) the best comparison country-year as predicted by our model (graphed as the blue line), (3) the nearest country-year by distance excluding own-country comparisons (graphed orange line), and (4) the nearest country-year by distance, allowing own-country year comparisons. The variable on the X-axis refers to the year when a census was taken. The variables are further described in Table 1. Source: Authors' calculations based on data from the *Integrated Public Use Microdata Series-International (IPUMS-I)*.

Figure 18: Prediction error with different comparison groups of *Same-Sex* on *Being economically active*, excluding sex-selecting countries



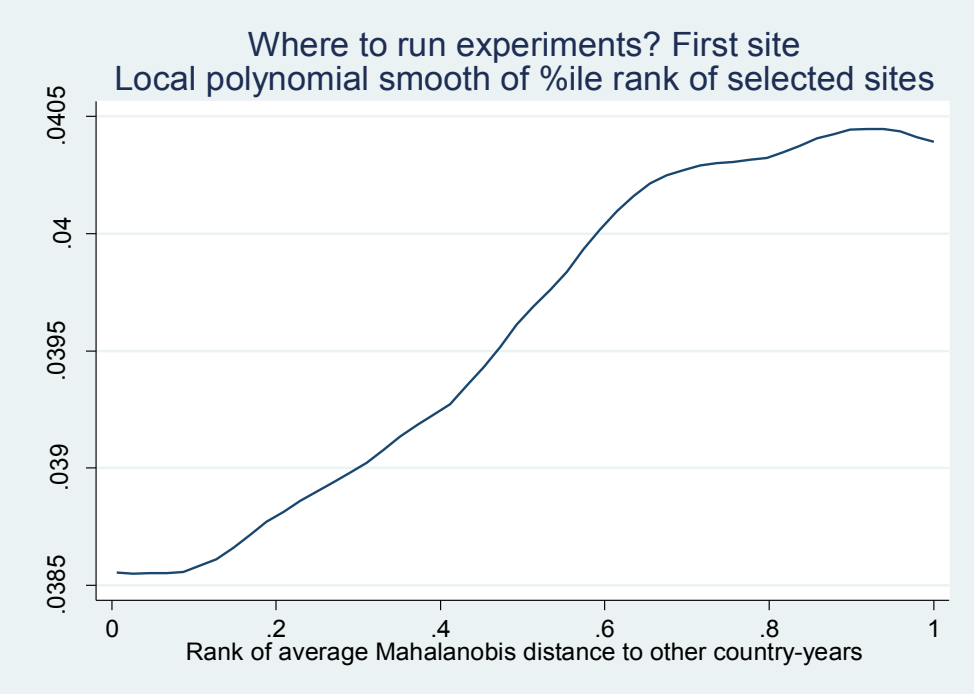
Notes: China, India, Nepal, and Vietnam are excluded from the analysis. The graph plots the prediction error over time based on the procedure described in Section 2 of the paper. These four groups of comparison countries are: (1) all the available country years, (graphed as the red line), (2) the best comparison country-year as predicted by our model (graphed as the blue line), (3) the nearest country-year by distance excluding own-country comparisons (graphed orange line), and (4) the nearest country-year by distance, allowing own-country year comparisons. The variable on the X-axis refers to the year when a census was taken. The variables are further described in Table 1. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

Figure 19: Mean prediction error on percentile of comparison country composite treatment-effect predictor, using one site to predict all others



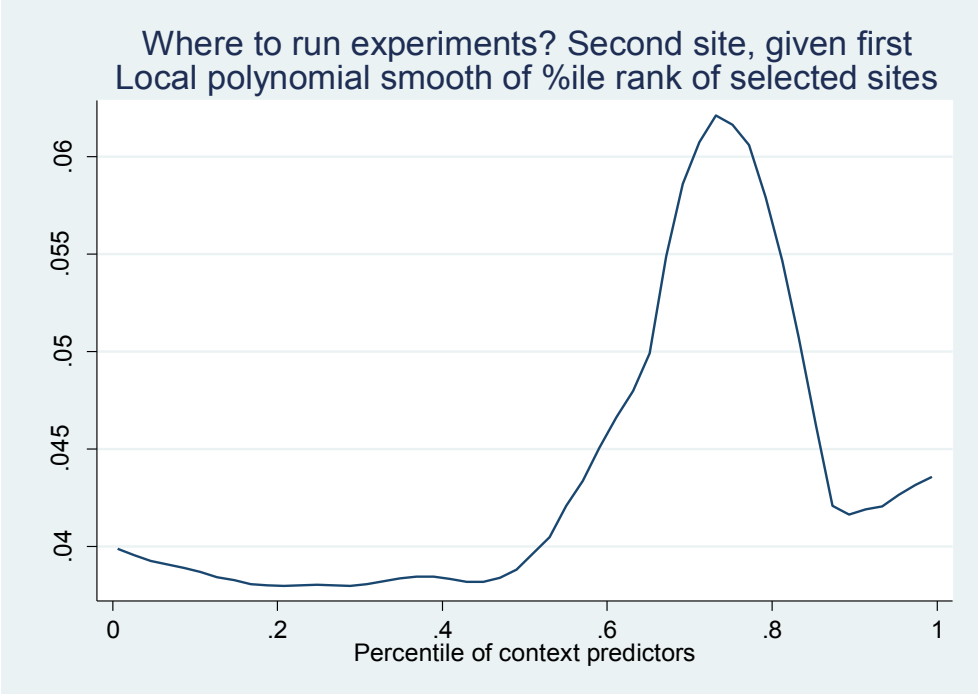
Notes: On the x-axis each country-year is ranked based on its percentile of a composite treatment effect predictor. The composite predictor is a weighted average country-year covariates weighted by their effect on the country-year treatment effect. The y-axis show the mean prediction error from using the site on the x-axis to predict all other country-years. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

