NBER WORKING PAPER SERIES

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

Rajeev Dehejia
Cristian Pop-Eleches
Cyrus Samii

Working Paper 21459
http://www.nber.org/papers/w21459

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2015, Revised March 2017

The authors thank Morris Chow for excellent research assistance; Ali T. Ahmed, Hunt Allcott, Joshua Angrist, Peter Aronow, James Bisbee, Gary Chamberlain, Drew Dimmery, Rachel Glennerster, and Raimundo Undurraga for valuable comments and suggestions; and seminar participants at the BREAD conference, Cowles Econometrics Seminar, EGAP, the Federal Reserve Board of Cleveland, the Federal Reserve Board of New York, GREQAM, IZA, Maastricht, NEUDC 2014, Columbia, Harvard, MIT, NYU, UCLA, UCSD, the World Bank, Yale, the 2014 Stata Texas Empirical Microeconomics Conference, and the Stanford 2015 SITE conference for helpful feedback. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

From Local to Global: External Validity in a Fertility Natural Experiment
Rajeev Dehejia, Cristian Pop-Eleches, and Cyrus Samii
NBER Working Paper No. 21459
August 2015, Revised March 2017
JEL No. C18,C31,C9,F63,J13

## ABSTRACT

We study issues related to external validity for treatment effects using 166 replications of the Angrist and Evans (1998) natural experiment on the effects of sibling sex composition on fertility and labor supply. The replications are based on census data from around the world going back to 1960. We decompose sources of error in predicting treatment effects in external contexts in terms of macro and micro sources of variation. In our empirical setting, we find that macro covariates dominate over micro covariates for reducing errors in predicting treatments, an issue that past studies of external validity have been unable to evaluate. We develop methods for two applications to evidencebased decision-making, including determining where to run the next experiment and whether policy-makers should commission new research or rely on an existing evidence base for making a policy decision.

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

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

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

## 1. Introduction

In recent decades across a wide range of fields in economics, such as labor, education, development, and health, the use of experimental and quasi-experimental methods has become widespread. The emphasis on experimental and quasi-experimental methods[1] was driven by an attempt to generate internally valid results. At the same time, the global scale of 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 studies in a limited range of contexts, external validity is mostly a matter of theoretical speculation. But with a large number of internally valid studies across a variety of contexts, it is reasonable to hope that researchers are accumulating generalizable knowledge, i.e., not just learning about the specific time and place in which a study was run but about what would happen if a similar intervention were implemented in another time or place.

The success of an empirical research program can be judged by the diversity of settings in which a treatment effect can be reliably predicted, possibly obviating the need for further experimentation with that particular treatment. This is the issue we address in this paper. More specifically, given internally valid evidence from "reference" settings, is it possible to predict the treatment effect in a new ("target") 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 micro variables? How might we judge whether an existing evidence base is adequate for informing new policies, thereby making further experiments with a given treatment unnecessary?

Although the issue of external validity has garnered the most attention recently in the context of randomized controlled trials, it is important to underline that the essential

[1] Throughout the remainder of the paper, we will use the term *experiments* broadly as referring to internally valid studies that use either true random experimental or quasi-experimental methods.

challenge of extrapolation is common to the broad set of methods used to identify treatment effects. Each of these methods has its own specific challenges for extrapolation. In this paper, as a starting point, we focus on reduced-form experiments or natural experiments. In ongoing and future work, we extend the analysis to other research designs (see for example Bisbee, Dehejia, Pop-Eleches, and Samii 2016 for a related analysis of the instrumental variables case).

Our approach in this paper is to use a natural experiment for which "replications" are, in fact, available for a wide variety of settings. 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. Replications of this natural experiment are recorded for a large number of countries over many years in censuses compiled in 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. Filmer, Friedman, and Schady (2009) estimate effects of sex composition on incremental fertility for mothers in different regions around the world using Demographic and Health Survey data. Their results show that the effect of sex composition on incremental fertility is apparent around the world, particularly in trying to make up for the absence of sons in early births.

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 natural experiment when estimated on a global scale. To the extent that fertility choices could be viewed as 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 an experiment or well-identified result 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 sites and also the

homogeneity in treatments and outcomes. What brings us back down to earth is that we have only a limited amount of information that we can use to characterize effect heterogeneity. As an empirical investigation, our task is to assess the external validity potential of this evidence base in extrapolating to new contexts. The evidence base consists of the set of studies and its limitations are defined by the variety of contexts that it covers and, crucially, the measured covariates that it includes. We approach the extrapolation problem as empiricists, using the available data in an agnostic and flexible manner. We examine how working through the extrapolation problem using the evidence base can inform how an experimental or quasi-experimental research program might optimally proceed. A complementary exercise, which we do not undertake in this paper, would be to use the evidence base to explain effect heterogeneity for the sake of theory development.

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), Gechter (2015), Pritchet and Sandefur (2013), Rosenzweig and Udry (2016), and Vivalt (2014). Allcott (2014), for example, 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 which opt in early into an experiment may be those most likely to benefit. 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 dominate the importance of individual characteristics when attempting to extrapolate. Rosenzweig and Udry (2016) instead examine the sensitivity of experimental estimates to aggregate shocks over time, e.g., the effect of agricultural prices or rainfall shocks on returns to investment in agriculture. We view our efforts as complementary. With a large number of (natural) experiments in our data set (166 repeated cross-sections from 61 unique countries) we are able to examine

empirically the relative importance of micro- vs. macro-level contextual variation for a range of settings and contextual variables.

Vivalt (2014) uses a random effects meta-analysis to study sources of effect heterogeneity 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 distinguish issues of extrapolation 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 (experiment – in our case country-year – level) covariates. We define and estimate an "external validity function" that characterizes the quality of an evidence base's predictions for a target setting. We examine the relationship between prediction error and individual and context covariates; while both are potentially useful in reducing prediction error from external comparisons, in our application context variables dominate. This is an important finding, since methodological work to date has tended to focus on accounting for variation in micro-level variables. But even the meaning of micro-level variables depends on context (e.g., a 35-year-old woman in a lower income country may have different potential outcomes than a woman of the same age in a high income country). Our analysis and empirical results indicate the need to take context-level heterogeneity into consideration.

Finally, we present two applications to evidence-based decision-making. In the first, we use the external validity function to determine the best location of a new experiment. Specifically, choosing among our 166 country-year sites, we ask which location would minimize mean squared prediction error for the other sites? In the second application, we ask when a policy decision maker should choose to run an experiment in a target setting rather than use extrapolated estimates of the treatment effect from an

existing evidence base. For both applications, pre-treatment covariate data proves to be crucial. Questions of external validity motivate the collection of rich covariate data even when an experiment or natural experiment does not require it for internal validity.

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 questions of external validity, rather than relying only on theoretical speculation. Focused consideration of external validity goes back at least to Campbell (1957), whose approach is taken up by Shadish et al. (2002). Debates in the classical literature omit a formal statement of how external validity may be achieved. More recently, Hotz, Imbens, and Mortimer (2005), Stuart et al. (2011), and Hartman et al. (2015) 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 using covariates to identify causal effects under “strong ignorability” (Rosenbaum and Rubin, 1983). 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. Making use of such identifying conditions requires measuring statistical relations between covariates and treatment effects that are invariant as we move from the reference to the target

population (Heckman and Vytlacil, 2007; Pearl and Bareinboim, 2014). We review these conditions in the next section.

Hotz et al. (2005), Stuart et al. (2011), and Hartman et al. (2015) apply various approaches to extrapolation from one site to another, 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. Because these previous studies work with only a small number of sites, they focus on micro-level differences across sites. Our analysis addresses both micro-level differences and macro-level differences (that is, country-year-level contextual characteristics). 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. We avoid this is issue in the current discussion, as we focus only on reduced form or intention-to-treat effects. We address extrapolation with instrumental variables in related work (Bisbee, Dehejia, Pop-Eleches, and Samii 2016).

Our analysis is 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.

## 3. Analytical framework

We are interested in using the results of existing experiments to inform our expectations of what might happen in a new, external context. We draw a sample of $C + 1$ contexts (e.g., country-years) from a broader population of contexts, $S$, (e.g., the entire global history of country-years). The sampled contexts are denoted as sets $S_c$ for $c = 1,\ldots,C + 1$, and from each context we have a random sample of units indexed by $i$, where $i \in S_c$ for some c. Our interest is in the effects of a binary unit-level treatment variable, $T_{ic} = 0,1$, which is governed by context-specific random assignment distributions, $P_c$. Following the current program evaluation literature (e.g., Imbens and Rubin, 2015), each unit $i$ in context $c$ possesses potential outcomes associated with the treatment values, $Y_{ic}(0)$ and $Y_{ic}(1)$, respectively. Units are also characterized by a unit-level vector of covariates, $W_{ic}$, taking values in the support $\mathcal{W}$. We also suppose that each context is characterized by a vector of covariates, $V_c$, taking values in the support $\mathcal{V}$.

We designate the first sampled context, $S_1$, to be the "target" context to which we aim to make inferences. Define the context level variable, $D_c = \mathrm{I}(c = 1)$, to denote whether or not the context $c$ is the target context. The support of unit-level covariates in $S_1$ is given by $\mathcal{W}_1$. The rest of the $C$ contexts are "reference" contexts for which units are subject to variation in the $T_{ic}$ assignments. That is, the reference contexts are denoted as the set $S_r = \{S_c : \mathrm{Var}_{Pc}[T_{ic}] > 0\}$. The analysis here considers the case where $S_1$ and $S_r$ are disjoint, so that there is no treatment variation in the target site. In that case, $D_c = 1$ if and only if $i \in S_1$, and so $D_c = 0$ for $i \in S_r$. From the $C$ reference contexts we also obtain data on outcomes, which are given as

(1) $\quad Y_{ic} = T_{ic}\, Y_{ic}(1) + (1 - T_{ic})\, Y_{ic}(0).$

Note that this embeds the so-called "stable unit treatment value assumption" (SUTVA) (Imbens and Rubin, 2015, pp. 9-12). We assume throughout that SUTVA would hold in the target contexts as well. We suppose that the data from the reference contexts are from randomized experiments such that the following condition holds:

(C0) $\quad T_{ic} \perp\!\!\!\perp ( Y_{ic}(0), Y_{ic}(1) ) \mid D_c = 0.$

Our target of inference is the average treatment effect in $S_1$, which we denote as

(2) $\quad \tau_1 = E[Y_{ic}(1) - Y_{ic}(0) | D_c = 1]$ .

where $E[.]$ takes expectations over $S$.

