Three Essays on Government Health Expenditure

Joseph L Dieleman

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Hendrik Wolff, Chair

Michael Hanlon

Phillip Brock

Program Authorized to Offer Degree:

Economics

University of Washington

Abstract

Three Essays on Government Health Expenditure

Joseph L Dieleman

Chair of Supervisory Committee:

Hendrik Wolff

Economics

In 2010, governments spent over $4.05 trillion on health. This is 7% of the world economy. This dissertation measures the health outcomes achieved by this spending, how this spending changes upon the receipt of health aid, and the methods used to analyze this spending. The first chapter assesses the effect of government health expenditure on health outcomes. I use a large cross-country panel and a new instrument to test how domestic government health expenditure affects under-five mortality. I find that the average spending-to-outcome elasticity is -0.34, although it ranges between -0.04 and -0.61 depending on country-level characteristics. Countries with larger GDP per capita and more civil liberties, political rights, and democracy have the elasticities furthest from zero, although countries with the largest under-five mortality rates have the largest effects when measured in deaths averted. The second chapter assesses the crowding out of domestic government health expenditure upon the receipt of development assistance for

health channeled to the government. Using a cross-country dynamic panel, I use Difference GMM to simultaneously measure the displacement and replacement rates. I find that increases in development assistance lead to the displacement of government resources, but the reduction in development assistance does not lead to the replacement of resources. Combined with the estimates from the first chapter, these findings suggest that net effect of health aid is less than previously measured. The third chapter is an assessment three linear regression estimation methods used to control for unobserved heterogeneity that exists in most clustered data. This type of heterogeneity exists in most analyses of government health expenditure. This chapter explains the standard clustered data model, the traditional random- and fixed-effects estimators, and the 'within-between' estimator – a variant of a model proposed by Mundlak in 1978. Simulation is used to illustrate when each estimator is optimal. The chapter ends with recommendations for when each estimator is preferred.

University of Washington

Acknowledgements

Three Essays on Government Health Expenditure

Joseph L Dieleman

This work would not have been possible if it were not for the generous support of my family, friends, coworkers, and advisors. First and foremost, I am thankful to my partner Rachel whose grace, patience and support made even the lows of this process endurable; I am thankful to my parents for instilling in me a work ethic that made graduate study possible and a compass that makes studying real problems appealing; and I am thankful to my classmates, coworkers, and friends that helped make this work possible and even enjoyable. Thanks in particular to Dan, Monica, John, Paula, Kirk, Skylar, Joelle, Justin, Ella, James, and Seth. I am also grateful to advisors and mentors who helped shape these ideas and reviewed my work. Thanks in particular to Michael Hanlon, whose mentoring I have not taken for granted, and Christopher Murray, whose abilities to understand a problem and the global health context are uncanny. Furthermore, I am grateful for the additional members of my dissertation supervisory committee, Hendrik Wolff, Phillip Brock, and Robert Plotnick, who have also been exceedingly helpful and flexible. I am also appreciative for five years of financial support from the University of Washington's Department of Economics and the Institute for Health Metrics and Evaluation. Finally and most importantly, I am thankful to my Lord, Savior, and Creator for inexplicable privilege and opportunity.

Index

Introduction

Government health expenditure is a substantial means of financing health throughout the world. In 2010, governments spent roughly $4.05 trillion (2010 US dollars) on health. This is 7% of the global gross domestic product (GDP). 85% of this expenditure was in high-income countries, although low- and middle-income governments also spent relatively large amounts on health. On average, low- and middle-income governments spend 3% of their GDP and 10% of their government expenditure on health (World Health Organization, 2012; James et al., 2012).

These sizable expenditures reflect many years of growth. Over the last one and a half decades most countries, and all region aggregates, show nearly monotonic growth in government health expenditure measured in real US dollars and in real US dollars per capita. This growth reflects a clear notion that government health expenditure is a valuable commodity, often providing services for the poorest and implementing necessary public health systems (Gupta et al., 2003). This growth also reflects governments' efforts to achieve goals set by the Commission on Macroeconomics and Health and the Abuja Declaration (Sachs, 2001; Abuja Declaration, 2001).

According to the World Health Organization, health expenditure is "all expenditures for activities whose primary purpose is to restore, improve and maintain health for the nation and for individuals during a defined period of time (World Health Organization, 2003)." Government health expenditure, the portion of total health expenditure spent by any level of government, can be broken into two distinct portions: domestically and externally generated resources. The domestically generated portion of government health expenditure is government health

expenditure as source (GHES); while the funds that originate externally are known as development assistance for health channeled to the government (DAHG). Together GHES and DAHG make up total government health expenditure (GHE). By definition, GHE ≡ GHES + DAHG. For all geographic regions of the world DAHG is a very small fraction of GHE, with DAHG totaling only $4.20 billion (2010 US dollars). Still, for low-income countries, DAHG makes up 18% of GHE.

This dissertation is an exploration of GHES and the econometric methods used to study GHES. The first two chapters use a cross-country panel to address potential determinants and health outcomes associated with GHES. The first chapter assesses the effect that GHES has on health. A review of this literature shows contradictory conclusions. Prior studies are limited by small datasets, generally do not take advantage of panel data, do not disaggregate GHE by source, and many do not account for simultaneity. In this first chapter, I create a large panel spanning 186 countries over 16 years and apply a novel instrument. I find that on average a 1% increase in GHES leads to a 0.34% decrease in under-five mortality, although this estimate ranges from 0.04 to 0.61 depending on country-level characteristics. Countries with larger GDP per capita and more civil liberties, political rights, and democracy have the elasticities furthest from zero, although countries with the largest under-five mortality rates have the largest effects when measured in deaths averted. These results show that economic development, democratization, and government health spending can work in concert to significantly reduce the burden of premature mortality.

The second chapter assesses the dynamic relationship between DAHG and GHES. This chapter has three primary conclusions. First, increases in DAHG lead to decreases in GHES. Governments substitute gains in resources caused by development assistance across sectors, even when development assistance is targeted for the health sector. Because aid is fungible, health aid crowds out government health spending. Second, this effect is persistent, meaning that short-run DAHG shocks have effects that persist beyond the period of the shock. Third, I show that one reason for the persistent DAHG effect is that this effect is asymmetric. The displacement caused by increases in DAHG is greater (in absolute terms) than the replacement of GHES caused by decreases in DAHG. This asymmetry has many practical implications, but most importantly, suggests that DAHG fluctuations can lead to financial gaps in the intermediate-term.

Finally, my third chapter uses simulation to evaluate three econometric methods used to control unobserved heterogeneity. Unobserved heterogeneity is prevalent in most clustered data and certainly prevalent in GHES data. The most common methods for dealing with unobserved heterogeneity are random- and fixed-effect estimation. A series of searches on academic databases suggests that health science researchers prefer random-effects estimation, while economists and political scientists prefer fixed-effects estimation. This chapter is meant to explore when each estimator might be most appropriate. In addition to exploring the traditional random- and fixed-effects estimators, I also explore a specification derived from Mundlak's 1978 Econometrica paper. His thesis is that when properly specified, fixed- and random-effects are asymptotically equivalent. I evaluate a variant of his specification, called the 'within-between' estimator, along with the traditional estimator, using 16,200 unique scenarios. These simulations show that in finite samples there are cases when each estimator is most appropriate,

although the cases in which random-effects estimation is preferred are less common and reserved for a few special cases. Fixed-effects estimation is generally preferred to random-effects, although in finite samples the within-between specification generally out-preforms both of the traditional estimators. The within-between specification, which is unbiased and much more flexible than the traditional fixed-effects estimator, marries the best characteristics of the two traditional estimators.

References

Abuja Declaration. 2001. Abuja Declaration on HIV/AIDS, Tuberculosis and Other Related Infectious Diseases. April 2001. Available at: http://www.un.org/ga/aids/pdf/abuja_declaration.pdf.

Gupta S, Verhoeven M, Tiongson ER. Public spending on health care and the poor. Health Economics 2003; 12; 685–696.

James SL, Gubbins P, Murray CJ, Gakidou E. Developing a comprehensive time series of GDP per capita for 210 countries from 1950 to 2015. Population Health Metrics 2012; 10; 12.

Sachs JD. Macroeconomics and Health: Investing in Health for Economics Development. World Health Organization; 2001.

World Health Organization. Guide to Producing National Health Accounts 2003.

———. World Health Statistics 2012. World Health Organization: Geneva; 2012.

Chapter 1: The effect of government health expenditure on under-five mortality

Short title: Government health expenditure

Version: June 5, 2013

Author: Joseph L. Dieleman, M.A.
Department of Economics, University of Washington
Institute for Health Metrics and Evaluation, University of Washington

Correspondence: Joseph L. Dieleman
dieleman@uw.edu
Institute for Health Metrics and Evaluation
2301 5th Avenue, #600
Seattle, WA 98121 USA
+ 1 (206) 897-2800 (telephone)
+ 1 (206) 897-2899 (fax)

Abstract

The literature exploring the effectiveness of government health spending has been muddled by varied estimates. I explore this issue by addressing two potential reasons for the diverse estimates – ill-powered sample size and a simultaneity problem. I construct a panel containing 2,976 country-years of data and use a novel instrument to control for measurement error and reverse causation. I show that on average a 1% increase in domestic government health expenditure leads to a 0.34% decrease in under-five mortality. Moreover, this effect is heterogeneous. Countries with larger gross domestic product (GDP) per capita or more civil liberties, political rights and democracy have significantly larger health spending elasticities. Applying my point estimates to 2010 data suggests countries' elasticities range from -0.02% to -0.61%, depending on each country's policies and level of economic development. This result holds across various specifications and health indicators. Furthermore, I show my instruments are intuitively plausible and pass a series of empirical tests. While the elasticity is closest to zero in low-income countries, these results are particularly important for those countries because under-five mortality tends to be highest there. These results show that economic development, democratization, and government health spending can work in concert to significantly reduce the burden of premature mortality.

Keywords: government health expenditure, under-five mortality

1. Introduction

There is a popular consensus amongst health advocates that government health spending should improve population health. Low- and middle-income countries have been encouraged to increase government health expenditure (Sachs, 2001) and even private sector advocates recognize the need for some government regulation in the health sector (Hanson et al., 2008). Furthermore, known cost-effective interventions have not been universally adopted in many countries (Jamison et al., 2006a; Jamison et al., 2006b). By scaling up coverage of these interventions, government health expenditure appears capable of acquiring significant health gains and reducing the burden of disease.

A large literature has measured how government health expenditure translates into population health gains. However, the estimates produced by that research are varied and in some cases contradictory. This debate, which began in the 1990s, has continued into the current decade. How aggregate government health spending influences health remains unknown.

In this paper, I use a novel instrument and a significantly larger dataset to measure how government health spending affects under-five mortality. My data spans 186 countries and 16 years (1995-2010). My variable of interest is government health expenditure net of development assistance channeled to the government health sector. I call this variable government health expenditure as source (GHES). I prefer GHES over total government health expenditure (GHE) because foreign aid expended by the government and domestically sourced spending are not fully fungible, often have distinct objectives, and likely have unique health effects (Lu et al.,

2010; Farag et al., 2009; Fernandes Antunes et al., 2012; Sijpe, 2012; Piva and Dodd, 2009). I employ government education expenditure as an instrument for GHES. By simultaneously controlling for the effect of education outcomes on health outcomes, I ensure that my instrument is exogenous. (The Hansen-J test, difference-n-Sargen test, and including the excluded instruments directly in the second stage of the estimation confirm this.) Furthermore, this instrument is not 'weak.' When the excluded instruments are tested in my preferred model, the F-statistic ranges from 16.4 to 324.0.

I estimate the influence of GHES on under-five mortality and find a significantly negative effect. I show these findings are robust across a variety of health indicators, model specifications, and currency conversion methods. Across 186 countries from 1995 to 2010, I estimate a 1% increase in GHES per capita resulted in a 0.34% decrease in under-five mortality. Furthermore, I show that this effect is heterogeneous. Countries with more civil liberties, political rights and democracy have significantly larger spending-to-outcome elasticities (in absolute terms). Similarly, high-income countries achieve a larger elasticity as well. Applying my point estimates to 2010 data suggests that elasticity ranges from -0.02% to -0.61% depending on country characteristics. While the elasticity is closest to zero in many low-income countries, these results are particularly important those countries. In these countries, government health spending per capita is lowest and under-five mortality is highest. Thus, significant gains in averting of under-five deaths are a distinct possibility.

2. Context

This literature has competing perspectives on whether government health expenditure influences health outcomes. Table 1 lists the most commonly cited cross-country analyses that investigate this topic. Two eminent papers from the 1990s champion the two competing perspectives. The first of these two is by Anand and Ravallion (1993), which analyzes public health expenditure's effect on mortality for 22 low- and middle-income countries. Anand and Ravallion conclude that a 1% increase in public health spending decreases premature death by 0.3%. A varied set of studies support these results. Gupta et al. (2002) and Bokhari and Gottret (2007) come to remarkably similar conclusions estimating elasticities of -0.29% and -0.33%, respectively.