Figure 20: Mean prediction error on average Mahalanobis distance of the comparison country-year to all target country-years



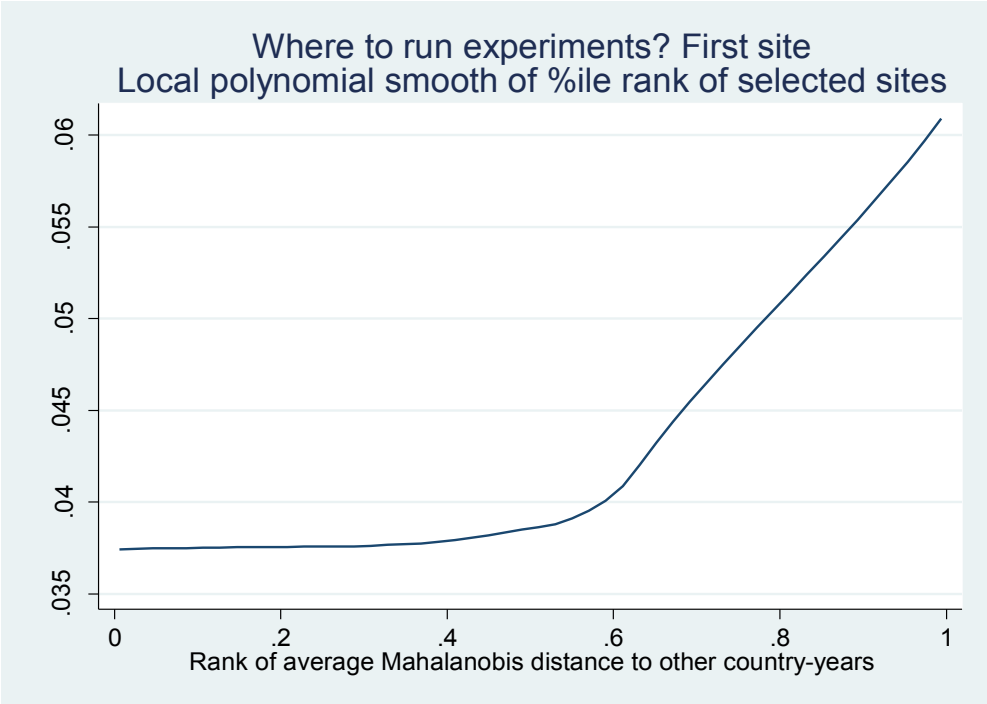
Notes: On the x-axis each country-year is ranked based on its avearge Mahalanobis distance to all other country-years. The y-axis show the mean prediction error from using the site on the x-axis to predict all other country-years. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

Figure 21: Mean prediction error, given the first comparison site, on percentile of composite treatment-effect predictor covariate, using two sites to predict the others



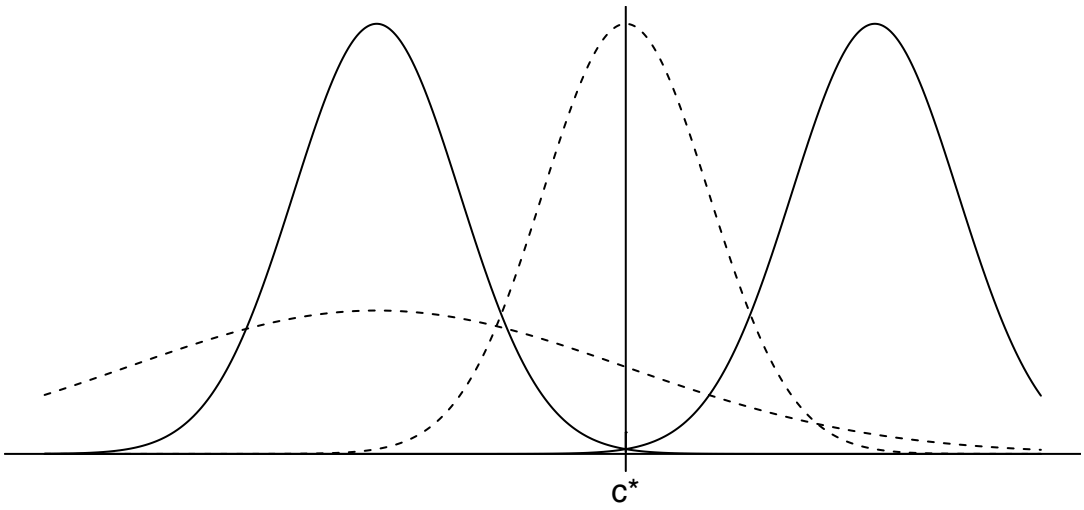
Notes: On the x-axis each country-year is ranked based on its percentile of a composite treatment effect predictor. The composite predictor is a weighted average country-year covariates weighted by their effect on the country-year treatment effect. The y-axis show the mean prediction error from using the site on the x-axis to predict all other country-years. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

Figure 22: Mean prediction error, given the first comparison site, on average Mahalanobis distance of the comparison country-year to all target country-years, using two sites to predict others



Notes: On the x-axis each country-year is ranked based on its avearge Mahalanobis distance to all other country-years. The y-axis show the mean prediction error from using the site on the x-axis in addition to the first selected comparison site to predict all other country-years. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

Figure 23: To experiment or extrapolate? A graphical illustration of the decision problem



Notes: Solid line = experiment not warranted. Dashed line = experiment warranted.