Following Hotz et al. (2005), we work under identifying assumptions on the selection of the target context relative to the reference contexts. First is what Hotz et al. refer to as the "unconfounded location" assumption:

(C1) $D_c \perp\!\!\!\perp ( Y_{ic}(0), Y_{ic}(1) ) \mid ( V_c, W_{ic} )$.

Unconfounded location implies, effectively, random assignment of individuals to reference versus target sites conditional on covariate values. The assumption is standard in the current literature on external validity (e.g., Hartman et al., 2015; Stuart et al., 2011), although it is stronger than it needs to be. In fact, the following assumption, when combined with C2 below, is sufficient for non-parametric identification of $\tau_1$ with data from $S_r$:

(C1b) $E[Y_{ic}(1) - Y_{ic}(0) | D_c = 1, W_{ic}, V_c] = E[Y_{ic}(1) - Y_{ic}(0) | D_c = 0, W_{ic}, V_c]$.

For extrapolating causal effects, the differences between C1 and C1b could be meaningful. Whereas C1 implies that the full joint potential outcome distributions are invariant conditional on the covariates, C1b allows for conditional mean levels of potential outcomes to vary across sites, so long as mean effects are invariant. For certain applications, such as in the biomedical sciences, this weaker "effect invariance" assumption may be reasonable.

The second assumption is the covariate overlap assumption:

(C2) $\delta < \Pr(D_c = 0 | V_1, W_{ic} = w) < 1 - \delta$,

for all $w \in \mathcal{W}_1$ and where the conditioning on $V_1$ restricts us to cases where the context-level covariates for the target context are within the support of the reference contexts.

Conditions C1 (or C1b) and C2 imply that data on effects in contexts for which $D_{ic} = 0$ are sufficient to identify effects in context for which $D_c = 1$ (Hotz et al. 2005, Lemma 1).[2] That is, by C0-C2 we have

(3) $$E[Y_{ic}(1) - Y_{ic}(0) | D_c = 1] = \int_{\mathcal{W}_1} E[Y_{ic}(1) - Y_{ic}(0) | D_c = 0, W_{ic} = w, V_1] \, dF(w | D_c = 1)$$
$$= \int_{\mathcal{W}_1} \{E[Y_{ic} | T_{ic} = 1, D_c = 0, W_{ic} = w, V_1] - E[Y_{ic} | T_{ic} = 0, D_c = 0, W_{ic} = w, V_1]\} \, dF(w | D_c = 1),$$

where the first equality is due to C1 and iterated expectations, and the terms on the last line are identified and estimable given C2.

[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.

Given this set up, we can decompose potential outcomes in terms of context-level and unit-level variation. For $t$ = 0,1, by the linearity of expectations, the following identity holds:

(4) $$Y_{ic}(t) = \alpha(t) + f(V_c, t) + g(V_c, W_{ic}, t) + \eta_c(t) + \phi_c(W_{ic}, t) + \nu_{ic}(t),$$

which involves two sets of components. The first are structural components common to units over the entirety of $S$, namely $\alpha(t) = E[Y_{ic}(t)]$, $f(V_c, t) = E[Y_{ic}(t)|V_c] - \alpha(t)$, and $g(V_c, W_{ic}, t) = E[Y_{ic}(t)|V_c, W_{ic}] - E[Y_{ic}(t)|V_c]$. The second are components characterizing variation idiosyncratic to $S_c$, namely $\eta_c(t) = E_c[Y_{ic}(t)] - E[Y_{ic}(t)|V_c]$, $\phi_c(W_{ic}, t) = E_c[Y_{ic}(t)|W_{ic}] - E[Y_{ic}(t)|V_c, W_{ic}]$, and $\nu_{ic}(t) = Y_{ic}(t) - E_c[Y_{ic}(t)|W_{ic}]$, where $E_c[.]$ is expectations on $S_c$. The structural components are equivalent to "fixed effects" in the terminology for mixed-effects models, while the idiosyncratic components are equivalent to "random effects" (Greene, 2008, pp. 233-243; Jiang, 2007).

Conditioning on covariates, by C1 we have the following for the potential outcomes that make up $\tau_1$:

(5) $$E[Y_{ic}(t)|D_c = 1, W_{ic} = w, V_1] = \alpha(t) + f(V_1, t) + g(V_1, w, t),$$

where the "random effects" terms are expectation zero because of the conditional randomness of $D_c$. Given the random sampling from $S$ that C1 implies and assuming linearity of $f(.)$, $g(.)$, and $\phi_c(.)$, the components in (5) would be identified under the usual random effects restrictions. Moreover, with full data in the target contexts, we could obtain estimates for the random effect components for all contexts. Dehejia (2003) performs an analysis along these lines for a set of labor training programs. In the current setting, however, we do not have the outcome or treatment data from the target setting and so our attention focuses first on the structural components. Under C1 we can write our estimand as,

(6) $$\tilde{\tau}_1 = \big(\alpha(1) - \alpha(0)\big) + \big(f(V_1, 1) - f(V_1, 0)\big) + \Big[\int_{\mathcal{W}_1} \big(g(V_1, w, 1) - g(V_1, w, 0)\big)\, dF(w|D_c = 1)\Big],$$

Expressions (4) and (6) also imply that the (ex post) prediction error for $\tau_1$ is given by the context-level random effects. The distribution of the random effects defines the uncertainty in relating $\tilde{\tau}_1$ to $\tau_1$, which we can characterize in terms of a prediction interval around estimates of $\tilde{\tau}_1$. For example, if the random effects are independent and normal, then over repeated selection of target contexts, the distribution of $\tau_1$ values

would be normal, centered on $\tilde{\tau}_1$ and have variance equal to the sum of the variances of the random effect components.

Constructing an estimator for $\tilde{\tau}_1$ requires specifications for $f(.)$ and $g(.)$. Similar to the non-parametric approach proposed by Hotz et al. (2005), we use series regressions of the unit-level covariates, context-level covariates, and their interactions (Newey, 1994). That is, for units in the reference contexts with $T_{ic} = t$, we fit the following series regressions with treatment-covariate interactions:

$$\mu_{ic}(W_{ic}, V_c, t) = \alpha_0 + t\widetilde{\alpha_{1,0}} + \sum_p P_0^p(V_c) + t\sum_p \widetilde{P_{1,0}^p}(V_c) + \sum_q P_0^q(W_{ic}) + t\sum_q \widetilde{P_{1,0}^q}(W_{ic}) + \sum_r P_0^r(V_c, W_{ci}) + t\sum_r \widetilde{P_{1,0}^r}(V_c, W_{ci}), \quad (7)$$

where $p$, $q$, and $r$ index the order of the series of approximating functions that specify higher order powers and interactions. That is, for an $m$-length vector of regressors, the approximating functions take the form, $P^s(X) = \sum_{(v_1,\ldots,v_m)\in\Omega_s} \prod_{j=1}^m \beta_{s,(v_1,\ldots,v_m)} X_j^{p_m}$ for vectors of non-negative integers ($v_1$,…,$v_m$) such that $\sum_{j=1}^m v_j = s$. The terms with tilde superscripts measure differences between treated and control conditional potential outcome means (hence the 1,0 subscripting). The approximating functions with tilde superscripts are treatment variable-covariate interaction terms, and therefore measure the ways that treatment effects vary on the basis of covariate differences across individuals and contexts. We fit the model using ordinary least squares, and the order of the approximating functions is determined using LASSO regularization, as in Belloni et al. (2014). Our resulting estimate of the treatment effect in the target context, $S_1$, based on the fit from the reference contexts $S_r$, is given by

$$\hat{\tau}_{1r} = \frac{1}{n_{S_1}} \sum_{i\in S_1} [\hat{\mu}_{ic}(W_{i1}, V_1, 1) - \hat{\mu}_{ic}(W_{i1}, V_1, 0)]. \quad (8)$$

We also define the "prediction error" between $S_1$ and $S_r$ as,

$$\zeta_{1r} = \hat{\tau}_{1r} - \tau_1. \quad (9)$$

The quantity $\zeta_{1r}$ 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 $\zeta_{1r}$ is an unobservable quantity since one does not know the true effect, $\tau_1$. The empirical exercise that we carry out is one where we can actually produce estimates of $\tau_1$ for the various country-year contexts in the data set. We can then use these as benchmarks to assess extrapolations from other country-year contexts. The distribution of $\zeta_{1r}$ is governed by

the combination of (i) random variation in selection of contexts and treatment assignment and (ii) bias resulting from failures of C0-C2 and misspecification in (7).[3] Below when we study the distribution of $\zeta_{1r}$ values for different combinations of target and reference contexts, we take into account such sources of random variation. Moreover, our setting is such that C0 can be assumed to hold and the covariate set is relatively parsimonious, in which case C3 is also uncontroversial. What remains in question, then, is C1. Our descriptive analysis of $\zeta_{1r}$ values below therefore provides an informal test of C1.

As Hotz et al. (2005) and also Gechter (2015) indicate, an implication of C1 is that the $Y_{ic}(0)$ distributions are invariant across the contexts conditional on $(V_c, W_{ic})$. If the $Y_{ic}(0)$ distributions are observable across contexts, this allows for another test of C1 (although such a test cannot speak, empirically, to the validity of C1b).

In testing for the quality of extrapolations from reference to target contexts, 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 approximately 28,000 dyads consisting of hypothetical target and reference country-years. In the cumulative analysis, the reference set includes country-year contexts in years prior to that of the target country-year.

## 4. A global natural experiment

There are two main challenges for assessing 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 are 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

[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. Thus, conditional on covariates, by standard arguments, our estimates of $\hat{\tau}_{1r}$ and $\tau_1$ are statistically independent and asymptotically normal (e.g., Abadie et al., 2014; Freedman, 2008; Lin, 2013), in which case our estimate of $\zeta_{1r}$ is also asymptotically normal.