A number of studies test and find that GHE affects health, although the effect is heterogeneous and in some cases only present for some countries or subpopulations. Bidani and Ravallion (1997), Or (2000), Wagstaff (2003) and Gupta et al. (2003) use cross-country data but disaggregate national populations into subpopulation with their own distinct elasticities. Three of these studies show GHE has the largest effects on the poor (relative to the non-poor) (Bidani and Ravallion, 1997; Wagstaff, 2003; Gupta et al., 2003). Or (2000) assesses distinct effects across sexes in high income countries, and finds nearly identical elasticities for women and men, -0.17% and -0.18% respectively. Rajkumar and Swaroop (2008) also assess heterogeneous health effects of GHE. Rather than looking at subpopulations within countries, Rajkumar and Swaroop measure heterogeneity across countries. They show on average a 1% increase in GHE leads to a 0.18% decrease in under-five mortality, but that this elasticity grows (shrinks) in absolute terms as a country has less (more) corruption.

Representing an alternative perspective, Filmer and Pritchett (1999) concludes public expenditure on health explains less than one-seventh of 1% of the variation in mortality in low- and middle-income countries. Earlier work by Musgrove (1996) and Jamison (1993) agree, with Musgrove stating, "multivariate estimates of the determinants of child mortality give much the same answer: income is always significant, but the health share in gross domestic product (GDP), the public share in health spending, and the share of public spending on health in GDP never are." More recently, Filmer et al. (2000), Baldacci et al. (2003), McGuire (2006) come to a similar conclusion. Baladacci et al. (2003) shows a traditional model identifies a significant GHE effect, but their preferred latent variable specification does not.

The methods, models, and conclusions in this literature vary, although they share several key features. First, this literature hinges on small datasets. Anand and Ravalian (1993) and Filmer and Pritchett (1999) complete their analyses using less than 100 observations. Small samples size appears to be the norm, rather than the exception in this literature, as even the largest datasets have less than 500 observations. (See Table 1.) Moreover, only one study, which is restricted to high-income countries, takes advantage of the panel data to address unexplained cross-country heterogeneity (Or, 2000).

My second concern is that analyses measure the effect of GHE on health outcomes. Not parsing out foreign aid spent by the government confounds two potentially non-uniform effects. Testing the effect of GHES rather than GHE is important because health aid channeled to the government is not completely fungible and is often intended for specific projects within the

health sector (Lu et al., 2010; Farag et al., 2009; Fernandes Antunes et al., 2012; Sijpe, 2012; Piva and Dodd, 2009).

My third concern is that the previous literature struggles to control for simultaneity (reverse causation) between GHE and health outcomes. While the objective of these studies is to identify how government spending might curb adverse health, it is also clear that adverse health can cause increases in health spending. These countervailing effects will attenuate measurement of the effectiveness of GHE. While much of this literature ignores this problem, two studies formally address it using instrumental variables. I also apply and instrumental variables approach, but rely on an alternative instrument. My experimentation with the instruments relied upon previously leads us to question their validity. (See subsection 4.2 for more details on this mater.)

To summarize the context of this analysis, there are two decades of studies that attempt to identify the effect that GHE has on health outcomes. These studies have identified important heterogeneities, but have not rendered a consensus on the average effect of increasing government health spending. I believe that the disparate perspectives can be explained by data and simultaneity. What follows is an attempt to measure the effect of GHES on health outcomes, utilizing a sample an order of magnitude larger than any sample used previously, while also controlling for simultaneity.

3. Data

For my analysis, I initially consider all 187 countries included in the 2010 Global Burden of Disease (Lozano et al., 2012). I omit Iraq from my analysis because my regression analysis shows it to be a distinct outlier. For the remaining 186 countries, I are able estimate GHES for 1995 through 2010.

3.1 Government health expenditure as source

I collect total government health expenditure (GHE) data from the World Health Organization (WHO) (World Health Organization, 2012). These data are available for most countries starting in 1995. For my sample, 77 country-years of data are not reported by the WHO. I utilize the data reported in nominal local currency units (LCUs). [1] I convert the nominal LCU series into real 2010 LCUs using country-specific GDP deflators, obtained from the International Monetary Fund (IMF) (International Monetary Fund, 2012). For countries that the IMF does not report on, I use deflators reported by the World Bank (The World Bank). I then covert the real LCU series into real purchasing power parity (PPP) dollars using PPP exchange rates published by the WHO (World Health Organization, 2012). 135 country-years of GHE data are lost in currency conversion because of incomplete deflator and exchange rate series.

For each country-year, I disaggregate GHE by subtracting development assistance for health channeled to the government (DAHG). By definition, GHE ≡ GHES +DAHG (World Health

[1] The WHO also reports GHEA in an assortment of nominal currencies and units, such as US dollars per capita and international dollars per capita. Since the WHO documentation does not explain how it conducted the currency conversions, I use the most basic (untransformed) form of data, which are the LCUs.

Organization, 2003; Lu et al., 2010). Figure 1 shows that for low-income countries, DAHG is sizable and worth considering apart from GHES. DAHG is obtained from Financing Global Health 2012 (Institute for Health Metrics and Evaluation, 2013). DAHG data is reported in 2010 US dollars, converted to 2010 PPP dollars, and exists through 2010 for 155 of my 186 countries. The remaining 31 countries for which DAHG is not available are high-income countries, which generally receive DAHG only for research and development. For these 31 countries, I assume DAHG spent on health interventions equals zero, and GHES equals GHE.

3.2 Health outcomes

I employ the probability of dying (per 1,000 people) under the age of five as my health metric and dependent variable. I choose under-five mortality for three reasons. First, this choice is common in the literature, and thus my conclusions are comparable to previous works (Filmer and Pritchett, 1999; Gupta et al., 2003; Wagstaff, 2003; McGuire, 2006; Bokhari et al., 2007; Baldacci et al., 2008; Rajkumar and Swaroop, 2008). Second, under-five morality is more accurately measured than adult mortality; and third, it has been shown to be more responsive to changes in policy and socioeconomic conditions (Filmer et al., 1999; McGuire, 2006). I use under-five mortality data estimated for the Global Burden of Disease 2010 project (Rajaratnam et al., 2010; Wang et al., 2012)

To test the robustness of my findings, I consider a number of additional health outcomes. These alternative dependent variables are: neonatal, post-neonatal, adult female, and adult male mortality rates and the malnutrition rate. Table A1 reports the sources of these data.

3.3 Missingness

I have perfectly complete GHES records the high-income countries belonging to the Organisation of Economic Cooperation and Development. On the other hand, I have 212 GHES missing observations for primarily low- and lower middle-income countries. Thus, missingness is correlated with economic development. To avoid selection bias resulting from these missing observations, I employ multiple imputation (MI). MI is a "best practice" approach to dealing with missing values, as it is robust, unbiased, and efficient (Rubin, 2004). MI is increasingly utilized in economics and is well established in other quantitative disciplines. In MI, missing values are predicted conditional on the observed values. Rather than computing a single imputation, MI creates a set of imputed data sets. Each unique data set is complete and an estimate of the true series.[2] I depend on a set of 35 covariates for my MI procedure. Table A1 in the appendix lists my covariates, their sources, and which covariates are log-transformed for MI. For my regression analysis, I analyze 20 imputed data sets and then aggregate coefficients according to rules set by Rubin (2004). These rules incorporate variation within and across the imputed data.

[2] There is a diverse set of algorithms used to estimate missing values. I rely on an expectation maximization (EM) algorithm that employs bootstrapping, a process developed by King *et al.* (2001). EM is an iterative method used for maximizing complex likelihood functions. The EM with bootstrapping provides replicable imputation for cross-sectional, time-series, or panel data sets containing missing values (Honaker *et al.,* 2010). For my analysis, MI is operationalized in *Amelia II: A Program for Missing Data* (Honaker *et al.,* 2011). The algorithm assumes that missing observations are missing at random (MAR), meaning that conditional on all the covariates, missingness is randomly distributed. The MI procedure also assumes a multivariate normal distribution across the entire data set. To better approximate this distribution, GHES per capita, and the majority of our covariates, enter our imputation process log-transformed.

4. Measuring the Effect of GHES on Health

4.1 Primary Estimation

A principal issue in this literature is whether government health expenditure has a significant health effect even when controlling for economic development. To control for economic development, I include GDP per capita and amount of maternal education attained. I also include DAHG, development assistance for health not channeled to the government (DAHNG), GGE per capita net of GHE and HIV prevalence. The HIV epidemic has been shown to have large effects on many countries, especially those in sub-Saharan Africa (Adetunji, 2000). All financial variables are in 2010 PPP per capita terms and are natural log transformed to account for diminishing marginal returns.

My baseline model is represented by equation (1), in which 5q0 represents under-five mortality rate, GHES is government health expenditure as source, DAHG is development assistance for health channeled to the government, DAHNG is development assistance for health not channeled to the government, GDP is gross domestic product, EDU is the average number of years of education attainment for women 25 years and older and HIV is the HIV prevalence. Unobserved country-specific time-invariant affects are represented by α, while time shocks and technology are controlled by γ. GHES for many countries has increased steadily over my time. Not controlling for time would conflate time, technology, and GHES.

$$5q0_{it} = \beta_1 GHES_{it} + \beta_2 DAHG_{it} + \beta_3 DAHNG_{it} + \beta_4 GGE_{it} + \beta_5 GDP_{it} + \beta_6 EDU_{it} + \beta_7 HIV_{it} + \alpha_i + \gamma_t + \upsilon_{it} \quad (1)$$

I control for the possibility of reverse causation by instrumenting GHES. While perfect instruments rarely exist, I am unconvinced the instruments used previously in this literature are valid. In subsection 4.2 I show that when applied to my data and specification this skepticism is warranted. As an alternative, I use government education expenditure to instrument GHES. Government spending across social sectors is highly correlated, and education spending is a good predictor of health spending. In my data, the correlation between these two types of government expenditures is 0.93. In addition to being a good predictor of GHES, I believe government education spending (as specified in my equation) is also exogenous. This is because my estimation already controls for the affect education has on health by including maternal education as a covariate. It is well-know that investing in education leads to positive health externalities. Still, it takes many years and is through education outcomes that education spending translates into any health gains. Certainly those five years old and younger, whose mortality I am measuring, are only affected by education gains of those older than them, whose education was budgeted years in advance. Moreover, by including maternal education, I am already controlling for the true effect through which education spending would affect health.

I explore my instrument thoroughly (and the instruments used elsewhere in this literature) in subsection 4.2. In that subsection I include alternative instruments, and show that when I include education spending directly in the second stage of my analysis it does not significantly affect under-five mortality (p-value = 0.84). Furthermore, I calculate the C-statistic and Hansen-J statistics to empirically test the exogeneity of the instrument. My instrument passes both tests, with p-values of 0.72 and 0.43, respectively.

I control for unobserved country-level heterogeneity using fixed effects estimation for all reported estimates (Schaffer, 2012). In order to correct for heteroskedastic and autocorrelated residual, all standard errors reported in this paper are Newey-West adjusted using four lags. I choose four lags following the rule-of-thumb suggested by Greene (2007), which recommends using 0.25*T lags, where T = 16 years in my data. Using more lags does not affect which coefficient estimates are statistically significant.

In addition to testing the effect that government health spending has on health, I test for two types of heterogeneity. Rajkumar and Swaroop (2008) show that the spending-to-health elasticity increases (in absolute terms) as a country's government is rated less corrupt. The data used by Rajkumar and Swaroop is proprietary and unavailable to me, but I test a similar metric. Hadenius and Teorell (2005) combine data measuring civil liberties, political rights, and democracy to create a single metric, which they show to be better indicator of political democracy. I use an updated series that follows this method (Teorell et al., 2011). This variable, which I call democracy, is on a continuous ten-point scale with ten indicating a well-functioning democracy with many civil liberties and political rights. To consider a heterogeneous GHES effect, I interact GHES and the democracy index. I instrument this variable with the interaction of education expenditure and democracy, and include democracy as a covariate in my estimation.

I also consider a heterogeneous effect of GHES across income. In my data, the correlation between democracy and GDP is 0.43, so it is possible that the heterogeneity identified as a function of democracy is really a function of economic development. To test which is driving

the heterogeneity in the GHES effect, I include the interaction between GHES and GDP. To instrument this interaction I include the interaction of education expenditure and GDP, and include GDP squared as a covariate in my estimation.

Table 2 reports my primary regression results. Column 1 of Table 2 shows health-outcome elasticity to be 0.30%. These are coefficient estimates based on the fixed effect instrumental variable estimation of equation (1). The GHES coefficient estimate is statistically significant at the 99% confidence level.