Table 1: Summary Statistics

	Mean	S.D.	Obs
Panel A: Individual level variables			
Had more children	0.57	0.50	12,516,425
Economically active	0.45	0.50	12,504,095
First two children are same sex	0.50	0.50	12,516,425
Age	30.1	3.56	12,516,425
Education (own)	1.89	0.84	12,516,425
Education (spouse)	2.04	0.97	12,516,425
Age at first marriage	20.69	3.11	12,516,425
Difference in first two kids boys vs girls	0.024	0.02	12,516,425
Year	1994	12.27	12,516,425
Panel B: Individual level variables (weighted by sampling weights)			
Had more children	0.60	0.49	549,696,649
Economically active	0.49	0.50	549,696,649
First two children are same sex	0.50	0.50	549,696,649
Age	30.0	3.58	549,696,649
Educaiton (own)	1.69	0.82	549,696,649
Educaiton (spouse)	1.95	0.91	549,696,649
Age at first marriage	20.54	2.96	549,696,649
Difference in first two kids boys vs girls	0.505	0.24	549,696,649
Year	1991	10.62	549,696,649
Panel C: Country level variables			
Real GDP per capita	9879	472	166
Education	1.91	0.56	169
Age	20.70	1.06	169
Labor force participation (women with one child)	0.51	0.21	169
Sex imbalance between boys and girls	0.02	0.02	169
Panel D: Dyadic differences between country pairs			
Age	0.98	0.73	14,196
Education (own)	0.63	0.46	14,196
Education (spouse)	0.58	0.42	14,196
Real GDP per capita	10117	9635	14,196
Year	14	10	14,196
Geographic distance (km)	8179	4809	14,196

Notes: Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

Table 2: Heterogeneity tests

Outcome	Effect specification	N*	Q-test statistic** (p-value)	wSF-test statistic*** (p-value)
More kids	Country-year	142	13,998 (<.0001)	0.9345 (<.0001)
	Country-year-ed. category	533	15,573 (<.0001)	0.9433 (<.0001)
Economically active	Country-year	128	224.26 (<.0001)	0.948 -0.0002
	Country-year-ed. category	477	586.26 (<.0001)	0.8592 (<.0001)

Notes: *Number of studies, which varies over the two outcomes because of incomplete data over available samples for the economically active indicator.
**Q test of effect homogeneity.
***Inverse-variance weighted Shapiro-Francia (wSF) test for normality of effect estimates. The test statistic is the squared correlation between the sample order statistics and the expected values of normal distribution order statistics.

Table 3: Bias regressions for *Having more children* - with covariates

Difference between country pairs in:	Absolute bias	Absolute bias	Absolute bias	Absolute bias	Absolute bias	Absolute bias	Absolute bias	Absolute bias	Absolute bias	Absolute bias	Absolute bias
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	Excluding sex selectors (11)
Education of mother	0.0484***								0.0331***	0.00956	0.0215
	(0.0108)								(0.0111)	(0.0155)	(0.0170)
Education of father		0.0617***								0.0334**	0.0199
		(0.0109)								(0.0137)	(0.0155)
Age of mother			0.0252						0.0118	0.0132	0.0180
			(0.0360)						(0.0359)	(0.0356)	(0.0366)
Census year				0.0149***					0.0123***	0.0117***	0.00998***
				(0.00390)					(0.00376)	(0.00366)	(0.00354)
log GDP per capita					0.0240***				0.00877	0.00906	0.0122*
					(0.00749)				(0.00665)	(0.00659)	(0.00722)
Sex ratio imbalance						-0.00137			-0.00708	-0.00635	0.00383
						(0.00460)			(0.00543)	(0.00535)	(0.00885)
Labor force participaiton							0.0362***		0.0237***	0.0239***	0.0222***
							(0.00553)		(0.00531)	(0.00527)	(0.00590)
Distance in KM								0.0650***	0.0410***	0.0387***	0.0407**
								(0.0154)	(0.0138)	(0.0144)	(0.0153)
Distance squared								-0.0173***	-0.0108***	-0.0101***	-0.0105***
								(0.00382)	(0.00338)	(0.00351)	(0.00373)
Constant	0.145***	0.144***	0.183***	0.166***	0.154***	0.184***	0.144***	0.140***	0.0809***	0.0794***	0.0723***
	(0.0124)	(0.0114)	(0.00662)	(0.0100)	(0.00964)	(0.00712)	(0.00696)	(0.0139)	(0.0171)	(0.0170)	(0.0178)
Observations	28,561	28,561	28,561	28,561	27,556	28,561	28,561	28,561	27,556	27,556	24,025
R-squared	0.037	0.038	0.003	0.009	0.029	0.000	0.037	0.018	0.083	0.085	0.091

Notes: The table shows bias regressions as described in Sections 3 and 9 of the paper. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I). Column 10 excludes China, India, Vietnam, and Nepal.