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 application, we were able to use 166 country-year samples (from 61 unique countries) with information on fertility outcomes and 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 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 (see Bisbee, Dehejia, Pop-Eleches, and Samii 2016 for an effort to extrapolate the instrumental variables results).

For our empirical analysis, we implement essentially the same sample restrictions, data definitions, and regression specifications as those proposed in Angrist and Evans (1998).[4] 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 first stage and 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 (assumption C0 from Section 3). As we will see below, the 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 related models and empirical 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 context variables, obvious candidates are female labor force participation as a broad measure of employment opportunities for women in a given country (Blau and Kahn, 2001) and the total fertility rate. Since the goal of our exercise is extrapolation, we also include a number of macro variables that do not necessarily play a direct causal role in explaining

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

fertility and labor supply decisions but rather have been shown to be important in explaining broad patterns of socio-economic outcomes across countries; these include log GDP per capita, as a broad indicator of development, average education, and geographic distance between reference and target country (Gallup, Mellinger and Sachs, 1998).

Descriptive statistics for our 166 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. Effects are measured in terms the changes in the probability of having more kids and being economically active.

**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 complete-data 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.

Figures 1 and 2 also highlight the fact that not all country-year treatment effects are statistically significantly different from zero. In Figure 1, approximately three fourths

of treatment effects are significant at the 10 per cent level (and two thirds at the 5 per cent level). In Figure 2, approximately one tenth of the treatment effects are significant at standard levels. Given this, in our subsequent analysis, we weight the country-year treatment effects by the standard error of the treatment effect.

The next set of figures investigates whether any of the treatment effect heterogeneity documented in Figures 1 and 2 is 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 heterogeneity based on geographic region, indicating 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 and that this heterogeneity is associated with levels of development.

## 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 Cochran's Q tests for effect homogeneity (Cochran, 1954), 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 with degrees of freedom equal to the number of effects minus one, 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 that the correlation is 1, i.e., we can reject the hypothesis of normality. 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

Appendix figures 1-3 present results of tests for the unconfounded location assumption using the $Y_{ic}(0)$ distributions, in the same spirit of the tests used by Hotz et al. (2005) and Gechter (2015) but also accounting for our regression-based approach. They show how differences between predicted and actual mean $Y_{ic}(0)$ values vary as we allow for differences in one or another covariate value. At zero covariate difference, the graphs pass through the origin. As such, our covariate set allows for accurate prediction of conditional mean $Y_{ic}(0)$ values. This is what we would expect if unconfounded location holds.

## 7. Characterizing heterogeneity: external validity function and unconditional relationships

In this section, we characterize how prediction error changes with context covariates such education, 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). We conduct this descriptive exploration in terms of what we call an "external validity function", which characterizes how prediction errors vary in the context-level covariate differences. Specifically, we estimate the treatment effect in a target context using only unit-level covariates from the reference contexts. This yields a prediction error estimate, $\hat{\zeta}_c^W$, for each target context. We then evaluate how this prediction error varies in $|V_{ck} - \bar{V}_{rck}|$, where $\bar{V}_{rck}$ is the mean of the *k*th context level covariate from the reference contexts used to generate the prediction for site *c*. In the dyadic analyses that we do below, $\bar{V}_{rck}$ is simply equal to the value of the kth context level covariate. For some of the analyses below, we construct a context-level covariate from unit-level covariates by taking the context mean.

In Figures 7 to 10, we present local linear regressions of prediction error at the dyad level on within-dyad covariate differences between the target and reference country-years. 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 standard deviation education difference (approximately one point on the four-point scale) error increases by approximately 0.1 (relative to the world treatment effect of 0.04 in Figure 1). The figure also plots +/- two standard errors of the external validity function, which is relatively flat over the range of -2 to +2 educational differences, but increases 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 a one standard deviation GDP per capita difference (approximately $10,000). 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 prediction error.

## 8. Characterizing heterogeneity: conditional relationship

In this section we continue our characterization of heterogeneity by estimating the multivariate relationship between prediction error and the full range of dyadic 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) to (9) of Table 3, we run the prediction error regressions one covariate at a time, giving us essentially the unconditional prediction error. Most covariates (measured as standard deviations of reference-target differences in education, education of spouse, age of the mother, year of census, log GDP per capita and labor force participation) are statistically significant, with a one standard deviation covariate difference increasing prediction error by 0.05 to 0.1, an order of magnitude approximately between one and two times the treatment effect (with differences in mother's age and total fertility rate leading to even larger errors). Geographical distance notably is not statistically significant.

In columns (10) to (11) of Table 3, we estimate multivariate prediction error regressions. Five main observations can be drawn from the results. First, the constant in the regressions is close in magnitude to, and not statistically significantly different from, zero, matching the finding from Figures 7-10 that when covariate differences between the reference and target location are small prediction error is also small. This is consistent with the unconfounded location (assumption C1). Second, many of the variables are statistically significant, although we note that education and labor force participation lose significance once the other controls are included. Third, the size of the prediction error due to covariate differences is generally large given an average treatment effect in the sample of 0.04. Fourth, it is noteworthy that the effects of GDP per capita and total fertility rate are negative in column (10). Since the unconditional effect of GDP per capita differences is positive in column (5), this reflects the counter-intuitive nature of the variation identifying the conditional coefficient: variation in GDP per capita conditional on a similar education, age, and labor force participation profile of women is presumably quite limited. At the same time, the coefficient on the difference in total fertility rate is negative both unconditionally (in column (8)) and conditionally (column (10)). This implies that the treatment effect is decreasing in total fertility rate, so comparing a reference country-year to a target country-year with a lower total fertility leads to

negative prediction error (under-estimation of the treatment effect). Fifth, the sex ratio imbalance enters positively, implying that it is indeed important to consider the degree of sex selectivity within countries when extrapolating the treatment effect. This remains true even when we drop the most notable sex-selectors from the sample (China, India, Nepal, and Vietnam, column (11)). Furthermore, dropping sex-selecting countries does not meaningfully change the estimated coefficients on covariate differences.

The results in Table 4 for the effect of *Same-Sex* on *Economically active* are similar in three respects. First, the constant is not statistically significantly different from zero at least when all covariates are included in columns (10) to (11), again consistent with unconfounded location (C1). Second, the magnitude of prediction error generated by reference-covariate target differences is large relative to the treatment effect. Third, covariate differences enter both positively (sex ratio imbalance, total fertility rate) and negatively (age of the mother, calendar year, and labor force participation of women) both unconditionally and conditional on other covariates. This reflects different patterns of treatment effect heterogeneity: a positive coefficient on the reference-target covariate difference implies that the treatment effect is increasing in the covariate (so if the target country has a higher value of the covariate, one overestimates the treatment effect in the reference country), a negative coefficient the opposite.

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-year 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 all of the other country-years are treated as reference sites. Pooling the data from the reference 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 contribute in pushing prediction error towards zero, dominating the scenario of no covariates. In the density plots, inclusion of covariates brings in the tails toward zero, and in the CDF plot the error distribution is drawn toward zero. However, the contribution of the macro variables is much stronger and almost completely removes the error. The results in Figure 12, which use *Economically active* as the outcome variable of interest, provide an even starker 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 is that a set of easily available cross-country variables has the potential to be useful in analyzing of external validity. This also raises concerns about generating extrapolations solely on the basis of micro-level data, an issue that Hotz et al. (2005), Stuart et al. (2011), and Hartman et al. (2015) were unable to investigate due to the limitations of their evidence bases.

Finally, we obtain similar results on the importance of context-level covariates when we use the LASSO regularization to specify the approximating functions characterized in expression (7) above. Appendix figure 4 shows the solution paths for the interaction terms in the series expansion. The solution path reveals that an error-minimizing specification (in terms of Mallow's Cp-statistic) is quite sparse in the interaction terms retained. Moreover, macro-level and macro-micro interaction terms dominate the LASSO solution paths through to the error-minimizing specifications. Even

in the fully saturated specification, the macro and macro-micro interaction terms that we have included dominate in terms of explanatory power (evident in looking at the standardized coefficient values displayed all the way to the right in the graphs of the full LASSO solution paths, Panels A and B). These results confirm two impressions arising from the exploratory analysis above: first, much of the effect heterogeneity is attributable to macro-level variation, and second, to the extent that micro-level variables matter in explaining effect heterogeneity, the influence of these micro-level variables is strongly moderated by macro-level moderation (e.g., the age of mothers moderates treatment effects, but in a manner that differs depending on macro context).

**9. The accumulation of evidence and prediction error**

Our results so far imply that with sufficient covariate data, particularly macro covariates, we can extrapolate the treatment effect with small expected prediction error. We now consider if and how the accumulation of experiments over time improves our ability to extrapolate to new settings or alternatively how well we are able to extrapolate with only a small experimental evidence base. We consider both expected prediction error as well as uncertainty estimates.

In Figure 13, we model the effect of *Same-Sex* on *Had more children* 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 country-year dyads available by 1980; fit the model to these dyads; and then estimate the prediction error for this sample. In Figure 13, we plot the resulting average prediction error values over time. The pattern shows that as we add more data to the model over time average prediction error eventually approaches zero. A second striking pattern is that prediction error 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 pertain to in-sample predictive accuracy or model fit, and not to 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-years observed in each year on the x-axis (e.g., the U.S. in 1980), we calculate the prediction error from taking a treatment effect estimate from a

reference sample and using that as an extrapolation estimate for the target. We choose the reference sample in four different ways: (1) all country-years available up to that year, excluding the target (graphed as the red line); (2) the best (lowest prediction error) reference country-year as selected by the prediction-error model (from Table 3) fit to data from prior years (graphed as the blue line); (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 all available country-years (in red) versus the best reference country-year selected by the model (in blue) confirms that when using our model we get much lower prediction error compared to the estimate from pooling all the samples available. Second, the pattern of prediction error over time from using the best model-selected reference 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 model-selected reference country-year 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 type (4) 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, the type (3) 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. This illustrates the risks of rules of thumb compared to a model based approach. *A priori*, allowing own-country comparisons seems plausible, but own-country comparisons are usually at least 10 years apart and our model easily accommodates the tradeoff between these competing factors.

This is underlined in looking at Figure 16, where the model outperforms both rules of thumb when the available reference samples are sparse.

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 the first or second birth. The patterns are similar.