Column 2 of Table 2 tests for heterogeneity across levels of democracy, and shows several interesting features. First, like in column 1, I identify a statistically significant effect of government health expenditure on under-five mortality. Second, democracies have a lower level of under-five mortality. As a country moves up the 10-point index by a single point, under-five mortality decreases by 2%. Third, the interaction between GHES and democracy is statistically significant and negative, meaning the effectiveness of GHES is greater in more democratic countries. The left-most panel of Figure 2 shows the range of estimated GHES elasticities based on the country-year democracy estimates. The thin vertical line shows the mean of this distribution. This figure shows that when I allow the GHES elasticity to vary as a function of level of democracy, the elasticity ranges from -0.24% to -0.33. (The mass of countries with an elasticity of -0.33 exists because of the preponderance of countries that score ten, the upper bound, of the democracy index.)

Column 3 of Table 2 allows the GHES elasticity to vary as a function of GDP. It shows that, like the level of democracy, economic development impacts the effectiveness of GHES, with increases in GDP meaning a larger elasticity (in absolute terms). The middle panel of Figure 2 shows that as GDP varies across my sample, I measure GHES elasticity estimates ranging from -0.02% to -0.64%. Thus, the effect of GDP has an even more dramatic effect than democracy on the effectiveness of GHES. This is a reasonable relationship because as GDP increases so does GHES, for most countries. Thus a 1% change in GHES for a country with a GDP sizably larger than another will be an increase of many more resources. An alternative way to think about this result is that economic development is good for health, but development in countries where the government is contributing significantly to financing health leads to greater gains from development.

Finally, column 4 of Table 2 includes both sets of interactions. Since both interaction terms are statistically significant, I choose this specification as my preferred estimation. Equation (2) represents this specification. In this equation, DEM is the Hadenius and Teorell measurement of political democracy (2005). The right-most panel of Figure 2 shows that when the GHES elasticity is free to vary based on the level of democracy and GDP, the GHES elasticity ranges from -0.02% to -0.61%, with a mean of -0.34%. Figure 3 illustrates the point estimate and the 95% confidence interval for the 30 most populated countries in the world using 2010 data. I see that for most of these countries there is a sizable GHES to health effect. Figure 3 also shows there is a great deal of heterogeneity in the estimated elasticities. As countries have lower GDP or level of democracy, they shift right in this figure and I estimate an elasticity closer to zero.

Still, 163 of the 186 countries have an elasticity estimate that is statistically significant at the 95% confidence level.

$$(2) \quad 5q0_{it} = \beta_{1.0} GHES_{it} + \beta_{1.1}(GHES_{it} * DEM_{it}) + \beta_{1.2}(GHES_{it} * GDP_{it}) + \beta_2 DAHG_{it} + \beta_3 DAHNG_{it} + \beta_4 GGE_{it} + \beta_{5.0} GDP_{it} + \beta_{5.1} GDP_{it}^2 + \beta_6 EDU_{it} + \beta_7 HIV_{it} + \beta_8 DEM_{it} + \alpha_i + \gamma_t + \upsilon_{it}$$

Column 4 of Table 4 also shows that the estimated effects of the covariates are primarily what is reflected elsewhere in the literature (Filmer et al., 1999; Cleland et al., 1988; Gakidou et al., 2010; Thomas et al., 1991). Increases in development assistance channeled to the government, GDP per capita, and maternal education are all associated with statistically significant reductions in under-five mortality. HIV prevalence is strongly associated with increases under-five mortality.

Two estimated coefficients in column 4 stand out as different from what is expected. First, the coefficient estimate for general government expenditure (net of GHE) is significant and positive, suggesting more GGE leads to increases in under-five mortality. One potential explanation for this is the fact that the two forms of government spending that most directly affect health, health and education, are accounted-for elsewhere in this model. Health spending is included as my variable of primary interest (and netted from GGE), while education outcomes, even more predictive than education spending, is included has a control.

The second estimated coefficient that does not meet expectations is development assistance for health channeled to the non-government sector. This coefficient has the opposite sign of what is

expected, but is not statistically significant. Similarly, the coefficient estimate for DAHG, while negative, as I expect, is relatively small in magnitude. In principle, one would hope that both types of health aid have a significant effect on lowering mortality, although aid effectiveness is debatable (Dodd et al., 2007; Doucouliagos and Paldam, 2009; Rajan and Subramanian, 2008). Furthermore, it is also possible that my failure to identify such effects is simply because I am not accounting for the reverse causation between aid and health. Indeed, some health aid is distributed to the countries where health indicators are marking increases in adverse health outcomes. This is especially true for HIV (Brooks et al., 2013). Thus, one potential reason why my estimation fails to identify sizable effects of these aid variables is because the estimates are heavily attenuated due to simultaneity.

In summary, my primary finding is that on average a 1% increase in domestic government health spending causes a 0.34% decrease in under-five mortality. Moreover, this elasticity varies dramatically, and more democratic and economically developed countries have a much larger degree of GHES effectiveness. When applied to 2010 data, the estimated elasticity ranges from -0.2% to -0.61%. My average estimate, which capitalizes on a substantially larger dataset, is very similar to the effect identified by Anand and Ravallion (1993), Gupta et al. (2002), and Bohari and Grottet (2007). Like Rajkumar and Swaroop (2008), I find that the GHES elasticity varies with policy environment, but I find the effect of economic development is substantially larger than the effect of policy. In the subsections that follow I confirm these estimates. First I scrutinize my instruments (subsection 4.2), and second, I provide a series of robustness checks (subsection 4.3).

4.2 Instrumentation

Filmer and Pritchett (1999) instrument public health spending with an indicator marking if the country is an oil-exporter, number of years since independence, and average of each countries' neighbors' public health and military spending. In an age of increasing heterogeneity it seems that neighbors' average spending might not capture the important year-over-year changes in a country's own health spending. Furthermore, the mechanism connecting oil exportation and years since independence to government expenditure (and not directly to health outcomes) isn't clear. To test these concerns, I attempt my analysis using instruments similar to those used by Filmer and Pritchett. I construct an indicator that classifies a country as an oil exporter if 10% of more of its GDP is oil rents; I calculate years-since-independence using data from the CIA; and I calculate the average military expenditure and health expenditure for each country's Global Burden of Disease region net of their own expenditures. Thus, rather than assessing contiguous neighbors I assess small geographic regions. I attempt to estimate equation (1) using these instruments. Since years since independence would be perfectly correlated with time when using a fixed effects model I estimate the equation for each year separately. I find that these instruments are quite weak when applied to my model. The weak identification F-statistic ranges from 0.7 to 3.9, depending on the year of data assessed.

Gupta et al. (2002) also use instrumentation to control for simultaneity. These authors use aid per capita and aid as a share of general government expenditure (GGE), military spending as a share of GGE, and GGE as instruments. My concern regarding these instruments is whether or not they are exogenous. Health aid that is off-budget (channeled through the non-governmental sector) is routinely directed towards health projects, and, thus may have its own effects on health

outcomes. I collect data that reflects the instruments used by Gupta et al. and use them to estimate equation (1). I see that the Gupta et al. instruments are indeed very good predictors of GHES (with weak instrument F-statistic of 168.8), but not all the instruments pass the conventional test for exogeniety. Military spending as a share of GGE and GGE have C-statistics with p-values of 0.00 and 0.08, respectively. Thus, it is unlikely that these instruments are valid for my analysis.

Returning to my own instruments, I apply four sets of tests to similarly evaluate validity. First, I consider the first stage of my instrumental variable estimation. These estimates, reported in Table 3, show that my instruments do a good job of explaining GHES, the interaction between GHES and democracy, and the interaction between GHES and GDP. Here I also report the F-statistic for the excluded instruments, ranging from 17.8 to 324.1. As all are above ten, I have reasonable confidence that my estimation does not suffer from any of the well-document problems associated with weak instruments.

Second, I use the Hansen-J statistic to test the exogeniety of my instruments. The null hypothesis of the Hansen-J test is that full instrument set is exogenous. Unfortunately, the Hansen-J test is only possible when an equation is overidentified. Since I have three variables that are instrumented, this means the test requires four instruments. To meet this criterion, I add the three-year lag of GHES as a fourth instrument. Since my reason for instrumenting GHES is to account for simultaneity, the GHES allocated three years in advance is presumably a valid instrument. Once this additional instrument is included, I complete the Hansen-J test and garner

a p-value of 0.43, suggesting that the null cannot be rejected and there is no indication that my full instrument set (including the new addition) is invalid.

Third, when my estimation is overidentified, I can afford to include an otherwise excluded instrument in my primary equation. Table 3 shows these results. The primary estimates vary only slightly, and the included instruments never have a p-value less than 0.50. This provides further evidence these instruments are exogenous.

Fourth, I calculate the C-statistic (or difference-in-Sargen statistic) for each of my three primary instruments. The C-statistic can test the exogeniety of individual instruments (whereas the Hansen-J tested the full instrument set). In order to apply this test, my equation must be more than over-identified, and have a total of five instruments. In addition to the three-year lag of GHES, I temporarily include the four-year lag of GHES as well. The p-values for the C-statistics for education expenditure, the interaction between education expenditure and democracy, and the interaction between education expenditure and GDP are 0.70, 0.63, and 0.15, respectively. Thus, I am unable to reject the hypothesis that each of the instruments is not exogenous.

4.3 Robustness

With confidence that my instruments are valid, I move on to assessing the sensitivity of my estimation. Column 1 of Table 4 shows the estimation of my preferred equation using un-imputed dated. Column 2 completes my analysis with all financial variables measured in 2010 US dollars rather than 2010 PPP dollars. Column 3 collapses my data such that all variables are

measuring the three-year average rather than annual data. Finally, column 4 uses simple fixed effect estimation without any instrumentation.

The estimates displayed in columns (1) through (3) in Table 4 confirm my primary results and show substantial stability. Even with only slightly more than half the number of observations, column (1) shows effects very similar to what was measured in column (4) of Table 2. Similarly, columns (2) and (3) show similar effects. Column (4) of Table 4 (which includes no instrumentation) confirms my expectation that there is substantial reverse causation. Not accounting for the simultaneity between GHES and health leads to noteworthy attenuation of the estimated elasticities.

Table 5 adds additional covariates common in this literature to my preferred specificaiton. I show the inclusion of these additional covariates does not change my primary results. Columns (1) through (6) include the proportion of households with access to an improved water source, the proportion of households with access to improved sanitation facilities, the number of deaths per 10,000 caused by conflict, the number of deaths per 10,000 caused by natural disaster, debt relief per capita, the share of the population living a city with population density greater than 1,000 per square kilometer and the total fertility rate. Table 5 shows that sanitation, war and natural disaster have statistically significant effects on under-five mortality.

Table 6 replaces under-five mortality with five alternative health outcomes. In columns (1) through (5) I consider neonatal, post-neonatal, adult female and adult male mortality, and the malnutrition rate. Each of these outcomes is natural log transformed so the coefficient estimates

are comparable to the elasticities estimated elsewhere. I see the primary results presented in Table 2 are reflected in columns (1) and (2) of Table 6. Neonatal and post-neonatal mortality responds to GHES in a very similar manner as under-five mortality. For malnutrition, (column (3)) I see that GHES has a significant effect, but there is less heterogeneity across democracy and economic develpoment. For adult mortality my estimates suggest that GHES is more effective in less democratic countries, especially for men (column (5)). One possible explanation might be that different forms of government on average favor different sorts of health projects, with democracies investing more in children and newborns, and countries with fewer rights and less democracy favoring adult males.

5. Conclusions

Two decades of contradicting results should not be interpreted as evidence that government health expenditure does not affect health. The previous research attempting to causally connect government spending and health outcomes have been hampered by ill-powered samples, simultaneity and underperforming instruments. I address these shortcomings by using a dataset six times larger than any used previously and introducing a new instrument. My estimates suggest government health expenditure does significantly reduce under-five mortality, with my primary result showing that an increase of GHES by 1% leads to a reduction of under-five mortality by 0.34%. Moreover, context matters. In democratic countries with larger GDP per capita the GHES elasticity extends upward (in absolute terms), while countries with fewer rights, less democracy, and less economic development has smaller elasticities. My preferred

specification suggests that in 2010 the GHES elasticity for the 186 countries in my sample range from -0.02% to -0.61%.

One established conclusion in the literature is that government health expenditure has heterogeneous effects within a population, with the gains in health achieved by health spending greatest for the poor (Bidani and Ravallion, 1997; Wagstaff, 2003; Gupta et al., 2003). I anticipate that this conclusion holds in my data, although addressing it is outside the bounds of this analysis. I believe the principal contribution of my work is to clarify a much debated issue and provide important insights on heterogeneity. To state that on average GHES does not lead to positive health gains is incorrect, although because of the wide-ranging heterogeneity, this may be true for a small subset of countries. These countries represent the minority my 186 country sample. On average GHES is capable of achieving significant gains, especially in middle- and upper-income countries with many civil liberties, political rights and a democratic political environment or in countries with high under-five mortality. These results offer the encouraging recommendation that economic development, democratization, and government health spending can work in concert to significantly reduce the burden of premature mortality.