Table 4: Bias regressions for *Being economically active* - with covariates

Difference between country pairs in:	Absolute bias	Absolute bias	Absolute bias	Absolute bias	Absolute bias	Absolute bias	Absolute bias	Absolute bias	Absolute bias	Absolute bias	Absolute bias
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	Excluding sex selectors (11)
Education of mother	0.00262								-0.00280	-0.0203	-0.00177
	(0.00963)								(0.00760)	(0.0163)	(0.00789)
Education of father		0.00318								0.0253	
		(0.00943)								(0.0197)	
Age of mother			-0.0687**						-0.0458**	-0.0455**	-0.0446**
			(0.0325)						(0.0201)	(0.0196)	(0.0199)
Census year				0.0222***					-0.00236	-0.00275	-0.00439
				(0.00620)					(0.00551)	(0.00557)	(0.00518)
log GDP per capita					-0.00599				-0.0244***	-0.0240***	-0.0247***
					(0.00577)				(0.00387)	(0.00377)	(0.00346)
Sex ratio imbalance						0.0240*			0.0218	0.0221	0.0506***
						(0.0129)			(0.0142)	(0.0141)	(0.0115)
Labor force participaiton							0.175***		0.172***	0.173***	0.169***
							(0.0137)		(0.0106)	(0.0109)	(0.0106)
Distance in KM								0.105***	0.0374***	0.0357***	0.0434***
								(0.0246)	(0.0131)	(0.0131)	(0.0146)
Distance squared								-0.0149***	-0.00332	-0.00286	-0.00554
								(0.00530)	(0.00313)	(0.00321)	(0.00349)
Constant	0.230***	0.230***	0.228***	0.207***	0.238***	0.218***	0.0435***	0.116***	0.00956	0.00812	-0.00300
	(0.00738)	(0.00861)	(0.00857)	(0.00775)	(0.00744)	(0.0104)	(0.0132)	(0.0234)	(0.0240)	(0.0246)	(0.0230)
Observations	29,486	29,486	29,486	29,486	29,486	29,486	29,486	29,486	29,486	29,486	26,069
R-squared	0.000	0.000	0.013	0.010	0.001	0.005	0.502	0.081	0.549	0.550	0.541

Notes: The table shows bias regressions as described in Sections 3 and 9 of the paper. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

Table 5: To experiment or to extrapolate? Prediction interval estimates for effects on "more kids"

		Prediction Interval		In-sample			Prediction Interval		In-sample
Country	**Year**	*Lower bound*	*Upper bound*	**estimate**	**Country**	**Year**	*Lower bound*	*Upper bound*	**estimate**
Argentina	1980	-0.0160	0.0378	0.0412	Malaysia	1980	-0.1812	0.1927	-0.0110
Argentina	1991	-0.0430	0.1400	0.0427	Malaysia	1991	-0.1073	0.1740	-0.0105
Argentina	2001	-0.0052	0.1016	0.0217	Malaysia	2000	-0.3416	0.3896	0.0088
Bolivia	1992	-0.0067	0.0531	0.0097	Mali	1987	-0.1861	0.1159	0.0151
Bolivia	2001	-0.0149	0.0541	0.0082	Mali	1998	-0.2015	0.1798	-0.0036
Brazil	1980	-0.3850	0.4455	0.0222	Mexico	1990	0.0049	0.1016	0.0245
Brazil	1991	-0.0468	0.1271	0.0303	Mexico	1995	0.0106	0.0675	0.0467
Brazil	2000	0.0058	0.0695	0.0361	Mexico	2000	0.0121	0.0580	0.0332
Chile	1982	-0.2452	0.2944	0.0487	Nepal	2001	-0.0276	0.0271	0.0269
Chile	1992	0.0088	0.0922	0.0349	Panama	1980	-0.1796	0.2050	-0.0133
Chile	2002	0.0186	0.0862	0.0264	Panama	1990	-0.1393	0.2463	0.0439
Colombia	1985	-0.0565	0.0880	0.0406	Panama	2000	0.0099	0.0725	0.0187
Colombia	1993	-0.0074	0.0953	0.0343	Peru	1993	-0.0058	0.0563	0.0183
Colombia	2005	-0.0154	0.0956	0.0404	Peru	2007	-0.0002	0.0750	0.0435
Costa Rica	1984	-0.0083	0.1406	0.0195	Philippines	1990	-0.0510	0.0859	0.0257
Costa Rica	2000	0.0183	0.0817	0.0029	Portugal	1981	-0.0827	0.0355	0.0391
Ecuador	1974	-0.0432	0.0280	0.0274	Portugal	1991	-0.0562	0.1696	0.0339
Ecuador	1982	-0.0384	0.0660	0.0261	Portugal	2001	0.0184	0.0887	0.0605
Ecuador	1990	0.0025	0.0628	0.0128	Rwanda	2002	-0.0567	0.0916	0.0403
Ecuador	2001	-0.0188	0.0673	0.0211	Senegal	1988	-0.0837	0.0688	0.0038
France	1975	-0.0409	0.0712	0.0316	Senegal	2002	-0.0909	0.0943	-0.0150
France	1982	-0.1378	0.2024	0.0313	South Africa	1996	-0.0326	0.0895	0.0244
France	1990	0.0438	0.1066	0.0380	South Africa	2001	-0.0121	0.0785	0.0209
France	1999	-0.1220	0.2424	0.0394	South Africa	2007	-0.0487	0.1319	0.0139
Ghana	2000	-0.0365	0.0460	0.0046	Spain	1991	-0.0280	0.1343	0.0629
Greece	1981	-0.0173	0.0970	0.0676	Spain	2001	-0.0394	0.1548	0.0300
Greece	1991	0.0277	0.1201	0.0585	Switzerland	1980	0.0230	0.1502	0.0554
Greece	2001	-0.0221	0.1659	0.0546	Switzerland	1990	-0.0957	0.2861	0.0603
Guinea	1983	-0.4256	0.3858	0.0209	Switzerland	2000	0.0391	0.1420	0.0416
Guinea	1996	-0.0957	0.0808	-0.0131	Tanzania	1988	-0.2844	0.4007	-0.0077
India	1987	-0.4052	0.4404	0.0290	Tanzania	2002	-0.0511	0.0650	0.0089
India	1993	-0.1041	0.1516	0.0300	Uganda	1991	-0.0622	0.0601	0.0099
India	1999	-0.0725	0.1164	0.0333	Uganda	2002	-0.0511	0.0460	0.0050
Iraq	1997	-0.0206	0.0574	0.0113	United States	1980	-0.0055	0.0775	0.0609
Israel	1995	-0.0182	0.1257	0.0002	United States	1990	0.0320	0.1023	0.0647
Italy	2001	-0.0092	0.1582	0.0273	United States	2000	-0.0347	0.1904	0.0598
Jordan	2004	-0.0224	0.1187	0.0203	United States	2005	0.0382	0.1201	0.0570
Kenya	1989	-0.1169	0.1254	0.0002	Venezuela	1981	-0.0171	0.0684	0.0413
Kenya	1999	-0.0616	0.0561	0.0037	Venezuela	1990	-0.1013	0.1939	0.0236
					Venezuela	2001	0.0124	0.0755	0.0852