Overall, we draw three conclusions from this analysis. First, without a sufficient number of experiments extrapolating the treatment effect is challenging; while the model-based approach performs well on average, in our data, its reliability is sensitive to year-to-year variation in the reference sample until around 1985 (by which point we have accumulated 54 country-year samples). Second, with a sufficiently large evidence base rules of thumb become more reliable. Third, in both rich and sparse data environment the model-based approach helps in trading off the pluses and minuses of the available reference country-years.

## 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 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 characterizing how the effect of an intervention varies around the world as in Imbens (2010, p. 420) or Rubin (1992), but with limited resources to implement new experiments. In this section we examine what the evidence base implies for the best location of new experimental sites, given this goal.

At the country-year level, our regressions above suggest that prediction error should be low for locations with low covariate distance to the evidence base. In assessing such covariate distance, the question is how to weight different covariates. With

knowledge of the estimates in Tables 3 and 4 (column (10) in each table) 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. Figure 19 provides confirmation for this intuition. We use 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 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 Mahalanobis distance accounts for redundancy due to correlations between regressors. It therefore accounts for all of the information in the linear external validity function specification that we can obtain without knowing the regression coefficients. 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.

If one had to choose only a single site to locate an experiment in order to learn about a collection of sites, the results show that choosing in a manner that minimizes Mahalanobis distance would be the most robust. If, however, the goal is to add new experiments is to an existing evidence base so as to characterize how effects vary, then these results recommend selecting sites that maximize Mahalanobis distance in the

covariates as specified in the external validity function. It is for such sites that the evidence base is unreliable in predicting treatment effects.

*10.2 To experiment or to extrapolate?*

Now suppose 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. Formally, this means 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 for the target site is entirely on one or another side of some critical threshold, $c^*$. We also assume 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. 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 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 the set of points in the covariate space corresponding to covariate values for $S_1$ recall that in expression (3) we defined the conditional effect estimate, $\hat{\tau}_{1r}$. Label the covariate distribution for $S_1$ as $C(S_1)$. We consider this relative repeated selection of a target context, yielding a distribution of target effects for which the $c$-th draw is given by $\tau_c$. In a manner analogous to expression (9), prediction error relative to an arbitrary target $\tau_c$ is given by

(10) $$\zeta_{cr} = \hat{\tau}_{1r} - \tau_c.$$

For the set of contexts with covariate distributed as $C(S_1)$, we have a distribution of $\zeta_{cr}$ values given $\hat{\tau}_{1r}$. This conditional distribution defines our uncertainty about the relationship between $\hat{\tau}_{1r}$ and $\tau_1$. The variance of this conditional distribution is driven by the estimation variability for $\hat{\tau}_{1r}$, which we can expect to be governed by the number of reference contexts available to us, as well as the "intrinsic" variability in the $\tau_c$ values conditional on $C(S_1)$, which we can expect to be governed by the richness of the covariate set. Thus, our prediction error will be a function both of the number of reference experiments *and* the richness of the covariate set.

Assuming $\hat{\tau}_{1r}$ for the conditional mean of effects at $C(S_1)$, 95% prediction interval for our target quantity, $\tau_1$, is given by

(11) $$PI_1 = \left[\hat{\tau}_{1r} - t_{0.025}\sqrt{Var[\zeta_{cr}|C(S_1)]}, \hat{\tau}_{1r} + t_{0.025}\sqrt{Var[\tau|C(S_1)]}\right]$$

in which case the solution to the decision problem is to experiment if $c^* \in PI_1$, and accept the existing evidence otherwise, where $t_{.025}$ is the appropriate .025 quantile value for the normalized conditional distribution of $\zeta_{cr}$. We work under a normal approximation (see fn. 3).

To estimate the conditional variance, $Var[\zeta_{cr}|C(S_1)]$, we proceed in two steps. We first use a leave-one-out approach to estimate $\hat{\zeta}_{cr}$ values for each of the country-year contexts in the evidence base. We then model the these $\hat{\zeta}_{cr}$ values in terms of our covariates, using a series specification analogous to what we used to model the conditional potential outcomes:

(12) $$E\left[\log\left(\hat{\zeta}_{cr}^2\right)|C(S_1)\right] = \alpha_\zeta + \sum_p P_\zeta^p(V_c) + \sum_q P_\zeta^q(\overline{W}_c) + \sum_r P_\zeta^r(V_c, \overline{W}_c),$$

where $\overline{W}_c$ is the mean of the unit-level covariates for country-year context $c$. The exponentiated predicted value at $(V_1, \overline{W}_1)$ is our estimate for $Var[\zeta_{cr}|C(S_1)]$. This

estimate will tend to be conservative, because we are working off of $\hat{\zeta}_{cr}$ estimates, which themselves are products of the respective $\hat{\tau}_{cr}$ values, rather than the true $\zeta_{cr}$ values.

Figure 24 shows the results of applying this approach to estimating the effects of *Same-sex* on *More kids*. Panel A shows how the cumulative reference sample evolves over time, eventually reaching our 142 complete-data country-year samples and about 10 million observations. Panel B shows the prediction intervals for target country-year (gray bars), arrayed by year. We also plot the actual effect estimates from those country-year samples (black dots) as a way to check on the accuracy of the procedure. The figure shows that the predictive intervals are informative, in that they do not span an extreme range, and they almost always cover the in-sample effect. The intervals become a bit tighter as the evidence base grows over time, although they do not collapse to zero. As a result, even for a decision rule based on a critical value of 0 ($c^* = 0$) and even with over 100 reference samples, the analysis would indicate the need for further experimentation.

That the intervals do not collapse to zero is expected because of the intrinsic variability, and this highlights the crucial role of covariate data for analyses that depend on external validity. Unlike the standard error of prediction, the intrinsic variability does not depend on the sample size in a strict sense. Rather, it is a function of the amount of variation left unexplained by the covariates, which remains fixed in this application. Panel C demonstrates this point clearly. The black line traces out the standard error of prediction), which tends toward zero as the reference samples accumulate. The gray dots show the estimates of the intrinsic variation, expressed in standard deviation units and thus on the same scale as the effect estimates. The intrinsic variation always dominates the standard error of prediction, and it remains quite large (relative to the size of the treatment effects) even as the sample size gets huge.

To tighten the intervals further, one would need to reduce the intrinsic variation. This would require either collecting more covariate data or finding ways to better use existing covariates to characterize the conditional effect distribution. Thus, even if rich covariate data are not needed for internal validity, this application shows the crucial role of covariate data in informing decisions that rely on external validity.

## 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 setting of interest using an evidence base from a reference context. 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 randomized and quasi-experiments are run and estimated globally, to our knowledge there is no one design that has been 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, we present 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 data. Fourth, we show that, given the accumulation of sufficient evidence, it is possible to draw externally valid conclusions from our evidence base, 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 for our application, macro-level context covariates dominate. Finally, we considered two applications for our approach. This first showed that experiments located near the middle of the covariate distribution tend to provide the most robust external predictions and that selecting on the maximum covariate Mahalanobis distance is optimal for learning about effect variability. The second 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, but that this depends crucially on the richness of available covariate data.

Prescriptively, we would draw four conclusions from our analysis about extrapolating experimental or quasi-experimental evidence from one setting to another. First, the reference and target setting must be similar along economically relevant dimensions, and particularly in terms of macro level features. In our analysis reference-target covariate differences of half a standard deviation created prediction error on the order of the treatment effect. Second, a sufficiently large experimental evidence base is

needed for reliable extrapolation; for our data, at least fifty country-year samples were needed before out-of-sample extrapolation became reliable. Third, given sufficient data, accounting for treatment effect heterogeneity in the evidence base is essential in extrapolating the treatment effect. Fourth, modeling treatment effect heterogeneity is important when extrapolating treatment effects in sparse data environments; in data-rich settings, rules of thumb might be sufficient.

While our conclusions are cautiously optimistic, it is important to underline both the caution and the inductive nature of our exercise. Our conclusions are circumscribed by the data and application we have considered. Nonetheless, 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 evidence from randomized experiments and quasi-experiments, 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.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2014), "High Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, Volume 28, Number 2, pp. 29-50.

Bisbee, James, Rajeev Dehejia, Cristian Pop-Eleches, and Cyrus Samii (2016), "Local Instruments, Global Extrapolation: External Validity of the Labor Supply - Fertility Local Average Treatment Effect," *Journal of Labor Economics* (forthcoming).

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.

Cochran, William G. (1954), "The Combination of Estimates from Different Experiments," *Biometrics*, Volume 10, Number , pp. 101-129.

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.

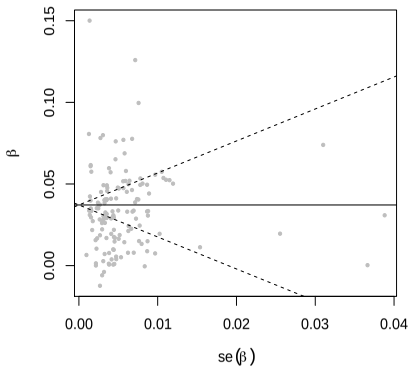
Dehejia, Rajeev (2003), "Was There a Riverside Miracle? A Hierarchical Framework for Evaluating Programs with Grouped Data," *Journal of Business and Economic Statistics*, Volume 21, Number 1, pp. 1-11.

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

Ebenstein, Avraham (2009), "When Is the Local Average Treatment Effect Close to the Average? Evidence from Fertility and Labor Supply," *Journal of Human Resources*, Volume 44, Number 4, pp. 955-975.
Filmer, Deon, Jed Friedman, and Norbert Schady (2009), "Development, Modernization, and Childbearing: The Role of Family Sex Composition," *World Bank Economic Review*, Volume 23, Number 3, pp. 371-398.
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.
Gechter, Michael (2015), "Generalizing the Results from Social Experiments: Theory and Evidence from Mexico and India," manuscript, Pennsylvania State University.
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.
Greene, William H. (2008), *Econometric Analysis, Sixth Edition*, Upper Saddle River, NJ: Pearson/Prentice Hall.
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 (2015), "From Sample Average Treatment Effect to Population Average Treatment Effect on the Treated: Combining Experimental with Observational Studies to Estimate Population Treatment Effects," *Journal of the Royal Statistical Society, Series A*, Volume 178, Number 3, pp. 757-778..
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.
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.