Acknowledgements

In addition to those thanked in the prefix, I would like to thank Annette Tardiff, Mark Anderson, and the participants at an ASHEcon conference seminar (Minneapolis, July 2012) for their valuable feedback regarding this chapter. Any remaining errors are of course my own.

Abbreviations

DAHG	Development Assistance for Health channeled to Governments
DAHNG	Development Assistance for Health channeled to Non-Governments
EDU	Maternal Education
EM	Expectation Maximization
GDP	Gross Domestic Product
GGE	General Government Expenditure (net of GHE)
GHE	total Government Health Expenditure
GHES	Government Health Expenditure as Source
HIV	Human Immunodeficiency Virus
IHME	Institute for Health Metrics and Evaluations
LCU	Local Currency Units
MAR	Missing At Random
MI	Multiple Imputation
WHO	World Health Organization

References

Adetunji J. Trends in under-5 mortality rates and the HIV/AIDS epidemic. BULLETIN-WORLD HEALTH ORGANIZATION 2000; 78; 1200–1206.

Anand S, Ravallion M. Human development in poor countries: on the role of private incomes and public services. The Journal of Economic Perspectives 1993; 7; 133–150.

Baldacci E, Clements B, Gupta S, Cui Q. Social Spending, Human Capital, and Growth in Developing Countries. World Development 2008; 36; 1317–1341.

Baldacci E, Guin-Siu MT, Mello LD. More on the effectiveness of public spending on health care and education: a covariance structure model. Journal of International Development 2003; 15; 709–725.

Bidani B, Ravallion M. Decomposing social indicators using distributional data. Journal of Econometrics 1997; 77; 125–139.

Bokhari FAS, Gai Y, Gottret P. Government health expenditures and health outcomes. Health Economics 2007; 16; 257–273.

Cleland JG, van Ginneken JK. Maternal education and child survival in developing countries: The search for pathways of influence. Social Science & Medicine 1988; 27; 1357–1368.

Dodd R, Schieber G, Cassels A, Fleisher L, Gottret P. Aid effectiveness and health 2007.

Doucouliagos H, Paldam M. The Aid Effectiveness Literature: The Sad Results of 40 Years of Research. Journal of Economic Surveys 2009; 23; 433–461.

Farag M, Nandakumar AK, Wallack SS, Gaumer G, Hodgkin D. Does funding from donors displace government spending for health in developing countries? Health Affairs 2009; 28; 1045.

Fernandes Antunes A, Xu K, James CD, Saksena P, Van de Maele N, Carrin G, Evans DB. General Budget Support: Has It Benefited the Health Sector? Health Economics 2012; n/a–n/a.

Filmer D, Hammer JS, Pritchett LH. Weak Links in the Chain: A Diagnosis of Health Policy in Poor Countries. The World Bank Research Observer 2000; 15; 199 –224.

Filmer D, Pritchett L. The impact of public spending on health: does money matter? Social Science & Medicine 1999; 49; 1309–1323.

Gakidou E, Cowling K, Lozano R, Murray CJ. Increased educational attainment and its effect on child mortality in 175 countries between 1970 and 2009: a systematic analysis. The Lancet 2010; 376; 959–974.

Greene WH. Econometric Analysis. 6th ed. Prentice Hall; August 17, 2007.

Gupta S, Verhoeven M, Tiongson ER. Public spending on health care and the poor. Health Economics 2003; 12; 685–696.

Hadenius A, Teorell J. Assessing alternative indices of democracy. Concepts & Methods Workinng Paper 6, IPSA 2005. http://www.concepts-methods.org/Files/WorkingPaper/PC%206%20Hadenius%20Teorell.pdf.

Hanson K, Gilson L, Goodman C, Mills A, Smith R, Feachem R, Feachem NS, Koehlmoos TP, Kinlaw H. Is Private Health Care the Answer to the Health Problems of the World's Poor? PLoS Med 2008; 5; e233.

Honaker J, King G. What to Do about Missing Values in Time-Series Cross-Section Data. American Journal of Political Science 2010; 54; 561–581.

Honaker J, King G, Blackwell M. Amelia II: A program for missing data. Journal of Statistical Software 2011; 45; 1–47.

Institute for Health Metrics and Evaluation. 2013. Financing global health 2012: End of the golden age? Seattle, United States; January 2013.

International Monetary Fund. 2012. World Economic Outlook, October 2012: Coping with High Debt and Sluggish Growth. Washington D.C.; 2012.

Jamison Dean T., Breman Joel G., Measham Anthony R., Alleyne George, Claeson Mariam, Evans David B., Jha Prabhat, Mills Anne, and Musgrove Philip, eds. Disease Control Priorities in Developing Countries. 2nd ed. World Bank Publications; April 2, 2006a.

———, eds. Priorities in Health: Disease Control Priorities Companion Volume. 1st ed. World Bank Publications; April 2, 2006b.

Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, Abraham J, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. The Lancet 2012; 380; 2095–2128.

Lu C, Schneider MT, Gubbins P, Leach-Kemon K, Jamison D, Murray CJ. Public financing of health in developing countries: a cross-national systematic analysis. The Lancet 2010; 375; 1375–1387.

McGuire J. Basic health care provision and under-5 mortality: A Cross-National study of developing Countries. World Development 2006; 34; 405–425.

Musgrove, Phil. Public and Private Roles in Health 1996. Techincal Report 339, World Bank.

Or Z. Determinants of health outcomes in industrialised countries: a pooled, cross-country, time-series analysis. OECD Economic Studies 2000; 30; 53–78.

Piva P, Dodd R. Where did all the aid go? An in-depth analysis of increased health aid flows over the past 10 years. Bulletin of the World Health Organization 2009; 87; 930–939.

Pritchett L, Summers LH. Wealthier is healthier. Journal of Human Resources 1996; 31; 841–868.

Rajan RG, Subramanian A. Aid and Growth: What Does the Cross-Country Evidence Really Show? Review of Economics and Statistics 2008; 90; 643–665.

Rajaratnam JK, Marcus JR, Levin-Rector A, Chalupka AN, Wang H, Dwyer L, Costa M, Lopez AD, Murray CJ. Worldwide mortality in men and women aged 15–59 years from 1970 to 2010: a systematic analysis. The Lancet 2010; 375; 1704–1720.

Rajkumar AS, Swaroop V. Public spending and outcomes: Does governance matter? Journal of Development Economics 2008; 86; 96–111.

Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley-Interscience; June 9, 2004.

Schaffer ME. xtivreg2: Stata module to perform extended IV/2SLS, GMM and AC/HAC, LIML and k-class regression for panel data models. July 24, 2012.

Sijpe NV de. Is Foreign Aid Fungible? Evidence from the Education and Health Sectors. The World Bank Economic Review 2012. http://wber.oxfordjournals.org/content/early/2012/10/30/wber.lhs023.

Teorell J, Samanni M, Holmberg S, Rothstein B. 2011. The QoG Standard Dataset version 6Apr11. University of Gothenburg: The Quality of Government Institute; 2011. Available at: http://www.qog.pol.gu.se.

The World Bank. World Development Indicators.

Thomas D, Strauss J, Henriques M-H. How Does Mother's Education Affect Child Height? The Journal of Human Resources 1991; 26; 183–211.

Wagstaff A. Child health on a dollar a day: some tentative cross-country comparisons. Social Science & Medicine 2003; 57; 1529–1538.

Wang H, Dwyer-Lindgren L, Lofgren KT, Rajaratnam JK, Marcus JR, Levin-Rector A, Levitz CE, Lopez AD, Murray CJ. Age-specific and sex-specific mortality in 187 countries, 1970?2010: a systematic analysis for the Global Burden of Disease Study 2010. The Lancet 2012; 380; 2071–2094.

World Health Organization. Guide to Producing National Health Accounts 2003.

———. World Health Statistics 2012. World Health Organization: Geneva; 2012.

Tables and figures

Table 1: Previous cross-country analyses of government health expenditure

Authors	Year	Journal	Sample	Estimation Method	Dependent Variable	Variable of Interest	Estimate
Anand & Ravallion	1993	Journal of Economic Perspectives	22 low-/middle-income countries	OLS	log(80-life expectancy)	log public health spending per person	-0.30***
Bidani & Ravallion	1997	Journal of Econometrics	35 developing countries	random coefficient model	log life expectancy	log public health spending per person	0.13*** for poor, but no effect for non-poor
Filmer & Pritchett	1999	Social Science & Medicine	98 countries	IV(geographic neighbors' public health spending and defense spending, years since independence, and if country's main export is oil)	log 5q0	log share of public spending on health as a fraction of GDP	-0.13
Gupta, Verhoeven, & Tiongson	2002	European Journal of Political Economy	22 countries	IV(aid per capita, aid as percent of general government expenditure, military spending as percent of general government expenditure, total government expenditure)	log 5q0	log public health spending as share of GDP	-0.29**

Or	2000	OECD Economic Studies	21 OECD countries, 1970-1992	country-level fixed effects	all cause years of life lost	share of public expenditure in total health expenditure	-0.17*** for women, -0.18*** for men
Gupta, Verhoeven, & Tiongson	2003	Health Economics	67 countries	OLS	5q0	log public health spending per capita	-26.7*** for poor, -9.2** for nonpoor
Baldacci, Guin-siu, & de Mello	2003	Journal of International Development	94 developing and transitioning countries	latent variable model	5q0 and 1q0	public expenditure on health care as percent of GDP	-0.05
Wagstaff	2003	Social Science & Medicine	32 countries	OLS	log 1q0	log per capita public spending on health	-0.25* for poor
McGuire	2006	World Development	46 countries	OLS	log 5q0	log public health spending as percent of GDP	0.07
Bokhari & Gottret	2007	Health Economics	127 countries	GMM	log 5q0	log government health expenditure per capita	-0.33*
Rajkumar & Swaroop	2008	Journal of Development Economics	91 countries; 1990, 1997 & 2003	OLS	log 5q0	public health spending as share of GDP	-0.18** average; but varies depending on governance

Table 2: Primary regression results

	(1)	(2)	(3)	(4)
VARIABLES	Dependent variable: log under-five mortality			
GHES	-0.305***	-0.296***	-0.359***	-0.340***
	(0.099)	(0.100)	(0.095)	(0.097)
GHES*Democracy		-0.009***		-0.006*
		(0.003)		(0.003)
GHE*GDP			-0.100**	-0.087**
			(0.041)	(0.043)
DAHG	-0.034***	-0.033***	-0.029**	-0.026**
	(0.013)	(0.013)	(0.012)	(0.012)
DAHNG	0.011	0.009	0.007	0.008
	(0.009)	(0.009)	(0.010)	(0.010)
GGE (net of GHE)	0.114**	0.101**	0.104**	0.096**
	(0.045)	(0.045)	(0.041)	(0.042)
GDP	-0.196***	-0.205***	-2.150***	-2.007**
	(0.075)	(0.073)	(0.810)	(0.807)
GDP squared			0.119**	0.110**
			(0.050)	(0.050)
Maternal eduation	-0.153**	-0.196***	-0.237***	-0.240***
	(0.066)	(0.064)	(0.074)	(0.071)
HIV pervalence	0.031***	0.031***	0.032***	0.031***
	(0.010)	(0.010)	(0.010)	(0.010)
Democracy		-0.017**		-0.016**
		(0.007)		(0.007)
Observations	2,976	2,976	2,976	2,976
Number of countries	186	186	186	186

Newey West HAC standard errors based on 4 lags in parentheses

*** p<0.01, ** p<0.05, * p<0.1

All financial variables are measured in 2010 PPP dollars per capita and natrual log transformed.

All interaction are between demeaned variables (centered at zero).

Coefficient estimates for year indicators surpressed.

Fixed effects instrumental variable estimation for all columns.

Instruments: (1) Education expenditure; (2) education expenditure and education expenditure interacted with democarcy; (3) education expenditure and education expenditure interacted with GDP; (4) education expenditure, education expenditure interacted with democracy, and education expenditure interacted with GDP.