Notes: Prediction interval estimates were produced from least squares estimates of regression models for conditional means and variances, using micro-level data from one-percent extracts of the census samples and country-year level covariates from the Penn World Tables. Micro-level covariate include gender of first born child, age of mother and spouse, age of first and second child, whether first born were twins, and educational attainment of mother and spouse. Country-year level covariates include population density, log real GDP/per capita, government spending share of real GDP/capita, ethnic fractionalization index, female labor force participation rate, and year.

Appendix Table 1: Treatment effects and standard errors by country-year

Country	Year of census	Treatment effect for Having more kids	Standard error for Having more kids	Treatment effect for Economically active	Standard error for Economically active
Argentina	1970	0.0347	0.0213	0.0048	0.0159
Argentina	1980	0.0412	0.0080	-0.0033	0.0065
Argentina	1991	0.0427	0.0065	-0.0004	0.0069
Argentina	2001	0.0217	0.0095	-0.0008	0.0096
Armenia	2001	0.1222	0.0207	-0.0157	0.0239
Austria	1971	0.0369	0.0171	-0.0031	0.0170
Austria	1981	0.0531	0.0174	-0.0258	0.0194
Austria	1991	0.0364	0.0172	-0.0043	0.0200
Austria	2001	0.0297	0.0186	-0.0371	0.0219
Belarus	1999	0.0228	0.0118	-0.0194	0.0149
Bolivia	1976	0.0208	0.0172	-0.0221	0.0145
Bolivia	1992	0.0097	0.0149	-0.0046	0.0174
Bolivia	2001	0.0082	0.0146	-0.0127	0.0165
Brazil	1960	0.0135	0.0065	0.0018	0.0039
Brazil	1970	0.0145	0.0052	-0.0009	0.0036
Brazil	1980	0.0222	0.0050	0.0049	0.0044
Brazil	1991	0.0303	0.0043	-0.0023	0.0042
Brazil	2000	0.0361	0.0044	-0.0020	0.0046
Cambodia	1998	0.0311	0.0102	0.0018	0.0101
Chile	1970	0.0410	0.0131	-0.0041	0.0095
Chile	1982	0.0487	0.0125	0.0041	0.0093
Chile	1992	0.0349	0.0112	-0.0139	0.0091
Chile	2002	0.0264	0.0128	-0.0057	0.0125
China	1982	0.0671	0.0035	-0.0032	0.0028
China	1990	0.1243	0.0035	-0.0013	0.0026
Colombia	1973	0.0113	0.0082	-0.0056	0.0060
Colombia	1985	0.0406	0.0077	-0.0098	0.0079
Colombia	1993	0.0343	0.0074	0.0004	0.0069
Colombia	2005	0.0404	0.0074	0.0063	0.0062
Costa Rica	1973	-0.0337	0.0266	-0.0042	0.0203
Costa Rica	1984	0.0195	0.0244	-0.0193	0.0183
Costa Rica	2000	0.0029	0.0219	0.0193	0.0186
Cuba	2002	0.0567	0.0132	-0.0107	0.0164
Ecuador	1974	0.0274	0.0143	0.0089	0.0107
Ecuador	1982	0.0261	0.0128	0.0019	0.0108
Ecuador	1990	0.0128	0.0122	0.0104	0.0117
Ecuador	2001	0.0211	0.0125	0.0039	0.0123

Appendix Table 1 continued: Treatment effects and standard errors by country-year