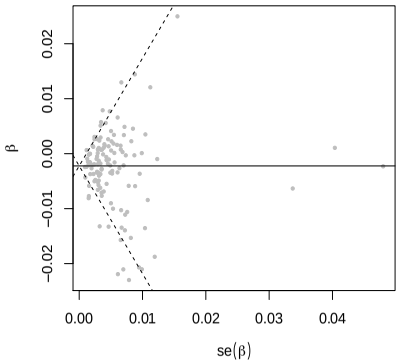
Imbens, Guido W., and Donald B. Rubin (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge: Cambridge University Press.
Jiang, Jiming (2007), *Linear and Generalized Linear Mixed Models and Their Applications*, New York, NY: Springer.
La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert Vishny (1998), "Law and Finance," Journal of Political Economy, Volume 106, pp. 1113-1155.
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.
Newey, Whitney K. (1994), "Series Estimation of Regression Functionals," *Econometric Theory*, Volume 10, Number 1, pp. 1-28.
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.
Rosenbaum, Paul R., and Donald B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, Volume 70, Number 1, 41-55.
Rosenzweig, Mark, and Christopher Udry (2016), "External Validity in a Stochastic World: Evidence from Low-Income Countries," manuscript, Yale University.
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.
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.
Vivalt, Eva (2014), "How Much Can We Generalize from Impact Evaluation Results?", manuscript, New York University.

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



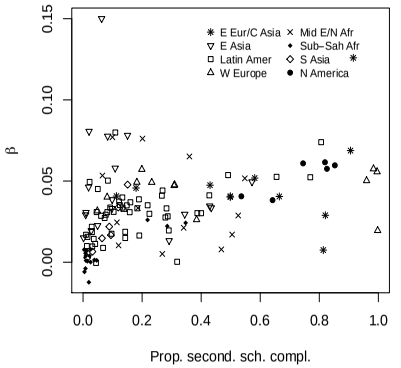

Notes: The funnel plot in this figure is based on data from 142 census samples. Source: Authors' calculations based on data from the *Integrated Public Use Microdata Series -International (IPUMS-I)* .

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



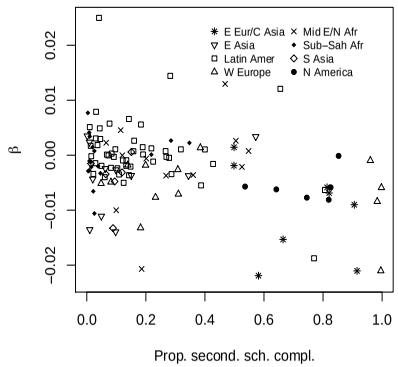

Notes: The funnel plot in this figure is based on data from 128 census samples. Source: Authors' calculations based on data from the *Integrated Public Use Microdata Series -International (IPUMS-I).*

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



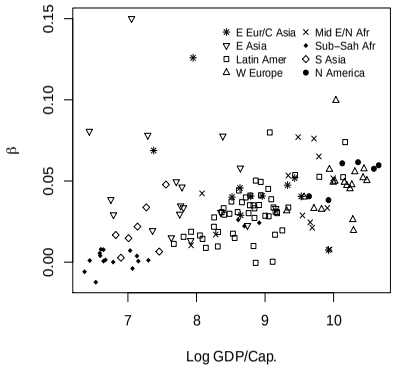

Notes: The graph plots the size of the treatment effect of *Same-Sex* on *Having more children* by the proportion of women with a completed secondary education based on data from 142 census samples. The graph also displays heterogeneity by geographic region. Source: Authors' calculations based on data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



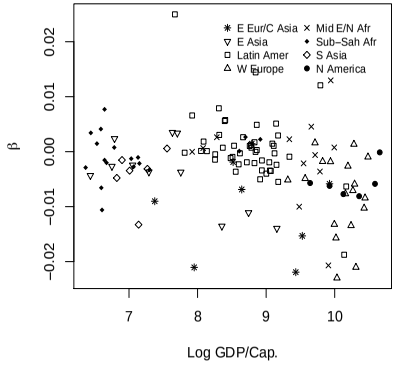

Notes: The graph plots the size of the treatment effect of *Same-Sex* on *Being economically active* by the proportion of women with a completed secondary education based on data from 142 census samples. The graph also displays heterogeneity by geographic region. Source: Authors' calculations based on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



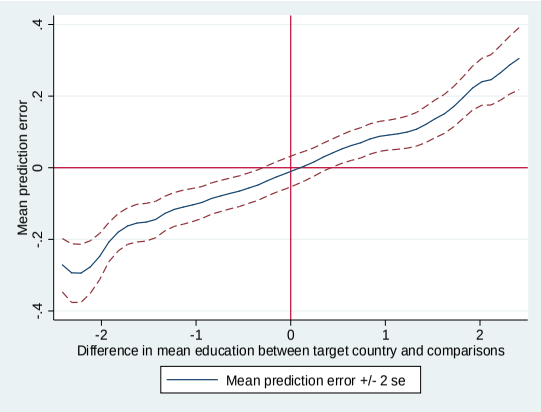

Notes: The graph plots the size of the treatment effect of *Same-Sex* on *Having more children* by log GDP per capita based on data from 142 census samples. The graph also displays heterogeneity by geographic region. Source: Authors' calculations based on data from the *Integrated Public Use Microdata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

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



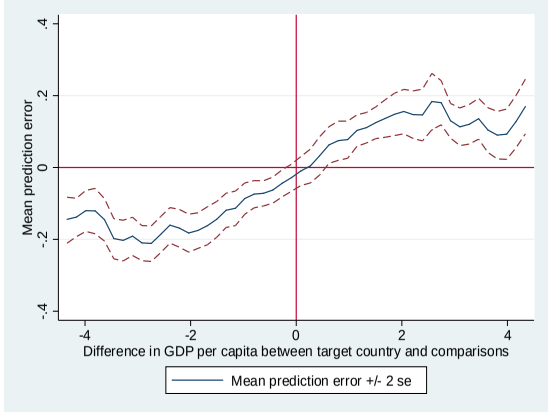

Notes: The graph plots the size of the treatment effect of *Same-Sex* on *Being economically active* by log GDP per capita based on data from 142 census samples. The graph also displays heterogeneity by geographic region. Source: Authors' calculations based on data from the *Integrated Public Use Microdata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

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



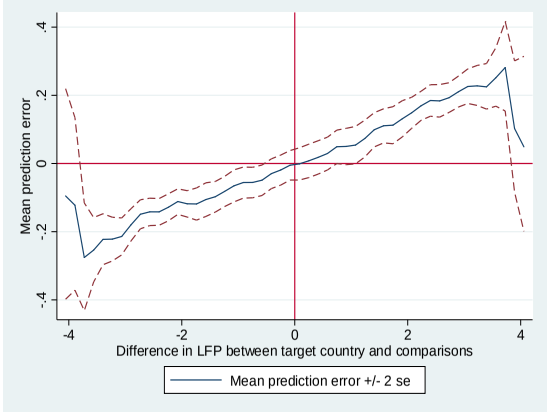

Notes: The graph plots the local polynomial regression of the dyadic prediction error against the standarized education difference between target and comparison country, where the education difference is standardized by its standard deviation (0.83). The variables are further described in Table 1. Source: 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 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).

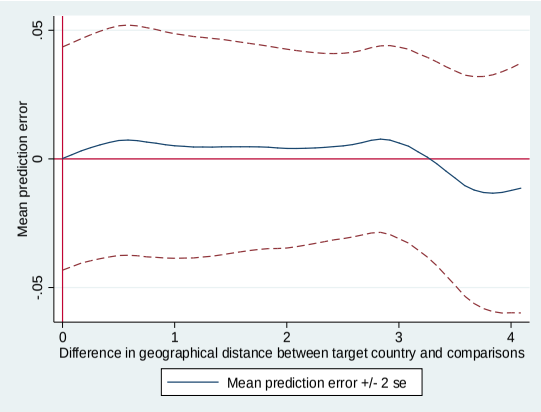
Figure 9: Unconditional external validity function: local linear regression of prediction error on standardized differences in women's labor force participation




Notes: The graph plots the local polynomial regression of the dyadic prediction error 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 10: Unconditional external validity function: local linear regression of prediction error on standardized geographical distance

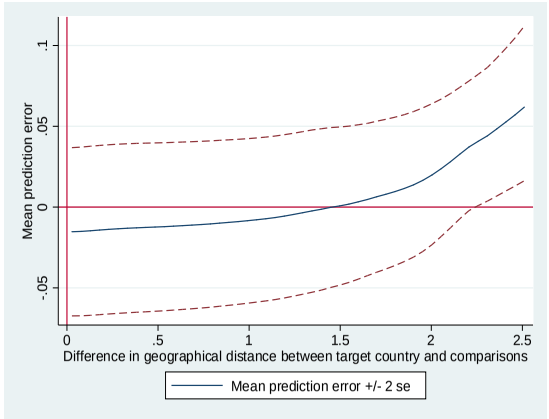
Panel A: All country-year dyads




Notes: The graph plots the local polynomial regression of the dyadic prediction error 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).