Table 3: First-stage estimates

VARIABLES	(1) GHES	(2) GHES*Dem	(3) GHES*GDP
DAHG	-0.090***	0.114**	0.083***
	(0.014)	(0.050)	(0.020)
DAHNG	0.047***	-0.061	-0.073***
	(0.015)	(0.051)	(0.023)
GGE	0.315***	-0.795***	-0.216*
	(0.051)	(0.189)	(0.117)
GDP	-0.400	-8.078***	-12.932***
	(0.532)	(2.668)	(1.317)
GDP squared	0.048	0.488***	0.780***
	(0.030)	(0.163)	(0.075)
Maternal education	-0.148	-1.326**	-0.561**
	(0.129)	(0.533)	(0.223)
HIV prevalence	0.000	-0.005	0.014
	(0.007)	(0.031)	(0.011)
Democracy	0.021**	-0.234***	-0.013
	(0.008)	(0.069)	(0.012)
Education expenditure	0.175***	0.426**	0.019
	(0.031)	(0.181)	(0.056)
Edu exp*Democracy	-0.001	0.981***	0.024**
	(0.005)	(0.034)	(0.011)
Edu exp*GDP	-0.007	-0.480***	0.283***
	(0.027)	(0.148)	(0.063)
F-statistic of excluded instruments	17.8	324.0	16.4
Observations	2,976	2,976	2,976

Newey West HAC standard errors based on 4 lags in parentheses

*** p<0.01, ** p<0.05, * p<0.1

All financial variables are measured in 2010 PPP dollars per capita and natrual log transformed.

All interaction are between demeaned variables (centered at zero).

Coefficient estimates for year indicators surpressed.

Fixed effects instrumental variable estimation for all columns.

Table 4: Hansen J test and including otherwise excluded instruments

VARIABLES	(1)	(2)	(3)	(4)	(5)
	Dependent variable: log under-five mortality				
GHES	-0.274***	-0.242***	-0.283***	-0.321***	-0.307***
	(0.062)	(0.072)	(0.066)	(0.106)	(0.078)
GHES*Democracy	-0.004	-0.004	-0.020	-0.001	-0.004
	(0.004)	(0.004)	(0.023)	(0.006)	(0.004)
GHE*GDP	-0.073*	-0.067*	-0.095*	-0.155	-0.083*
	(0.041)	(0.041)	(0.056)	(0.140)	(0.050)
DAHG	-0.023***	-0.021**	-0.020**	-0.021**	-0.023**
	(0.008)	(0.009)	(0.010)	(0.009)	(0.010)
DAHNG	0.001	0.000	0.000	-0.001	0.000
	(0.008)	(0.008)	(0.008)	(0.009)	(0.008)
GGE (net of GHE)	0.079***	0.073***	0.067**	0.083***	0.100***
	(0.028)	(0.028)	(0.031)	(0.031)	(0.032)
GDP	-1.644**	-1.540**	-2.047**	-2.719	-1.813*
	(0.785)	(0.784)	(1.025)	(1.943)	(0.925)
GDP squared	0.084*	0.078	0.109*	0.150	0.095*
	(0.048)	(0.048)	(0.063)	(0.119)	(0.057)
Maternal education	-0.317***	-0.303***	-0.360***	-0.375***	-0.341***
	(0.085)	(0.087)	(0.109)	(0.129)	(0.099)
HIV pervalence	0.017	0.017	0.017	0.019	0.010
	(0.016)	(0.016)	(0.016)	(0.017)	(0.017)
Democracy	-0.010*	-0.010*	-0.014*	-0.010*	-0.008
	(0.006)	(0.005)	(0.008)	(0.006)	(0.006)
Education expenditure		-0.013			
		(0.017)			
Edu exp*Democracy			0.016		
			(0.024)		
Edu exp*GDP				0.023	
				(0.036)	
Hansen-J p-value	0.432				0.283
Additional instrument	3-yr lag GHES	3-yr lag GHES	3-yr lag GHES	3-yr lag GHES	3- & 4-yr lag GHES
Observations	2,418	2,418	2,418	2,418	2,232

Newey West HAC standard errors based on 4 lags in parentheses

*** p<0.01, ** p<0.05, * p<0.1

All financial variables are measured in 2010 PPP dollars per capita and natrual log transformed.

All interaction are between demeaned variables (centered at zero).

Coefficient estimates for year indicators surpressed.

Fixed effects instrumental variable estimation for all columns.

Standard instruments: education expenditure, education expenditure interacted with democracy, and education expenditure interacted with GDP.

Table 5: Robustness checks

VARIABLES	(1)	(2)	(3)	(4)
	Dependent variable: log under-five mortality			
GHES	-0.326***	-0.359***	-0.343**	-0.043**
	(0.099)	(0.107)	(0.136)	(0.020)
GHES*Democracy	-0.002	-0.007**	-0.009	-0.008***
	(0.002)	(0.003)	(0.006)	(0.003)
GHE*GDP	-0.123*	-0.057	-0.083	-0.007
	(0.072)	(0.040)	(0.062)	(0.011)
DAHG	-0.014	-0.020	-0.027	-0.006
	(0.018)	(0.013)	(0.017)	(0.007)
DAHNG	0.005	0.026*	0.011	-0.001
	(0.012)	(0.015)	(0.018)	(0.008)
GGE (net of GHE)	0.078**	0.109**	0.103	0.011
	(0.037)	(0.046)	(0.066)	(0.020)
GDP	-3.043**	-1.538**	-2.088*	-0.564**
	(1.475)	(0.672)	(1.194)	(0.264)
GDP squared	0.170**	0.090**	0.112	0.015
	(0.087)	(0.044)	(0.074)	(0.017)
Maternal education	-0.317***	-0.268***	-0.301***	-0.137**
	(0.103)	(0.083)	(0.104)	(0.064)
HIV pervalence	0.044***	0.031***	0.028	0.030***
	(0.005)	(0.009)	(0.017)	(0.010)
Democracy	-0.004	-0.018**	-0.022*	-0.023***
	(0.005)	(0.008)	(0.012)	(0.006)
Check:	Raw (unimputed) data	2010 US dollars	Collapse into 3-year averages	Fixed-effects estimation (no instruments)
Observations	1,519	2,976	930	2,976

Newey West HAC standard errors based on 4 lags in parentheses

*** p<0.01, ** p<0.05, * p<0.1

All financial variables are measured in 2010 PPP dollars per capita and natrual log transformed.

All interaction are between demeaned variables (centered at zero).

Coefficient estimates for year indicators surpressed.

Fixed effects instrumental variable estimation for columnns (1) through (3).

Instruments: education expenditure, education expenditure interacted with democracy, and education expenditure interacted with GDP.

Table 6: Additional covariates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	Dependent variable: log under-five mortality						
GHES	-0.319***	-0.310***	-0.327***	-0.312***	-0.319***	-0.317***	-0.318***
	(0.096)	(0.095)	(0.097)	(0.096)	(0.097)	(0.096)	(0.096)
GHES*Democracy	-0.007*	-0.006*	-0.007*	-0.006*	-0.006*	-0.006*	-0.006*
	(0.003)	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
GHE*GDP	-0.069*	-0.078*	-0.067	-0.072*	-0.073*	-0.074*	-0.073*
	(0.040)	(0.042)	(0.042)	(0.042)	(0.042)	(0.042)	(0.042)
Water	0.095						
	(0.212)						
Sanitation		-0.207**					
		(0.092)					
War			0.041***				
			(0.010)				
Natural disaster				20.761***			
				(4.245)			
Debt relief					0.003		
					(0.010)		
Urban						0.079	
						(0.088)	
Fertility Rate							0.024
							(0.026)
Observations	2,790	2,790	2,790	2,790	2,790	2,790	2,790

Newey West HAC standard errors based on 4 lags in parentheses

*** p<0.01, ** p<0.05, * p<0.1

All financial variables measured in 2010 PPP dollars per capita and natrual log transformed.

All interaction are between demeaned variables (centered at zero).

Coefficient estimates for year indicators and regular covariates included in equation (2) surpressed.

Fixed effects instrumental variable estimation for all columns.

Instruments: education expenditure, education expenditure interacted with democracy, and education expenditure interacted with GDP.

Table 7: Alternative health indicators

VARIABLES	(1) Neonatal	(2) Post-neonatal	(3) Malnutrition	(4) Adult female	(5) Adult male
GHES	-0.304***	-0.406***	-0.161***	-0.208**	-0.187**
	(0.097)	(0.120)	(0.052)	(0.093)	(0.086)
GHES*Democracy	-0.008**	-0.008**	-0.001	0.001	0.004*
	(0.003)	(0.004)	(0.002)	(0.003)	(0.002)
GHE*GDP	-0.083**	-0.069	-0.032	0.003	-0.021
	(0.037)	(0.044)	(0.021)	(0.035)	(0.031)
DAHG	-0.017	-0.017	-0.006	-0.009	-0.011
	(0.011)	(0.015)	(0.006)	(0.010)	(0.009)
DAHNG	0.035***	0.027	0.019***	0.047***	0.039***
	(0.012)	(0.017)	(0.007)	(0.015)	(0.014)
GGE (net of GHE)	0.087**	0.100*	0.030	0.102**	0.099**
	(0.037)	(0.052)	(0.022)	(0.043)	(0.039)
GDP	-1.507**	-1.912***	-0.493	-0.012	-0.195
	(0.634)	(0.732)	(0.351)	(0.552)	(0.472)
GDP squared	0.087**	0.115**	0.016	-0.004	0.010
	(0.042)	(0.047)	(0.023)	(0.032)	(0.028)
Maternal education	-0.099	-0.265***	-0.113*	0.030	0.065
	(0.078)	(0.093)	(0.060)	(0.064)	(0.059)
HIV pervalence	0.023***	0.033***	-0.007***	0.054***	0.038***
	(0.006)	(0.009)	(0.002)	(0.009)	(0.006)
Democracy	-0.020***	-0.019**	0.002	0.005	0.006
	(0.008)	(0.008)	(0.004)	(0.006)	(0.006)
Observations	2,976	2,976	2,976	2,976	2,976

Newey West HAC standard errors based on 4 lags in parentheses

*** p<0.01, ** p<0.05, * p<0.1

All financial variables are measured in 2010 PPP dollars per capita and natrual log transformed.

All interaction are between demeaned variables (centered at zero).

Coefficient estimates for year indicators surpressed.

Fixed effects instrumental variable estimation for all columns.

Instruments: education expenditure, education expenditure interacted with democracy, and education expenditure interacted with GDP.

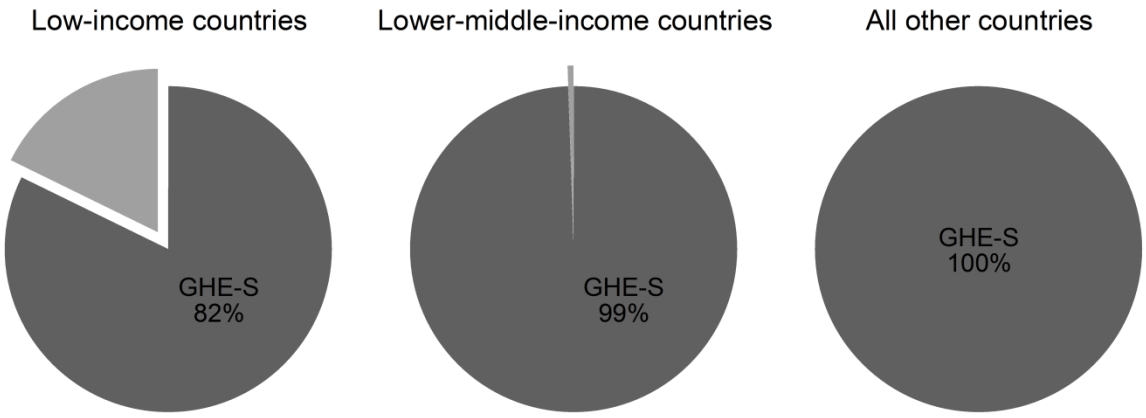
Table A1: Imputation covariates

Variables	Units	Source
Country specific linear time trends		
Government health expenditure as source* (plus 1- and 2-year lags and leads)	2010 PPP per capita	Calculated by authors from WHO data
General government expenditure*	2010 PPP per capita, net of total government health expenditure	Calculated by authors from WHO data
Development assistance for health channeled to government*	2010 PPP per capita	Institute for Health Metrics and Evaluation
Development assistance for health channeled to non-governmental organizations*	2010 PPP per capita	Institute for Health Metrics and Evaluation
Percentage of population living at over 1000 people per square km*	Proportion of population	Masters, et al.
Maternal education*	Average years of education for women older than 24 years	Gakidou, et al.
Adult male education*	Average years of education for men older than 24 years	Gakidou, et al.
War deaths*	Deaths per 1000	Institute for Health Metrics and Evaluation
Natural disaster deaths*	Deaths per 1000	Institute for Health Metrics and Evaluation
Population*	Count	Wong, et al.
Malnutrition*	Proportion of population	Institute for Health Metrics and Evaluation
Under-five mortality rate*	Deaths prior to 5 years per 1000 live births	Wong, et al.
Female adult mortality rate*	Deaths prior to 60 years per 1000 females reaching 15	Wong, et al.
Percentage of population with access to an improved water source*	Proportion of population	Institute for Health Metrics and Evaluation
Percentage of population with access to an improved sanitation facility*	Proportion of population	Institute for Health Metrics and Evaluation
Exchange rate	Official PPP exchange rate	World Health Organization

Total fertility rate*	Average number of children born to a women over her lifetime	Institute for Health Metrics and Evaluation
Government military expenditure*	2010 PPP per capita	World Bank
Government military expenditure*	As share of general government expenditure	World Bank
Government education expenditure*	2010 PPP per capita	World Bank
Government education expenditure *	As share of general government expenditure	World Bank
Male adult mortality rate*	Deaths prior to 60 years per 1000 males reaching 15	Wong, et al.
Debt relief*	2010 PPP per capita	World Bank
HIV/AIDs *	Prevalence	UN AIDS
Gross domestic product*	2010 PPP per capita	James, et al.
Gross domestic product growth rate	Percent change	Calculated by authors from James, et al.