Country	Year of census	Treatment effect for Having more kids	Standard error for Having more kids	Treatment effect for Economically active	Standard error for Economically active
Egypt	1996	0.0403	0.0041	-0.0040	0.0032
France	1962	0.0259	0.0099	-0.0012	0.0083
France	1968	0.0319	0.0097	0.0092	0.0088
France	1975	0.0316	0.0090	0.0073	0.0094
France	1982	0.0313	0.0085	-0.0026	0.0093
France	1990	0.0380	0.0101	0.0044	0.0110
France	1999	0.0394	0.0106	-0.0123	0.0121
Ghana	2000	0.0046	0.0108	-0.0067	0.0100
Greece	1971	0.0519	0.0139	-0.0172	0.0142
Greece	1981	0.0676	0.0125	-0.0061	0.0119
Greece	1991	0.0585	0.0127	0.0131	0.0146
Greece	2001	0.0546	0.0145	0.0168	0.0188
Guinea	1983	0.0209	0.0190	-0.0122	0.0211
Guinea	1996	-0.0131	0.0133	-0.0207	0.0147
Hungary	1970	0.0561	0.0187	NA	NA
Hungary	1980	0.0481	0.0155	NA	NA
Hungary	1990	0.0370	0.0165	-0.0355	0.0194
Hungary	2001	0.0176	0.0223	-0.0308	0.0253
India	1983	0.0126	0.0131	0.0263	0.0142
India	1987	0.0290	0.0130	-0.0349	0.0134
India	1993	0.0300	0.0143	-0.0204	0.0151
India	1999	0.0333	0.0143	-0.0256	0.0146
Iraq	1997	0.0113	0.0073	0.0043	0.0050
Israel	1972	0.0345	0.0224	-0.0021	0.0217
Israel	1983	0.0097	0.0190	NA	NA
Israel	1995	0.0002	0.0196	0.0154	0.0211
Italy	2001	0.0273	0.0107	-0.0090	0.0143
Jordan	2004	0.0203	0.0137	0.0102	0.0104
Kenya	1989	0.0002	0.0098	0.0185	0.0112
Kenya	1999	0.0037	0.0095	-0.0097	0.0101
Kyrgyz Republic	1999	0.0607	0.0162	0.0039	0.0181
Malaysia	1970	-0.0173	0.0237	-0.0114	0.0308
Malaysia	1980	-0.0110	0.0257	-0.0503	0.0286
Malaysia	1991	-0.0105	0.0192	-0.0047	0.0200
Malaysia	2000	0.0088	0.0190	-0.0226	0.0200
Mali	1987	0.0151	0.0129	-0.0224	0.0155
Mali	1998	-0.0036	0.0111	0.0143	0.0135

Appendix Table 1 continued: Treatment effects and standard errors by country-year

Country	Year of census	Treatment effect for Having more kids	Standard error for Having more kids	Treatment effect for Economically active	Standard error for Economically active
Mexico	1970	0.0078	0.0139	0.0079	0.0099
Mexico	1990	0.0245	0.0040	-0.0063	0.0032
Mexico	1995	0.0467	0.0196	-0.0054	0.0209
Mexico	2000	0.0332	0.0037	-0.0073	0.0035
Mongolia	1989	0.0449	0.0230	NA	NA
Mongolia	2000	0.0720	0.0243	0.0238	0.0268
Nepal	2001	0.0269	0.0066	-0.0041	0.0075
Pakistan	1973	0.0127	0.0095	-0.0030	0.0042
Pakistan	1998	0.0117	0.0029	NA	NA
Palestine	1997	0.0142	0.0167	0.0019	0.0101
Panama	1960	-0.0416	0.0506	0.0459	0.0435
Panama	1970	-0.0100	0.0288	0.0515	0.0263
Panama	1980	-0.0133	0.0265	-0.0090	0.0270
Panama	1990	0.0439	0.0268	-0.0146	0.0250
Panama	2000	0.0187	0.0261	0.0211	0.0241
Peru	1993	0.0183	0.0085	0.0064	0.0078
Peru	2007	0.0435	0.0089	0.0082	0.0089
Philippines	1990	0.0257	0.0045	-0.0093	0.0047
Philippines	1995	0.0372	0.0044	NA	NA
Philippines	2000	0.0287	0.0045	NA	NA
Portugal	1981	0.0391	0.0200	0.0358	0.0228
Portugal	1991	0.0339	0.0203	0.0048	0.0248
Portugal	2001	0.0605	0.0230	-0.0177	0.0283
Puerto Rico	1970	0.2339	0.0724	NA	NA
Puerto Rico	1980	0.0599	0.0316	NA	NA
Puerto Rico	1990	0.0370	0.0331	-0.0288	0.0334
Puerto Rico	2000	0.0801	0.0362	0.0129	0.0377
Puerto Rico	2005	NA	NA	NA	NA
Romania	1977	0.0502	0.0097	NA	NA
Romania	1992	0.0284	0.0094	-0.0103	0.0093
Romania	2002	0.0403	0.0100	0.0161	0.0126
Rwanda	1991	0.0014	0.0120	-0.0081	0.0050
Rwanda	2002	-0.0019	0.0136	0.0100	0.0102
Saint Lucia	1980	NA	NA	NA	NA
Saint Lucia	1991	NA	NA	NA	NA
Senegal	1988	0.0038	0.0124	-0.0205	0.0131
Senegal	2002	-0.0150	0.0124	0.0150	0.0137

Appendix Table 1 continued: Treatment effects and standard errors by country-year