Figure 10 (continued)

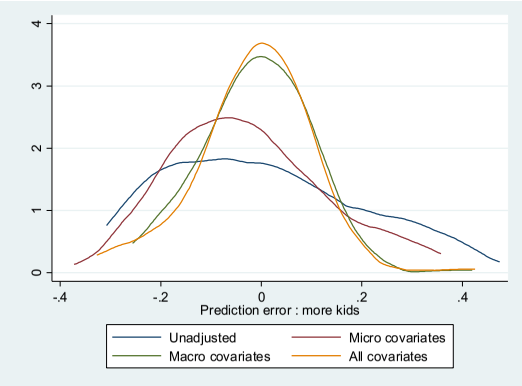
Panel B: All within-region country-year dyads




Notes: The graph plots the local polynomial regression of the dyadic prediction error 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*

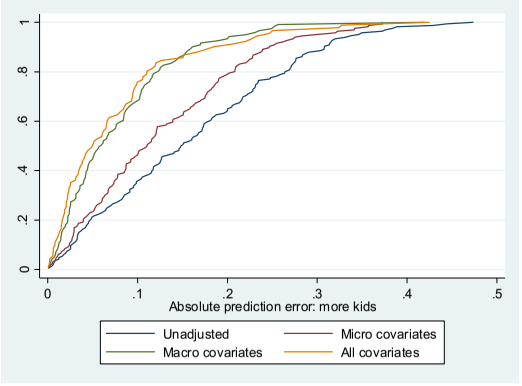
Panel A: Density estimate - prediction error




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 11 (continued)

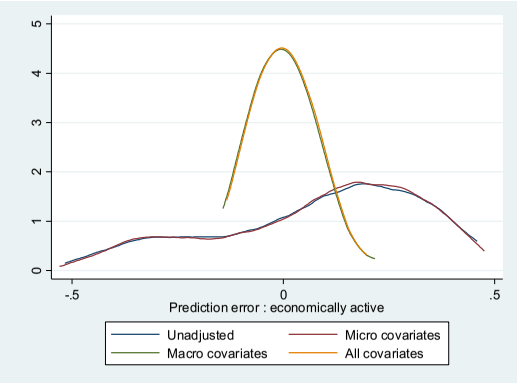
Panel B: CDF - absolute prediction error




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*

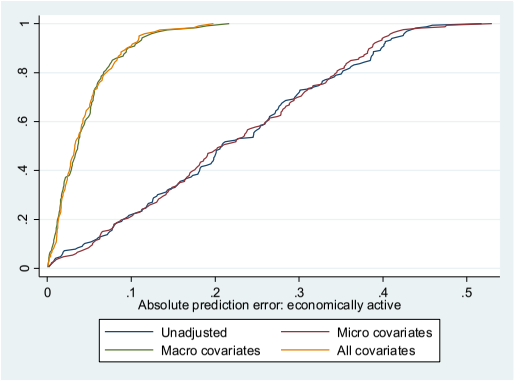
Panel A: Density estimate - prediction error




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).

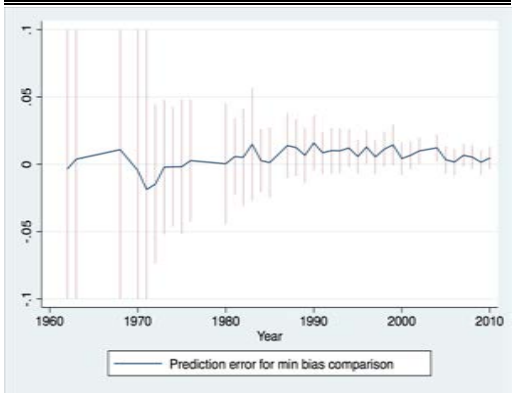
Figure 12 (continued)

Panel B: CDF - absolute prediction error



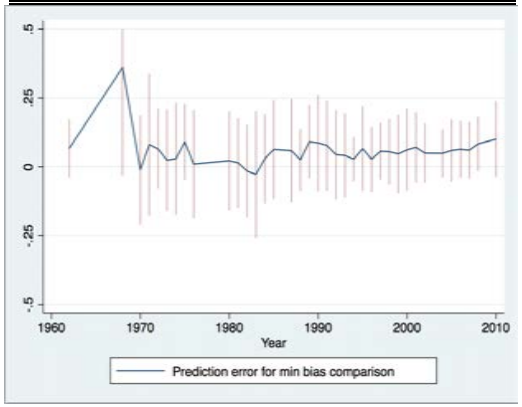

Notes: The graph plots the density estmates of the prediction error and CDF of the absoluteprediction error based on the procedure described in Section 9 of the paper. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

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



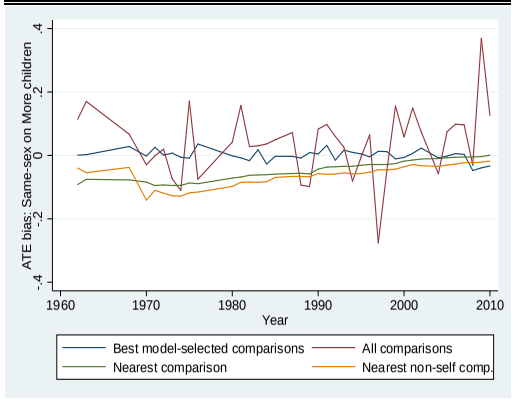

Notes: The graph plots the prediction error over time based on the procedure described in section 9 of the paper. The variable on the X-axis refers to the year when a census was taken. The variables are further described in Table 1. Source: Authors' calculations based on data from the *Integrated Public Use Microdata Series - International (IPUMS-I)* .

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



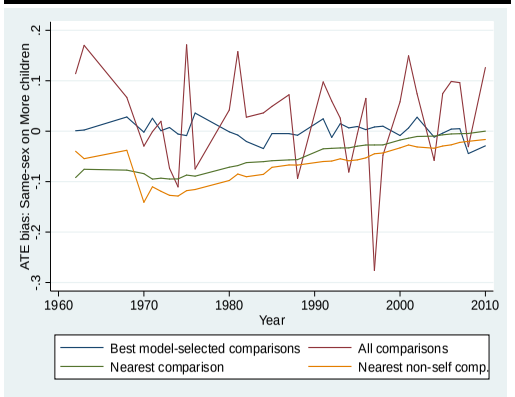

Notes: The graph plots the prediction error over time based on the procedure described in Section 9 of the paper. The variable on the X-axis refers to the year when a census was taken. The variables are further described in Table 1. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

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



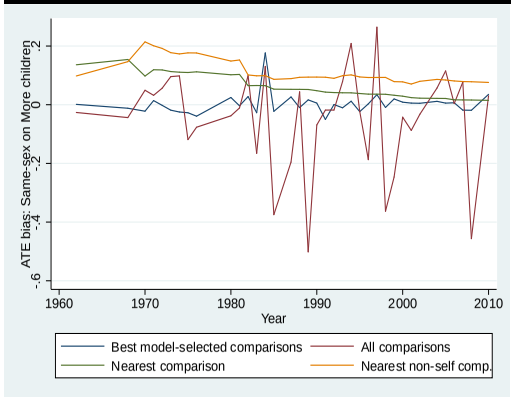

Notes: The graph plots the prediction error for target country-years availabe up to the year on the x-axis using the procedure described in Section 9 of the paper and four groups of reference countries: (1) all the available country years, (graphed as the red line), (2) the best comparison country-year 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*



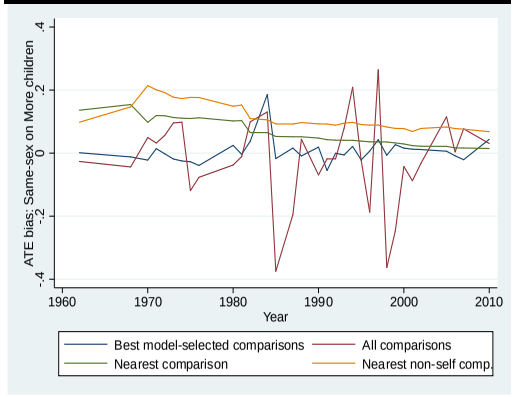

Notes: The graph plots the prediction error for target country-years availabe up to the year on the x-axis using the procedure described in Section 9 of the paper and four groups of reference countries: (1) all the available country years, (graphed as the red line), (2) the best comparison country-year 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



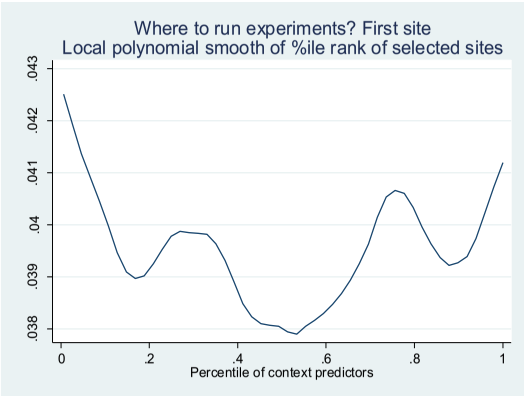

Notes: China, India, Nepal, and Vietnam are excluded from the analysis. The graph plots the prediction error for target country-years availabe up to the year on the x-axis using the procedure described in Section 9 of the paper and four groups of reference countries: (1) all the available country years, (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



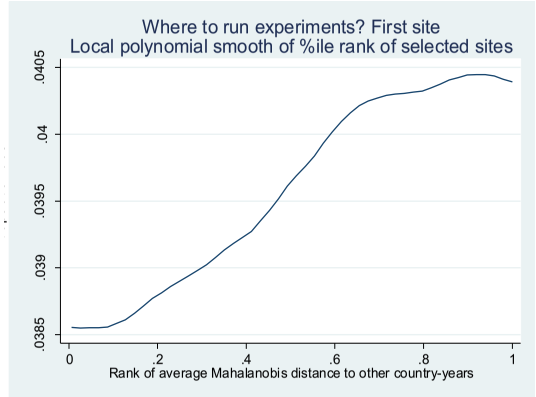

Notes: China, India, Nepal, and Vietnam are excluded from the analysis. The graph plots the prediction error for target country-years availabe up to the year on the x-axis using the procedure described in Section 9 of the paper and four groups of reference countries: (1) all the available country years, (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



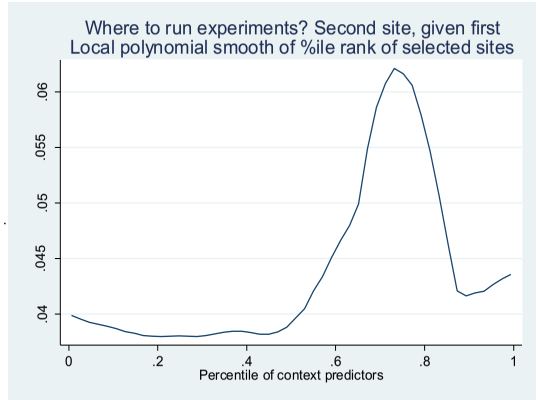

Notes: On the x-axis each country-year is ranked based on its percentile of a composite treatment effect predictor. The composite predictor is a weighted average country-year covariates weighted by their effect on the country-year treatment effect. The y-axis show the mean prediction error from using 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