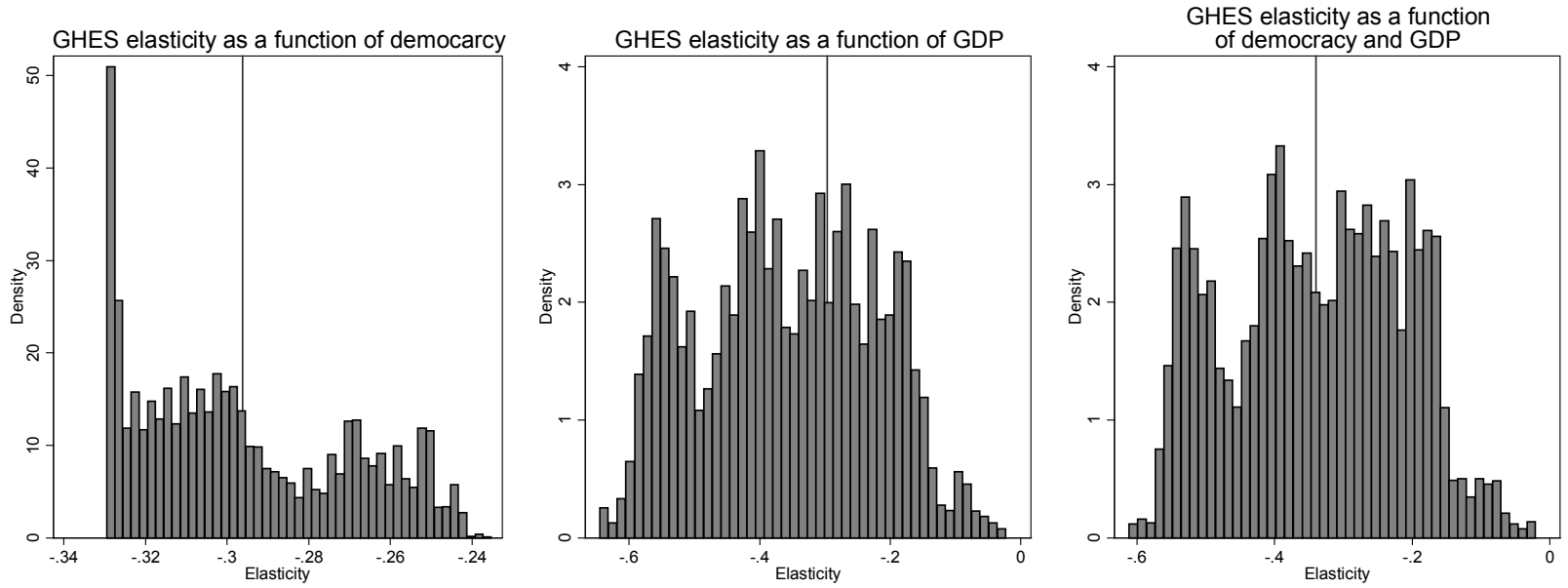
* indicates that the variables enters the imputation log-transformed.

Figure 1: Source of government health expenditure by income group



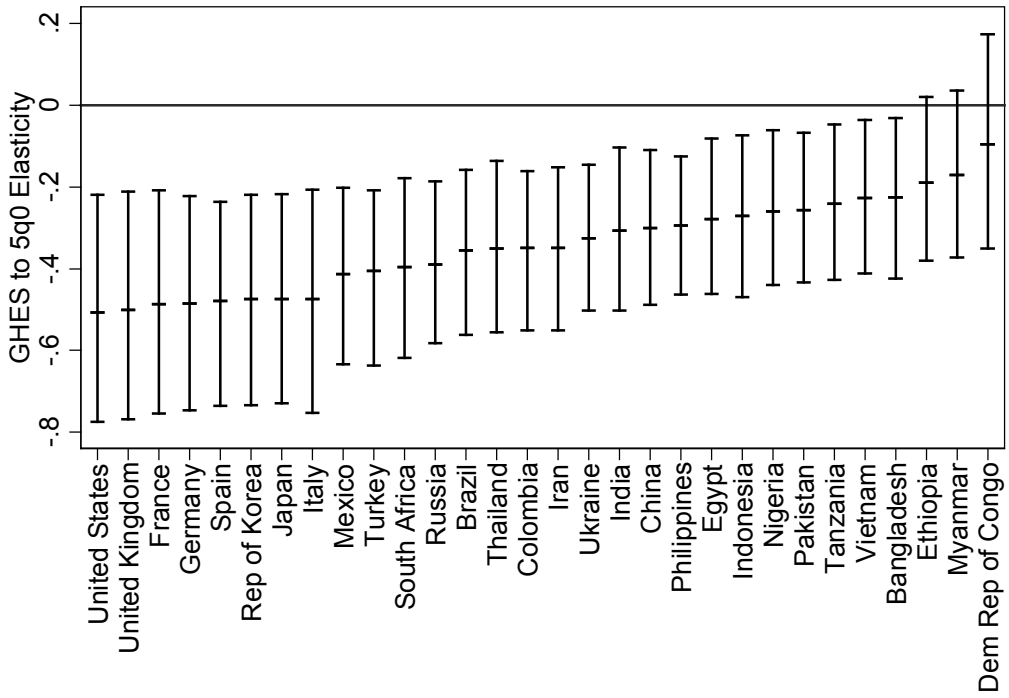
Source of government health expenditure (as agent). Light grey is development assistance for health channeled to government. Dark grey is government health expenditure as source.

Figure 2: Distribution of GHES elasticities for 186 countries between 1995 and 2010



Each panel shows the distribution of GHES elasticities allowing the elasticity to vary across a different dimension. The left-most panel allows the elasticity to vary as a function of how democratic a country is. The mean elasticity is -0.30%. The middle panel allows the elasticity to vary as a function of GDP. The mean elasticity in this case is -0.36%. The right-most panel allows the GHES elasticity to vary across both the level of democracy and GDP, and has a mean of -0.34%.

Figure 3: 2010 point estimates and 95% confidence intervals of the GHES elasticities for the 30 most populated countries in the world



The center dash for each country marks the estimated GHES elasticity specific to that country's level of democracy and GDP in 2010. The length of the line represents the 95% confidence interval of the estimated GHES elasticity. The countries included are the 30 most populous as of 2010. For each country, I take 1,000 random draws from the 20 imputed datasets of the level of democracy and GDP and multiply them by 1,000 draws from the estimated variance-covariance matrix.

Chapter 2: Measuring the displacement and replacement of government health expenditure

Running Header: The Displacement and replacement of GHES

Version: May 2, 2013

Author: Joseph L. Dieleman, M.A.
Department of Economics, University of Washington
Institute for Health Metrics and Evaluation, University of Washington

Correspondence: Joseph L. Dieleman
dieleman@uw.edu
Institute for Health Metrics and Evaluation
2301 5th Avenue, #600
Seattle, WA 98121 USA
+ 1 (206) 897-2800 (telephone)
+ 1 (206) 897-2899 (fax)

Abstract

This study measures the effect increases and decreases of health aid have on government health expenditure. Government health expenditure is a primary means of financing health services in many countries. Research assessing the relationship between government health expenditure and development assistance for health channeled to governments (DAHG) has produced contradictory conclusions. Moreover, no research has considered that this relationship may depend on whether DAHG is increasing or decreasing over time. I construct a panel of financial flows data spanning 132 countries and 16 years. General method of moments estimation provides credible evidence that DAHG causes the displacement of government health expenditure as source (GHES). In the short-run, $1 of DAHG displaces $0.65 (95% confidence interval, or CI: 0.12, 1.18) of GHES, leaving only $0.35 (CI: -0.18, 0.88) of DAHG additional. Furthermore, I show that GHES is dynamic, meaning that the current level of GHES is a function of past levels. This also means that short-run shocks can have persistent effects on GHES. To explain why, I disaggregate the average effect of DAHG by separately identifying the effects of increases versus decreases in DAHG. I find that $1 year-over-year increase in DAHG leads to a $0.63 (CI: 0.09, 1.15) decrease in GHES, while a $1 year-over-year decrease in DAHG does not have an effect on GHES that is statistically different from zero (CI: -0.40, 0.96). Simulation shows that the displacement of GHES between 1995 and 2010 reduced total government health expenditure by $46.7 billion of resources. Similarly, the irregular disbursement of DAHG reduced total government expenditure by $21.0 billion. Thus, this research highlights the cost of erratic aid disbursement and sizable substitution. Short-run fluctuations in DAHG have long-run implications for the amount of government resources available for health.

Keywords: government health expenditure, development assistance, displacement, replacement additionality, crowding out, aid fungibility

1. Introduction

Governments finance a sizable portion of population health initiatives throughout the world. Government resources fund public health systems, health services, and public insurance schemes. In 2010, governments spent roughly $4.05 trillion on health, although these resources are not distributed uniformly across the globe (World Health Organization, 2012; Institute for Health Metrics and Evaluation, 2013). 85% of this total is spent in high-income countries, where approximately 15% of the world's population lives (United Nations Population Division, 2011). To address health financing gaps in many low- and middle-income, domestically-generated government health expenditure is complemented by development assistance for health (DAH). In 2010, DAH totaled $28.2 billion, with 47% of the DAH that can be tracked to specific countries channeled to governments (denoted as DAHG) (Institute for Health Metrics and Evaluation, 2013).

For most donors, the implicit (and sometimes explicit) intention of DAHG is to increase the total amount of governmental spending on health. In other words, donors intend for governments to maintain their own health spending while they receive DAHG. However, governments have competing priorities and may use DAHG as a substitute for their own resources. So when a government receives DAHG, it might reduce its level of government health expenditure as source (GHES, or the domestically-generated portion of total government health expenditure). It then redirects those resources to a non-health sector. The displacement of GHES means that DAHG is not fully additional, and a dollar of DAHG does not lead to an increase of one dollar in total health expenditure. Some research shows this to be the case (Lu et al., 2010; Farag et al.,

2009; Fernandes Antunes et al., 2012). While these findings coincide with economic theory and empirical studies that assess other government sectors (Pack and Pack, 1993, 1990; Dodd et al., 2007; Feyzioglu et al., 1998; Swaroop et al., 1999; Oates, 1999; Wagstaff, 2011; Swaroop et al., 2000), there is other research which questions these conclusions (Roodman, 2012; Batniji and Bendavid, 2012; Garg et al., 2012; Van de Sijpe, 2012). In order to assess and predict resources for health and their potential health gains, knowing how financial flows interact is of the utmost important.

Furthermore, an assessment of the relationship between DAHG and GHES should provide more than just a single aggregated estimate of the effect. The rate at which governments displace GHES with DAHG might differ across many dimensions, including the type of DAHG or the organization of the country's health system. A third dimension that might dictate a government's rate of displacement is how the present amount of DAHG relates to previous amounts. Over (2009) shows donors allocate aid across sectors based on whether or not the donor's economy is growing. It is rational that aid recipients might also respond asymmetrically to increases and decreases in resources, especially if reallocating funds across sectors is costly. The prevailing assumption is that governments replace GHES upon decreases in DAHG at the exact same rate that they displace GHES upon increases in DAHG. If displacement does not equal replacement, then the adverse effects of a short-run DAHG shock could persist beyond the length of the shock itself.

To assess these issues, I construct a panel of cross-country GHES and DAHG data that is larger and more current than any other used for measuring the additionality of DAHG. I use general

method of moments (GMM) estimation to identify to what extent the effect of DAHG on GHES is causal and dynamic. Finding that a significant and long-run effect exists, I estimate a second model that is flexible enough to measure the effect of increases and decreases in DAHG separately. I test if these two effects are symmetric (equal), as unequal rates of displacement and replace can in part explain the persistence of the adverse DAHG effect.