Country	Year of census	Treatment effect for Having more kids	Standard error for Having more kids	Treatment effect for Economically active	Standard error for Economically active
Slovenia	2002	0.0161	0.0294	0.0254	0.0372
South Africa	1996	0.0244	0.0094	0.0010	0.0098
South Africa	2001	0.0209	0.0096	-0.0011	0.0097
South Africa	2007	0.0139	0.0216	-0.0133	0.0231
Spain	1991	0.0629	0.0106	-0.0050	0.0115
Spain	2001	0.0300	0.0128	0.0094	0.0174
Switzerland	1970	0.0299	0.0270	0.0068	0.0239
Switzerland	1980	0.0554	0.0244	-0.0246	0.0263
Switzerland	1990	0.0603	0.0268	-0.0204	0.0295
Switzerland	2000	0.0416	0.0291	-0.0508	0.0357
Tanzania	1988	-0.0077	0.0077	0.0077	0.0063
Tanzania	2002	0.0089	0.0063	-0.0192	0.0063
Thailand	1970	0.0129	0.0125	NA	NA
Thailand	1980	0.0694	0.0188	NA	NA
Thailand	1990	0.0705	0.0189	NA	NA
Thailand	2000	0.0543	0.0165	NA	NA
Uganda	1991	0.0099	0.0088	0.0024	0.0104
Uganda	2002	0.0050	0.0066	0.0073	0.0086
United Kingdom	1991	0.0646	0.0212	-0.0497	0.0239
United States	1960	0.0384	0.0098	0.0024	0.0083
United States	1970	0.0462	0.0095	0.0029	0.0095
United States	1980	0.0609	0.0043	-0.0116	0.0047
United States	1990	0.0647	0.0044	-0.0144	0.0048
United States	2000	0.0598	0.0048	0.0055	0.0052
United States	2005	0.0570	0.0116	-0.0035	0.0129
Venezuela	1971	0.0206	0.0107	0.0052	0.0091
Venezuela	1981	0.0413	0.0101	-0.0128	0.0093
Venezuela	1990	0.0236	0.0093	-0.0018	0.0080
Venezuela	2001	0.0852	0.0093	-0.0121	0.0090
Vietnam	1989	0.0300	0.0065	0.0042	0.0060
Vietnam	1999	0.0638	0.0075	-0.0007	0.0069

Source: Treatment effect and standard errors by country-year of *Same-Sex* on *Having more children* and *Being economically active*. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

Table A2: Mundlak estimation interaction term coefficients

Level of aggregation	Variable	Interaction term coefficient for having more children	SE for having more childrem	Interaction term coefficient for economically active	SE for Econ. active
Individual	Age mother	0.0010	(0.000)**	0.0000	0.0000
Individual	Age father	0.0000	(0.000)	0.0000	0.0000
Individual	Age mother second born	0.0000	(0.000)*	0.0000	0.0000
Individual	Age mother first born	0.0000	(0.000)	0.0000	(0.000)**
Individual	First born is twin	-0.0030	(0.010)	0.0020	(0.0050)
Individual	Education mother level2	0.0110	(0.002)**	0.0000	(0.0010)
Individual	Education mother level3	0.0160	(0.002)**	0.0000	(0.0020)
Individual	Education mother level4	0.0060	(0.002)*	0.0030	(0.0020)
Individual	Education spouse level2	0.0060	(0.001)**	0.0010	(0.0010)
Individual	Education spouse level3	0.0110	(0.002)**	0.0000	(0.0010)
Individual	Education spouse level4	0.0110	(0.002)**	0.0010	(0.0020)
Country-Year	Age mother	0.0070	(0.003)*	-0.0020	(0.0020)
Country-Year	Age father	0.0010	(0.001)	0.0010	(0.0010)
Country-Year	Age mother second born	0.0020	(0.001)	-0.0010	(0.0010)
Country-Year	Age mother first born	-0.0010	(0.002)	0.0010	(0.0010)
Country-Year	First born is twin	-0.0120	(0.032)	0.0080	(0.0150)
Country-Year	Education mother level2	0.0300	(0.030)	0.0280	(0.0170)
Country-Year	Education mother level3	-0.0590	(0.036)	-0.0250	(0.0230)
Country-Year	Education mother level4	0.0390	(0.071)	0.0460	(0.0340)
Country-Year	Education spouse level2	-0.0210	(0.033)	-0.0320	(0.0180)
Country-Year	Education spouse level3	0.0630	(0.039)	0.0260	(0.0240)
Country-Year	Education spouse level4	-0.1190	(0.093)	-0.0380	(0.0400)
Country	Population density	0.0000	0.000	0.0000	(0.000)**
Country-Year	log GDP per capita	0.0090	(0.002)**	0.0000	(0.0020)
Country	rssnat_kg (?)	0.0000	(0.000)	0.0000	0.0000
Country	Ethnic fractionalization	0.0000	(0.000)	0.0000	0.0000
Country	Region 2	-0.0160	(0.007)*	0.0080	(0.0040)
Country	Region 3	-0.0120	(0.010)	0.0090	(0.0060)
Country	Region 4	0.0120	(0.013)	0.0020	(0.0070)
Country	Region 5	-0.0170	(0.008)*	-0.0060	(0.0060)
Country	Region 6	-0.0230	(0.007)**	0.0050	(0.0040)
Country	Region 7	-0.0180	(0.010)	0.0060	(0.0060)
Country	Decade 1970	0.0010	(0.003)	-0.0020	(0.0020)
Country	Decade 1980	0.0100	(0.004)*	0.0000	(0.0020)
Country	Decade 1990	0.0070	(0.004)	-0.0010	(0.0020)
Country	Decade 2000	0.0010	(0.004)	0.0010	(0.0030)
	Cosntant	0.0000	(0.003)	0.0000	(0.0100)
	R-Squared	0.00		0.00	
	Number of obs.	8169580		6934850	

Notes: The table shows coefficients on interactions between the listed variable and the "same sex" treatment indicator from Mundlak regression described in Appendix 2 of the paper. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

* p <.05, ** p <.01, *** p<.001 in tests using heteroskedasticity robust standard errors.