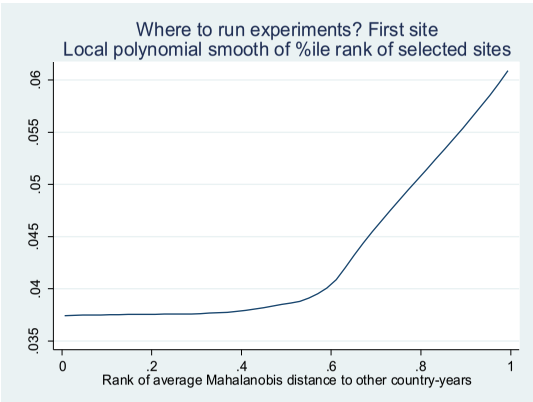

Notes: On the x-axis each country-year is ranked based on its avearge Mahalanobis distance to all other country-years. The y-axis show the mean prediction error from using the site on the x-axis to predict all other country-years. Source: Authors' calculations based on data from the Integrated Public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



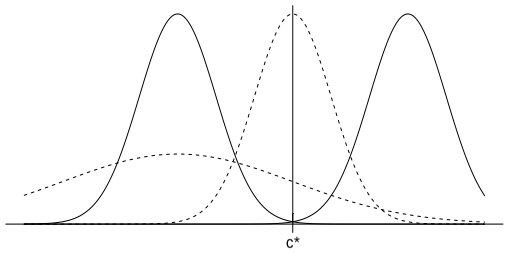

Notes: On the x-axis each country-year is ranked based on its percentile of a composite treatment effect predictor. The composite predictor is a weighted average country-year covariates weighted by their effect on the country-year treatment effect. The y-axis show the mean prediction error from using 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



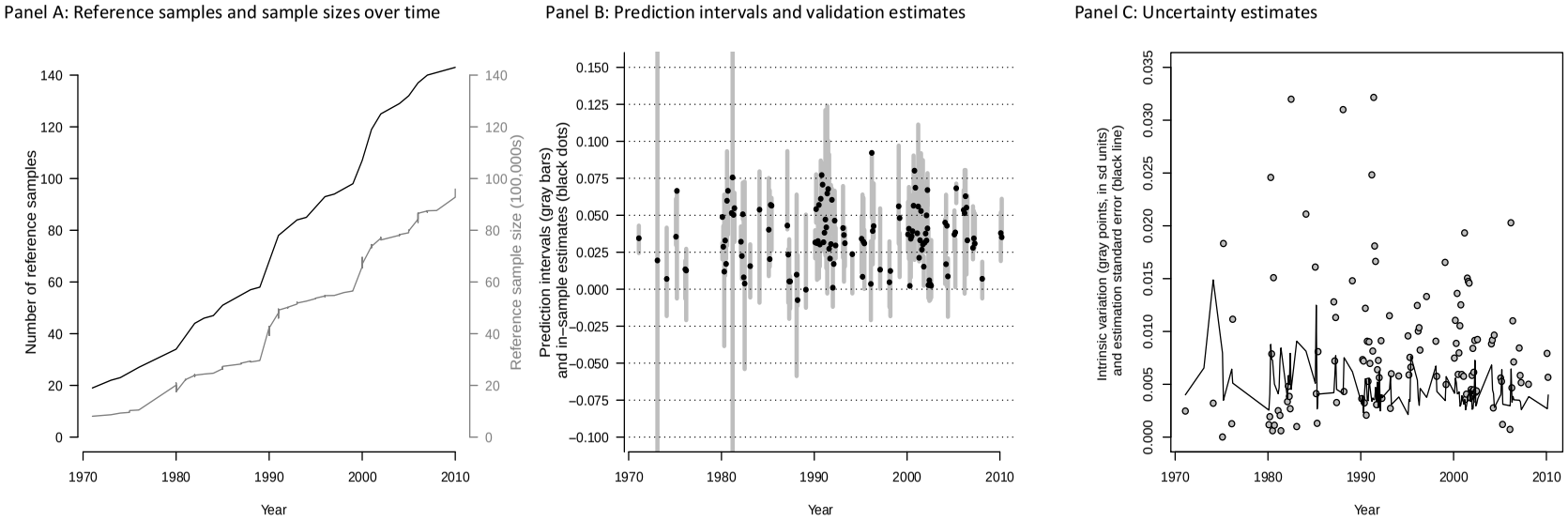

Notes: On the x-axis each country-year is ranked based on its avearge Mahalanobis distance to all other country-years. The y-axis show the mean prediction error from using the site on the x-axis in addition to the first selected comparison site to predict all other country-years. Source: Authors' calculations 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



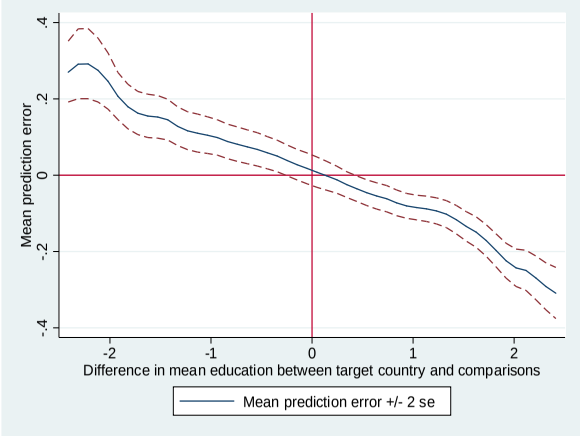

Notes: Solid line = experiment not warranted. Dashed line = experiment warranted.

Figure 24: To experiment or extrapolate? Sample, prediction intervals, and uncertainty estimates



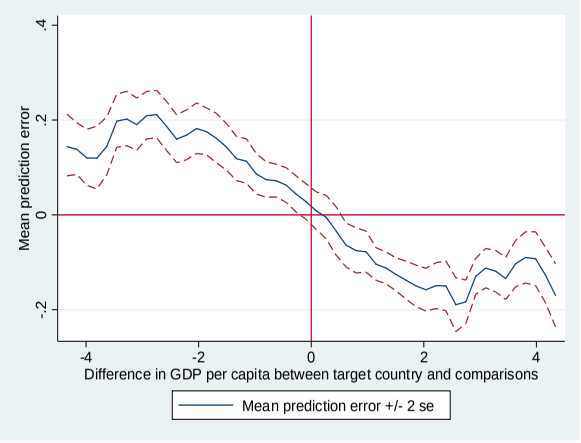

Notes: Panel A shows how the cumulative reference sample evidence base is growing over time in terms of number of country-year samples (black) and number of reference sample observations (gray). Panel B shows the estimated prediction interval for the effect of *Same-sex* on *More kids* for each target country-year (gray bars) and then, for validation, the actual effect estimates from those country-year samples (black dots). Panel C shows the estimation standard error for each target country-year (black line) and then the estimated intrinsic variation, that is, the estiamated standard deviation of the effect distribution at the point in the covariate space for the target country-year. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

Appendix Figure 1: Testing for unconfounded location: local linear regression of Y(0) prediction error on standardized differences in women's labor force participation



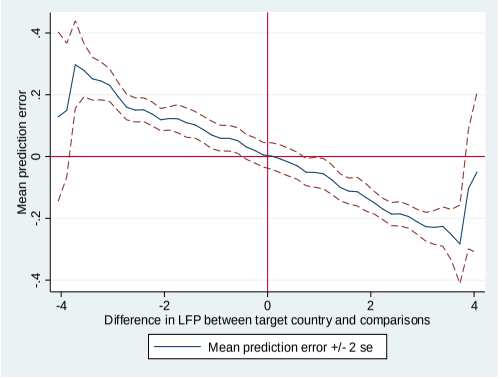

Notes: The graph plots the local polynomial regression of the difference between actual Y(0) and predicted Y(0) against the standardized education difference between target and comparison country, where the education difference is standardized by its standard deviation (0.82). The variables are further described in Table 1. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

Appendix Figure 2: Testing for unconfounded location: local linear regression of Y(0) prediction error on standardized differences in GDP per capita




Notes: The graph plots the local polynomial regression of the difference between actual Y(0) and predicted Y(0) against the standardized difference in GDP per capita between target and comparison country, where the education 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).

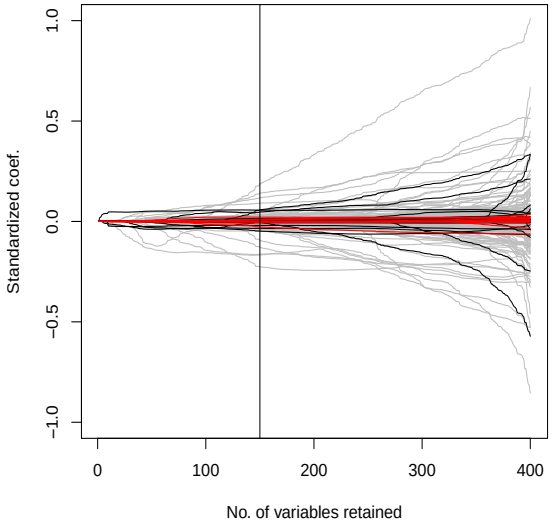
Appendix Figure3 : Testing for unconfounded location: local linear regression of Y(0) prediction error on standardized differences in GDP per capita



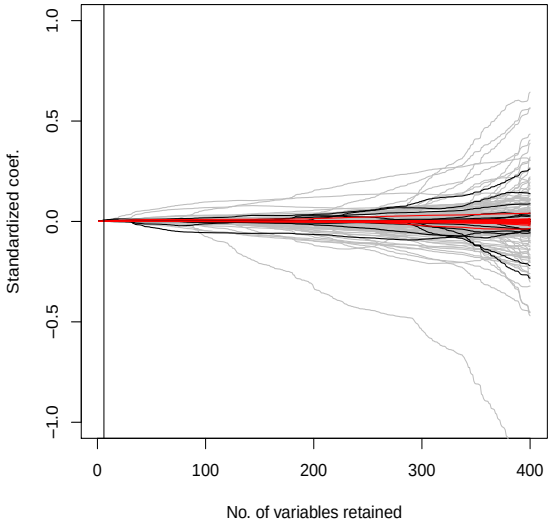

Notes: The graph plots the local polynomial regression of the difference between actual Y(0) and predicted Y(0) against the standardized difference in GDP per capita between target and comparison country, where the education difference is standardized by its standard deviation ($9680). The variables are further described in Table 1. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