2. Methods

2.1 Data

The data used in this study span 16 years (1995 through 2010) and 132 countries. DAHG is reported by the Institute for Health Metrics and Evaluation in Financing Global Health 2012: End of the Golden Age? (Institute for Health Metrics and Evaluation, 2013). The World Health Organization (WHO) reports total government health expenditure data, also known as government health expenditure as agent (GHE) (World Health Organization, 2012). By definition, GHE ≡ GHES + DAHG (Lu et al., 2010; World Health Organization, 2003). Thus, GHES is obtained by subtracting DAHG from GHE. All financial variables except GDP are measured as a share of the country's 16-year average GDP. This strategy enables us to correctly identify periods when GHES is increasing or decreasing in absolute terms. (If the variables were measured as a share of contemporaneous GDP, then periods when DAHG increased but at a slower rate than GDP would incorrectly appear to have declining levels of DAHG). Tables S1 through S4 in the supplementary appendix include a list of countries included in the study, a list of data sources and summary statistics (James et al., 2012; UNAIDS and WHO, 2010; EM-DAT

Universite Catholique de Louvain, Brussels, Belgium; Uppsala Conflict Data Program - Uppsala University; Institute for Health Metrics and Evaluation, 2013; World Health Organization, 2012).

2.2 Models

This analysis focuses on two models. Model (1) reflects the current literature's approach to evaluate the dynamic nature of the DAHG effect. The variables used on the right-hand side of model (1) are described in Table 1. One feature distinguishes model (2) from model (1). In model (2) DAHG is parsed into two distinct variables: DAHG and $DAHG^-$. $DAHG^-$ is the interaction between DAHG and a binary indicator marking a year-over-year decrease in DAHG. ($DAHG^-$ is thus zero for all country-years where DAHG is the same or larger than the previous year.) By modeling increases and decreases in DAHG separately, model (2) measures the displacement and replacement of GHES without assuming these two rates are equal. $\hat{\beta}_1$ is the estimated displacement rate (which follows an increase in DAHG), while $\left(\hat{\beta}_1 + \hat{\beta}_{1spread}\right)$ is the estimated replacement rate (which follows a decrease in DAHG).

$$(1)\ GHES_{it} = \alpha_i + \beta_1 DAHG_{it} + \beta_2 DAHNG_{it} + \beta_3 GGE_{it} + \beta_4 GDP_{it} + \beta_5 War_{it} + \beta_6 Disaster_{it} + \beta_7 Pop_{it} + \varepsilon_{it}$$

$$(2)\ GHES_{it} = \alpha_i + \beta_1 DAHG_{it} + \beta_{1spread} DAHG^-_{it} + \beta_2 DAHNG_{it} + \beta_3 GGE_{it} + \beta_4 GDP_{it} + \beta_5 War_{it} + \beta_6 Disaster_{it} + \beta_7 Pop_{it} + \nu_{it}$$

There is a great deal of country-level heterogeneity across my panel, which is addressed by including time-invariant, country-level fixed effects. Subscripts i and t indicate t country and year of each observation. War and natural disaster deaths are included to control for contextual shocks which also affect health expenditure, while population is included to adjust for the

diverse size of countries across my sample. A brief description of the method used to estimate models (1) and (2) is provided below, while a more complete description is reserved for the supplementary appendix.

2.3 Estimation

The Wooldridge test is used to identify autocorrelation in GHES. The null hypothesis is that no autocorrelation exists (Drukker, 2003; Wooldridge, 2001). I find that hypothesis is soundly rejected for both models ($p = 0.000$ for both modes). To deal with autocorrelation, I include a one-year lag of the dependent variable (LDV). To avoid introducing the bias common in short panels that include fixed-effects and a LDV, I use Difference GMM to estimate the model (Arellano and Bond, 1991; Arellano and Bover, 1995; Blundell and Bond, 1998; Nickell, 1981). Difference GMM first-differences each model across time to control of unobserved country level heterogeneity and uses internally derived instruments to control for endogenous variables. This estimator is ideal for repeated panels with few time periods where autocorrelation within countries, unobserved fixed-effects across countries and endogenous independent variables could otherwise bias the results (Roodman, 2009). In both models I treat DAHG and the LDV as endogenous because of the statistical relationship binding these variables to GHES.

A common problem with Difference GMM is the accumulation of too many instruments. Over-instrumenting can simultaneously over-fit endogenous variables and weaken the test used to assess the exogeneity of the instrument set (Roodman, 2007). I simultaneously employ two methods to address this problem. I "collapse" the instrument set (Roodman, 2007, 2009), and use principal components analysis to factor the instrument set (Bai and Ng, 2010; Roodman,

2009). Simulation shows that when combined, these techniques minimize bias (Mehrhoff, 2009).

Two tests support the validity of the internal instruments used in Difference GMM. First, I use the Arellano-Bond test to identify the degree that autocorrelation exists after country-level fixed effects are parsed out (Arellano and Bond, 1991; Roodman, 2009). If autocorrelation is present, then part of the instrument set is not valid. The null hypothesis is that no autocorrelation exists. I use lags twice removed (and greater) of DAHG and the LDV as instruments, so I test for first-order auto correlation in levels using the AR(2) test in differences. Second, I use the Hansen J test to test the exogeneity of the instrument set. The null hypothesis is that the full instrument set is exogenous. As the p-value for this test moves towards zero, causal claims become tenuous.

In addition to the internally derived instruments standard to Difference GMM, I include HIV prevalence rates and indicator variables marking the receipt of development assistance from the President's Emergency Plan for AIDS Relief and the Global Fund to Fight Aids, Tuberculosis and Malaria. These excluded instruments are predictive of DAHG, but when included directly in the model, none of these instruments affect changes in GHES. Therefore, they seem to affect GHES only via DAHG. Furthermore, excluding them from the estimation does not substantively change my results.

For model (2), a Wald test is used to assess the linear hypothesis that the coefficient $\hat{\beta}_{1spread}$ is statistically different from zero. The null hypothesis of this test is $\hat{\beta}_{1spread} = 0$. Rejecting this

hypothesis is the same as rejecting $\hat{\beta}_1 = \left(\hat{\beta}_1 + \hat{\beta}_{1spread}\right)$, and confirms governments displace and replace GHES at different rates.

2.4 Robustness

I assess the sensitivity of the results across a wide set of alternative datasets, estimation specifications, variable specifications, and estimation methods. For model (1) and (2), I complete more than 350 additional regressions. Each additional regression varies a single component of the estimation. These results provide evidence that the baseline results reported in this study are robust. For further description of these methods, primary results and sensitivity analyses, see Tables S5 through S20 and Figures S1 and S2 in the supplementary appendix.

3. Results

Table 2 reports my primary results – the displacement and replacement of GHES. Both models (1) and (2) show DAHG is a significant determinant of GHES. Model (1) shows that on average $0.65 (CI: 0.12, 1.18) of GHES is displaced for every $1 of DAHG. This leaves $0.35 (CI: -0.18, 0.88) of every $1 of DAHG additional.

Model (1) also shows that GHES is dynamic, as the LDV is significantly different from zero at the 99% confidence level. This means that on average this year's level of government spending is a function of last year's level of spending. The long-run effect (total loss of GHES over time for every $1 of DAHG) is estimated to be $1.39 (CI: 0.41, 2.37). $0.65 of GHES is sacrificed upon the receipt of $1 of DAH. Because GHES does not "bounce back" after being displaced

contemporaneous, the next year's level of spending remains $0.35 less than it would have been without the initial $1 of DAH, even without additional displacement. This is illustrated in Figure 1. GHES is dynamic, such that the effect of a hypothetical one-year DAHG shock in 1997 persists beyond the 1997. GHES only partially rebounds in 1998, and the simulation shows how a shock has an on-going effect on GHES beyond 2001.

Model (2) explores one reason why the effects of DAHG shocks persist. This model disaggregates the effects caused by increases in DAHG from decreases in DAHG. While the former are expected to decrease GHES, the latter (if it exists) will serve to increase it. Model (2) shows that an increase in DAHG leads to significant and sizable displacement: for every $1 increase in DAH, GHES decreases by $0.62 (CI: 0.09, 1.15). This is not statistically different than the estimate generated in model (1), but is statistically different from zero at the 95% confidence interval. Decreases in DAHG, on the other hand, have a relatively small effect on GHES. For every $1 decrease in DAHG, GHES increases by $0.27 (CI: -0.41, 0.96). This estimate is not statistically different from zero at any conventional level. Most important, there is some evidence that these two rates–the displacement rate and the replacement rate–are statistically different from each other ($p = 0.098$). This statistical assessment confirms it is likely that DAHG is being displaced at a larger rate than it is being replaced.

Figure 2 illustrates the cost of displacement. This figure, based on model (2), compares two different scenarios: GHE^{Actual} and GHE^{NoDis}. GHE^{Actual} is total (modeled) government health expenditure, including both domestic and foreign components based on actual aid disbursements. GHE^{NoDis} is a counterfactual that assumes a policy where recipient governments had not

displaced any DAHG between 1995 and 2010. I calculate the cost of displacement by subtracting GHE^{Actual} from GHE^{NoDis}. Over the course of time, the graph shows that aid accumulated since 1995 has an increasing effect on GHE. The one-year cost of displacement grows over time, to $7.9 billion in 2010, as the adverse effects of 15 years of displacement are exacerbated due to disproportionally low replacement rates and the dynamic nature of GHES.

Figure 3 illustrates an alternative comparison. This figure illustrates the cost of sporadic DAHG disbursements. As in Figure 2, GHE^{Actual} is the modeled value of GHE and includes the cost of displacing GHES without comparable replacement. In Figure 3, GHE^{Actual} is compared to GHE^{Smooth}, which is a second counterfactual. Rather than assuming recipients do not displace (as in Figure 2), Figure 3 assumes that donors disbursed DAHG at a monotonically increasing amount. Thus, GHE^{Smooth} does not acquire the loss associated with stunted replacement, since replacement never occurs when DAHG is only increasing. In this scenario, each country receives their own country-specific 16-year aggregate level of DAHG so that the two scenarios differ only in the timing over which the DAHG was disbursed. For Figure 3, the cost is sporadic DAHG is calculated by subtracting GHE^{Actual} from GHE^{Smooth}. Figure 3 shows the cost of displacement accrues over time with an increasing gap between GHE^{Actual} and GHE^{Smooth}. In 2010, the cost of 16 years of reduced spending leads to a one-year cost of $4.4 billion.

While not displacing DAHG (Figure 2) or disbursing DAHG at a stable rate (Figure 3) are two policy options that mitigate the adverse effects of displacement, a third policy option is recipient governments could displace and replace DAHG at an equal rate. While potentially more tractable politically, I calculate that the cost of asymmetric displacement and replacement rates

has led to a smaller cost. Figure 3 compares all three policy alternatives. Between 1995 and 2010, had recipient governments not displaced DAHG, $46.7 billion more resources would have been available as GHE for the 132 countries in my sample. Alternatively, had donors disbursed DAHG at an ever-non-decreasing rate, $21.0 billion of additional resources would have been available as GHE. Finally, had recipients displaced and replaced at the same rate, $8.4 billion more resources would have between this same period for GHE.

In addition to reporting the displacement and replacement rates, Table 2 also reports how other determinants affect GHES. GDP has a consistent and large effect on GHES. A 1% increase in GDP per capita translates to a 0.0001 increase in GHES (CI: 0.0000, 0.0213), where GHES is measured as a share of GDP. This confirms other reports that suggest that government health spending is a luxury good, and demand for GHES increases as incomes rise. DAHNG, GGE and natural disasters deaths are also positively associated with GHES, such that an increase in any of these variables translates into an increase in GHES. Model (2) shows that addition to the gains assigned to gains in GDP, a $1 increase in in the non-health sector government spending is associated with a GHES increase by roughly $0.02 (CI: -0.00, 0.04). Similarly, a $1 increase in DAHNG increases GHES by roughly $0.35 (CI: -0.06, 0.75). On the other hand, 1 death per 10,000 persons caused by armed conflict reduces GHES by $0.03 (CI: 0.01, 0.05). These results reflect relationships described previously in this literature and serve to validate my primary results (Lu et al., 2010; Farag et al., 2009; Xu et al., 2011).

To evaluate the validity of these results, I examine my instrument sets and specifications. I do not find any evidence of autocorrelation (after parsing out the country-level effects) in either

model (p-value = 0.20 and 0.53, respectively). Nor do I find any evidence of endogenous instruments (p-values = 0.48 and 0.53, respectively). Additionally, the supplementary appendix reports a series of over 350 sensitivity analyses, the vast majority of which support these results.