Table A3: Lasso regression solution path for *Having more children*

Step	Cp	R-squared	Level of variable	Variable
1	4625.279	0.0000		(intercept)
2	2498.716	0.0003	Country-Year	log GDP per capita
3	2350.882	0.0003	Country-Year	Age mother
4	1584.965	0.0004	Country-Year	Education mother level3
5	1376.348	0.0004	Individual	Education mother level3
6	1363.203	0.0004	Country	Region 5
7	1300.660	0.0004	Individual	Education spouse level3
8	982.563	0.0004	Individual	Education mother level2
9	984.033	0.0004	Country	Region 6
10	880.924	0.0005	Country-Year	Age mother second born
11	878.040	0.0005	Individual	Education spouse level4
12	834.695	0.0005	Country	Region 4
13	787.947	0.0005	Individual	Age mother
14	788.196	0.0005	Country	Decade 1980
15	731.651	0.0005	Country	Region 7
16	701.073	0.0005	Country-Year	Education spouse level3
17	681.351	0.0005	Individual	Education spouse level2
18	541.294	0.0005	Country	Decade 1970
19	489.224	0.0005	Country	Decade 1990
20	456.303	0.0005	Individual	Age mother second born
21	288.558	0.0005	Individual	Education mother level4
22	222.202	0.0005	Country-Year	Education mother level2
23	180.748	0.0005	Country	Population density
24	155.622	0.0006	Country	Region 3
25	146.025	0.0006	Country	Ethnic fractionalization
26	146.285	0.0006	Individual	Age father
27	141.768	0.0006	Individual	First born is twin
28	134.628	0.0006	Country	Region 2
29	83.319	0.0006	Country-Year	Education spouse level4
30	62.811	0.0006	Individual	Age mother first born
31	62.097	0.0006	Country-Year	Age father
32	63.022	0.0006	Country-Year	First born is twin
33	57.906	0.0006	Country-Year	Education mother level4
34	52.194	0.0006	Country	Government consumption share
35	41.047	0.0006	Country-Year	Education spouse level2
36	36.183	0.0006	Country-Year	Age mother first born
37	37.000	0.0006	Country	Decade 2000
	Number of obs.	8169580		

Notes: The table shows the solution path using the least angle algorithm to fit the lasso to the Mundlak regression as described in Appendix 2 of the paper. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

Table A4: Lasso regression solution path for *Economically active*

Step	Cp	R-squared	Level of variable more childrem	Variable more childrem
1	156.232	0.0000		(intercept)
2	136.994	0.0000	Country-Year	Education spouse level4
3	128.846	0.0000	Country-Year	Age mother
4	122.448	0.0000	Country	Region 5
5	105.865	0.0000	Country	Region 4
6	101.807	0.0000	Individual	Age mother second born
7	97.879	0.0000	Country-Year	Age father
8	79.261	0.0000	Country	Region 2
9	70.851	0.0000	Country	Region 6
10	71.677	0.0000	Country	Population density
11	67.411	0.0000	Individual	Education mother level4
12	65.604	0.0000	Country	Decade 1970
13	45.665	0.0000	Country	Decade 2000
14	43.180	0.0000	Country	Region 3
15	30.642	0.0000	Individual	Education spouse level4
16	26.652	0.0000	Country-Year	First born is twin
17	26.758	0.0000	Individual	Education spouse level2
18	28.009	0.0000	Country	Government consumption share
19	28.826	0.0000	Individual	Age father
20	28.031	0.0000	Country-Year	Education spouse level2
21	29.632	0.0000	Individual	Age mother
22	30.859	0.0000	Individual	First born is twin
23	29.847	0.0000	Country	Ethnic fractionalization
24	29.974	0.0000	Individual	Age mother first born
25	30.202	0.0000	Country-Year	Age mother first born
26	31.884	0.0000	Individual	Education spouse level3
27	32.108	0.0000	Country-Year	Education spouse level3
28	33.936	0.0000	Country	Region 7
29	35.914	0.0000	Country-Year	Decade 1990
30	37.716	0.0000	Individual	Education mother level3
31	35.905	0.0000	Country-Year	Education mother level2
32	37.253	0.0000	Country-Year	Education mother level4
33	35.150	0.0000	Country-Year	Age mother second born
34	36.404	0.0000	Country-Year	Education mother level3
35	36.277	0.0000	Country	Decade 1980
36	35.162	0.0000	Country	log GDP per capita
37	37.000	0.0000	Individual	Education mother level2
	R-Squared	0.00		
	Number of obs.	6934850		

Notes: The table shows the solution path using the least angle algorithm to fit the lasso to the Mundlak regression as described in Appendix 2 of the paper. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).