Appendix Figure 4: LASSO solution paths for series approximation interaction terms

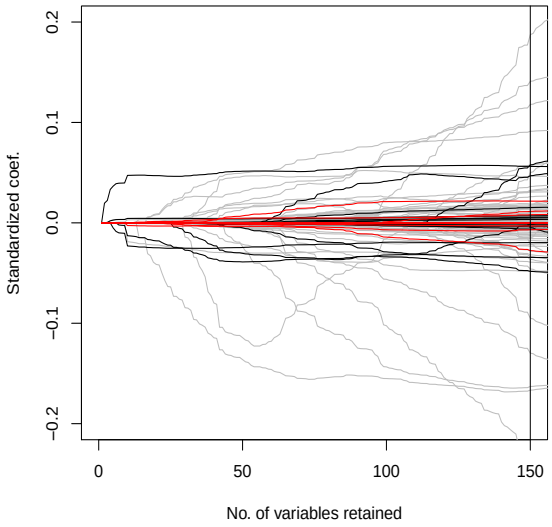
Panel A: Full solution path for "more kids" interaction terms



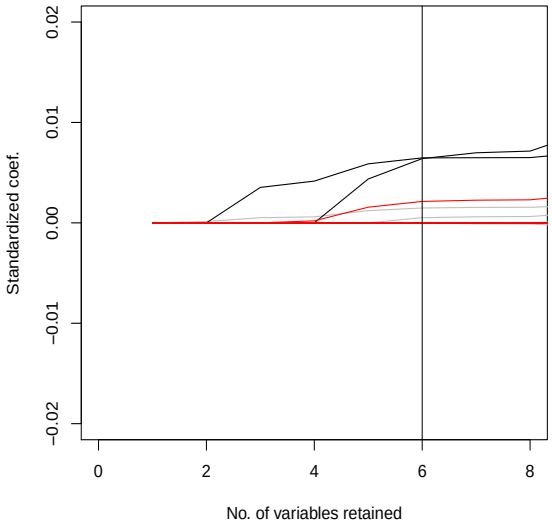

Panel B: Full solution path for "economically active" interaction terms




Panel C: Zoom view of solution path for "more kids" interaction terms




Panel D: Zoom view of solution path for "economically active" interaction terms




Notes: The graphs plot, on the y-axis, standardized coefficient values for treatment-covarate interaction terms in the series approximation for the more kids (left) and economically active (right) outcomes, and on the x-axis, the number of variables retained under LASSO regularization as one loosens the penalty parameter from including only an intercept (at left in each graph) to including all terms in the series (at right in each graph). The black vertical line shows the point at which the specification minimizes Mallow's $C_p$-statistic. Panels A and B show the full solution path through the full saturated second-order series expansion, while panels C and D zoom to the neighborhood where Cp is minimized. Micro-level covariates are colored red, macro-level covariates are colored black, and macro-micro interactions are colored gray for the lines drawing out the coefficient values in the solution paths. Source: Authors' calculations based on data from the *Integrated Public Use Microdata Series-International (IPUMS-I)* .

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 | 166 |
| Age | 20.70 | 1.06 | 166 |
| Labor force participation (women with one child) | 0.51 | 0.21 | 166 |
| Sex imbalance between boys and girls | 0.02 | 0.02 | 166 |
| *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. Degrees of freedom are 141 for More kids and 127 for Economically active.
***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: Extrapolation prediction error regressions for *Having more children* - with covariates

| | Prediction error | Prediction error | Prediction error | Prediction error | Prediction error | Prediction error | Prediction error | Prediction error | Prediction error | Prediction error | Prediction error |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Standardized Difference between country pairs in: | | | | | | | | | | | Excluding sex selectors |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) |
| Education of mother | 0.102*** | | | | | | | | | -0.0315 | -0.0265 |
| ($\sigma$=0.83) | (0.0119) | | | | | | | | | (0.0198) | (0.0213) |
| Education of father | | 0.112*** | | | | | | | | -0.00990 | -0.0120 |
| (σ=0.84) | | (0.0154) | | | | | | | | (0.0224) | (0.0236) |
| Age of mother | | | 0.515*** | | | | | | | 0.318*** | 0.308*** |
| (σ=0.83) | | | (0.0245) | | | | | | | (0.0343) | (0.0360) |
| Census year | | | | 0.0534*** | | | | | | 0.0139*** | 0.0140*** |
| (σ=11.5) | | | | (0.00636) | | | | | | (0.00371) | (0.00382) |
| log GDP per capita | | | | | 0.0645*** | | | | | -0.0207*** | -0.0188*** |
| (σ=9680) | | | | | (0.00840) | | | | | (0.00653) | (0.00680) |
| Sex ratio imbalance | | | | | | 0.00217 | | | | 0.0159** | 0.0232** |
| (σ=0.02) | | | | | | (0.0117) | | | | (0.00766) | (0.0100) |
| Labor force participaiton | | | | | | | 0.0678*** | | | -0.00786 | -0.0113* |
| (σ=0.22) | | | | | | | (0.00898) | | | (0.00625) | (0.00652) |
| Total fertiltiy rate | | | | | | | | -0.302*** | | -0.255*** | -0.253*** |
| (σ=1.54) | | | | | | | | (0.0150) | | (0.0154) | (0.0161) |
| Distance in KM | | | | | | | | | 0.0159 | 0.00872 | 0.00820 |
| (σ=4809) | | | | | | | | | (0.0271) | (0.0106) | (0.0112) |
| Distance squared | | | | | | | | | -0.00454 | -0.00232 | -0.00182 |
| | | | | | | | | | (0.00759) | (0.00346) | (0.00371) |
| Constant | -0.0107 | -0.00596 | -0.00465 | 0.00132 | -0.0124 | -0.000375 | -0.00290 | -0.00327 | -0.00980 | -0.00424 | -0.00323 |
| | (0.0168) | (0.0169) | (0.0117) | (0.0181) | (0.0166) | (0.0177) | (0.0169) | (0.0115) | (0.0179) | (0.0104) | (0.0107) |
| Observations | 28,561 | 28,561 | 28,561 | 28,561 | 27,556 | 28,561 | 28,561 | 28,561 | 28,561 | 27,556 | 24,025 |
| R-squared | 0.184 | 0.141 | 0.549 | 0.117 | 0.173 | 0.000 | 0.148 | 0.638 | 0.000 | 0.723 | 0.724 |

Notes: The table shows extrapolation prediction error regressions as described in Sections 3 and 8 of the paper. The left-hand-side variable is reference-to-target prediction error in the country-year dyad. The right-hand-side variables are standardized referene-to-target differences in covariates, where the standardization is given in parentheses. Standard errors are clustered at the target-country-year level. Column 11 excludes China, India, Vietnam, and Nepal. Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

Table 4: Extrapolation prediction error regressions for *Being economically active* - with covariates

| | Prediction error | Prediction error | Prediction error | Prediction error | Prediction error | Prediction error | Prediction error | Prediction error | Prediction error | Prediction error | Prediction error |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Standardized Difference between country pairs in: | | | | | | | | | | | Excluding sex selectors |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) |
| Education of mother | -0.0243 | | | | | | | | | 0.104*** | 0.100*** |
| ($\sigma$=0.83) | (0.0180) | | | | | | | | | (0.0144) | (0.0155) |
| Education of father | | -0.0140 | | | | | | | | -0.0954*** | -0.0858*** |
| ($\sigma$=0.84) | | (0.0191) | | | | | | | | (0.0154) | (0.0156) |
| Age of mother | | | -0.247*** | | | | | | | -0.0126 | -0.000542 |
| ($\sigma$=0.83) | | | (0.0647) | | | | | | | (0.0386) | (0.0379) |
| Census year | | | | -0.0824*** | | | | | | -0.0116** | -0.0103** |
| ($\sigma$=11.5) | | | | (0.00969) | | | | | | (0.00439) | (0.00428) |
| log GDP per capita | | | | | -0.00477 | | | | | 0.0498*** | 0.0452*** |
| ($\sigma$=9680) | | | | | (0.0103) | | | | | (0.00348) | (0.00355) |
| Sex ratio imbalance | | | | | | 0.0518** | | | | -0.00407 | 0.00604 |
| ($\sigma$=0.02) | | | | | | (0.0196) | | | | (0.0111) | (0.0125) |
| Labor force participaiton | | | | | | | -0.191*** | | | -0.214*** | -0.213*** |
| ($\sigma$=0.22) | | | | | | | (0.00861) | | | (0.00587) | (0.00597) |
| Total fertiltiy rate | | | | | | | | 0.148*** | | 0.000714 | -0.00155 |
| ($\sigma$=1.54) | | | | | | | | (0.0364) | | (0.0225) | (0.0215) |
| Distance in KM | | | | | | | | | 0.0403 | 0.0114 | 0.0174 |
| ($\sigma$=4809) | | | | | | | | | (0.0349) | (0.0107) | (0.0115) |
| Distance squared | | | | | | | | | 0.00221 | -0.00182 | -0.00375 |
| | | | | | | | | | (0.00972) | (0.00186) | (0.00238) |
| Constant | 0.0657* | 0.0662* | 0.0513 | 0.0565* | 0.0668* | 0.0658* | 0.0296* | 0.0503 | -0.00918 | 0.0125 | 0.0123 |
| | (0.0385) | (0.0395) | (0.0308) | (0.0327) | (0.0397) | (0.0391) | (0.0164) | (0.0313) | (0.0332) | (0.0157) | (0.0151) |
| 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 | 0.006 | 0.001 | 0.069 | 0.160 | 0.001 | 0.022 | 0.735 | 0.086 | 0.031 | 0.825 | 0.816 |

Notes: The table shows extrapolation prediction error regressions as described in Sections 3 and 8 of the paper. The left-hand-side variable is reference-to-target prediction error in the country-year dyad. The right-hand-side variables are standardized reference-to-target differences in covariates, where the standardization is given in parentheses. Standard errors are clustered at the target-country-year level. Column 11 excludes China, India, Vietnam, and Nepal.
Source: Authors' calculations based on data from the Integrated Public Use Microdata Series-International (IPUMS-I).

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).