4. Discussion

When development assistance displaces government expenditure, the sector-specific net gain is not what it could have been sans displacement. On average in the short run, only $0.35 (CI: -0.18, 0.88) of every $1 of DAHG is additional, while $0.65 (CI: 0.12, 1.18) worth of GHES is displaced. Furthermore, reductions in GHES persist over time, even beyond the period in which DAHG is being disbursed. This research tests why this might be, hypothesizing that displacement of GHES might not be asymmetric. I find some evidence of this, as the replacement rate (associated with decreasing DAHG) is less than the original displacement (associated with increasing DAHG). This suggests that a one-year shock of $1 leaves a government health system with a financing gap much larger than it first appears as it takes time for government to reallocate resources back to the health system. These results imply that fluctuations and unanticipated reductions in DAHG have serious ramifications for total health sector expenditure in the long-run. At a time when the supply of DAH has stagnated, the possibility that displacement imposes financing constraints on the health sector has important and timely ramifications (Leach-Kemon et al., 2011; Institute for Health Metrics and Evaluation., 2011; Glassman, 2012).

These conclusions lead to three important questions about government behavior. First, why does displacement occur? Economic theory suggests that welfare-maximizing governments allocate resources according to the marginal gains associated with achieving their priorities. If the government receives large amounts of development assistance for health relative to development assistance for non-health sectors, then a rational government displaces GHES (Pack and Pack, 1993, 1990; Oates, 1999; Feyzioglu et al., 1998; Ooms et al., 17; van de Walle and Mu, 2007). This argument assumes that development assistance is fully fungible and that moving resources across sectors is costless. As a consequence health aid "crowds-out" governments' own resources for health. In the global scale up of development assistance, assistance to health grew at faster rate than assistance to other sectors (Institute for Health Metrics and Evaluation., 2011, 2009). This suggests the scale-up in health that outpaced the growth of aid to other sectors could be one explanation for displacement. Still, confirmation of this theory requires detailed financial data that is not currently available. To evaluate relative flows, researchers would need data tracking government expenditure and development assistance channeled to governments disaggregated across all sectors. Without these data, my ability to fully understand this phenomenon is limited.

The second question related to government behavior is why the GHES displacement rate does not equal the replacement rate. One intuitive answer is that it is easier to share a windfall than reallocate funds across sectors during a time of austerity. In general, governments are constrained by finite budgets that must be allocated across many competing priorities. Ceteris paribus, an increase in DAHG leads to a net gain in government resources. This research shows that these gains are not restricted to the health sectors and the cost of sharing resources across

sectors when net gains are positive is relatively small. Indeed, when DAHG increases, 65% (CI: .12, 1.18) of it is effectively shared with non-health sectors. On the other hand, decreases in DAHG cause the government's budget to contract. Returning displaced funds back to the health sector would necessitate cutting funds from other sectors. This research suggests that is relatively difficult, and thus policy makers are less successful at achieving it.

The third question related to government behavior is if the displacement and replacement estimated here is similar to other estimates of aid displacement? While to the best of my knowledge no papers assess asymmetric fungiblity, there is a related empirical literature that assumes symmetry across increases and decreases in aid. This literature can be disaggregated into three different levels. The first level, which is the most general, assesses how all-sector aid affects total government expenditure and revenue. Here, the more recent literature suggests that aid has positive effect on government expenditure and may also have a positive effect on domestic revenue (Pack and Pack, 1990, 1993; Clist and Morrissey, 2011; Feyzioglu et al., 1998; Devarajan et al., 2002).

The second level of this literature assesses how all-sector aid affects individual government sectors, with several analyses assessing the health sector. Here the literature is more mixed. A number of papers show a positive aid effect (Morrissey, 2012; Pack and Pack, 1990), with estimates ranging up to a $0.55 gain in GHE for $1 of aid (Pack and Pack, 1993). On the other hand, several other analyses were not able to measure an effect that all-sector aid has on health spending relationship (Fernandes Antunes et al., 2012; Van de Sijpe, 2012; Swaroop et al., 2000; Feyzioglu et al., 1998; Swaroop et al., 1999).

The third level of assessment in this literature is most related to this present study. These papers consider how health-sector specific aid affects the government health sector. Three papers that assess this relationship are not able to identify estimates that are statistically different from zero (Swaroop et al., 1999; Feyzioglu et al., 1998; Swaroop et al., 2000). Still, four other papers are able to identify statistically significant measurements that can be compared to those estimated here. In 2010 Lu et al. used a smaller, but very similar dataset also from IHME to show that $1 of DAHG leads to $0.46 decrease in GHES. That same year Farag et al. (2010) used a similarly sized dataset from a different source to estimate that $1 increase in DAH leads to a decrease in GHES between $0.27 and $0.61. In 2012, Fernandes Antunes et al. estimated a similar relationship, leading to a $0.42 reduction in GHES, while Van de Sijpe (2012) created and used a unique data set, and estimated that (a variable comparable to) DAHG leads to a reduction in $0.16 GHES. Placing my own work into this literature, my estimates show slightly more substitution, showing a reduction of $0.65 of GHES.

This research has several limitations. First, the estimated effects reflect a global average for 132 countries across 16 years. It would be inappropriate to apply results from this study to any one country in any single year. In an attempt to identify heterogeneous effects across regions, I conduct subset analyses focusing separately on Sub-Saharan Africa, South and Eastern Asia, and Latin America (see the supplementary appendix). Small samples prevent precise estimates, but these analyses generally confirm displacement occurs. Second, this study makes no claims about population health or welfare. This study confirms that DAHG displaces GHES with long-standing effects, but it would be incorrect to assume that substitution has a specific effect on

population health or welfare. While it is often assumed that fewer resources in a government's health sector translate to worse health, it remains possible that displacement has a positive effect on population health or welfare via alternative government sectors (McGillvray and Morrissey, 2001; Wagstaff, 2011). If displaced GHES is substituted into other pro-health sectors such as education, positive health gains might still be achieved (Lu et al., 2010; Sridhar and Woods, 2010; Gakidou et al., 2010). Without project-level data about where displaced funds are being channeled, the welfare and health effects related to subadditionality are ambiguous. A final limitation is in my ability to detect a causal effect. Since I measure GHES by netting DAHG from GHE, identifying the direction of the causal connection is exceedingly difficult. I believe my estimates provide evidence of a causal path connecting development assistance for health to government health expenditure. My estimates, extensive set of sensitivity analyses (see supplementary appendix), and economic theory inform my position that that increases in DAHG lead to decreases in GHES. Still, my evidence is not equivalent to the definitive proof that would be provided by a randomized control trial or natural experiment. This research uses the best available statistical methods, expenditure data, and health aid data. However, Difference GMM is a fragile and complicated estimation method, and future work should continue to critically assess their applicability to this issue.

This research leads to five policy recommendations. First, donors and recipients should work together to maintain stable flows of development assistance. Stable financing has positive implications for projects and allow for more effective long-term planning (Cavagnero et al., 2008; Woods, 2005). Moreover, stable disbursement of DAHG is integral to maintaining consistent levels of health financing and preventing financial gaps caused by displacement of

GHES without subsequent replacement. Figure 3 shows that there are significant costs associated with sporadic DAHG disbursement and there would be significant gains if DAHG were disbursed smoothly.

The second recommendation is that recipient governments equate the displacement and replacement rates of GHES. While competing demands for new resources will always exists, incomplete replacement only punishes the health sector. Fluctuating aid disbursements does not need to lead to such distortion. The asymmetry between displacement and replacement are, presumably, a function of internal politicking rather than a means of increasing population welfare.

The third recommendation is that policy-makers, donors, and researchers place more emphasis on tracking domestic and international financial flows. Data limitations force researchers to rely on complicated estimation methods. State-of-the-art estimation methods, specification tests, and sensitivity analyses are a poor substitute for better data. More granular measurements of development assistance and more comparable and transparent measurements of government expenditure could lead to more exact measurement of displacement and replacement. Moreover, improved data would permit analysts to address a number of new questions, including identifying heterogeneous effects of DAHG and the welfare implications of displacement. By improving the tracking of financial flows, researchers, donors, and policy-makers could become more aware of which countries are substituting their own resources and where displaced funds are going. A comprehensive database on financial flows related to health would empower

researchers to more thoroughly assess which types of flows lead to the greatest health gains and measure the health losses (or gains) associated with displacement and replacement.

The forth recommendation is that cost-effectiveness analyses of health projects incorporate the long-term costs related to financial flows. Evidence from this study indicates that short-term financing may create long-term financing gaps. Even if future flows are heavily discounted, short-term projects with external, unsustainable funding may have consequential long-term financial consequences. The effectiveness analyses of these projects should assess if the long-term gains associated with the project offset long-term financial costs imposed on the broader public health system.

Finally, the fifth recommendation is that donors might consider how the health projects they finance align with the country's priorities. If displacement is a forgone conclusion, then it would be advantageous for donors to align their projects with the governments' priorities. This conforms to the Paris Declaration which encourages "increasing alignment of aid with partner countries' priorities, systems and procedures and helping the strengthening of their capacities (Paris Declaration, 2005)". Without such alignment, displacement will shift health service away from those set by ministries of health in favor of the donor's priorities. Given the potentially narrow focus of donors and the relatively broad set of tasks for which public health systems are responsible, substituting domestic resources with DAHG may have problematic outcomes. On the contrary, if displacement is not a foregone conclusion, then donors could finance new projects without affecting the provision of existing service. The complex interplay between the

amount of financial displacement and the displacement of actual health services is important for donors to consider. It is an area of research that deserves more attention.

This study confirms that a significant portion of government health expenditure as source is substituted out of the health sector upon the receipt of development assistance for health. Further, it shows that this effect is neither symmetric nor static. Increases in DAHG have a larger effect on GHES than do decreases in DAHG, and thus the adverse effects of receiving DAHG can persist over time. This study provides an important reminder of the value of maintaining stable financing for health. The set of known cost-effective health interventions is expanding, yet the global burden of disease remains high and disproportionate. Financial resources for health have the potential to positively impact population health, so understanding how financial flows interact with each other is of critical importance.

Acknowledgements

In addition to those thanked in the prefix, I would like to thank David Roodman, Jack Langenbrunner, the participants at a Health System Strengthening conference seminar (Beijing, October 2012), and the participants at a PHEnOM weekly seminar (Seattle, January 2013) for their valuable feedback regarding this chapter. Any remaining errors are of course my own.

Abbreviations

AR(2)	Arellano-Bond test for autocorrelation
CI	95% Confidence interval
DAH	Development assistance for health
DAHG	Development assistance for health channeled to governments
DAHG	Development assistance for health not channeled to governments
GDP	Gross domestic product
GGE	General government expenditure, net of government health expenditure
GHE	Total government health expenditure (as agent, also known as GHEA)
GHES	Government health expenditure as source
GMM	General method of moments
HIV	Human immunodeficiency virus
IHME	Institute for Health Metrics and Evaluation
LDV	Lag dependent variable
WHO	World Health Organization

References

Arellano M, Bond S. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. The Review of Economic Studies 1991; 58; 277.

Arellano M, Bover O. Another look at the instrumental variable estimation of error-components models* 1. Journal of Econometrics 1995; 68; 29–51.

Bai J, Ng S. Instrumental variable estimation in a data rich environment. Econometric Theory 2010; 26; 1577–1606.

Batniji R, Bendavid E. Does Development Assistance for Health Really Displace Government Health Spending? Reassessing the Evidence. PLoS Med 2012; 9; e1001214.

Blundell R, Bond S. Initial conditions and moment restrictions in dynamic panel data models. Journal of Econometrics 1998; 87; 115–143.

Cavagnero E, Lane C, Evans D, Carrin G. Development assistance for health: should policy-makers worry about its macroeconomic impact? Bulletin of the World Health Organization 2008; 86; 864–870.

Clist P, Morrissey O. Aid and tax revenue: Signs of a positive effect since the 1980s. Journal of International Development 2011; 23; 165–180.

Devarajan S, Miller M, Swanson E. Goals for development: History, prospects, and costs. World Bank Policy Research Working Paper No. 2819 2002.

Dodd R, Schieber G, Cassels A, Fleisher L, Gottret P. Aid effectiveness and health 2007.

Drukker D. Testing for serieal correlation in linear panel-data models. The Stata 2003; 3; 168–77.

EM-DAT Universite Catholique de Louvain, Brussels, Belgium. The OFDA/CRED International Disaster Databse. Available at: http://www.emdat.be/database.

Farag M, Nandakumar AK, Wallack SS, Gaumer G, Hodgkin D. Does funding from donors displace government spending for health in developing countries? Health Affairs 2009; 28; 1045.

Fernandes Antunes A, Xu K, James CD, Saksena P, Van de Maele N, Carrin G, Evans DB. General Budget Support: Has It Benefited the Health Sector? Health Economics 2012; n/a–n/a.

Feyzioglu T, Swaroop V, Zhu M. A panel data analysis of the fungibility of foreign aid. The World Bank Economic Review 1998; 12; 29.

Gakidou E, Cowling K, Lozano R, Murray CJ. Increased educational attainment and its effect on child mortality in 175 countries between 1970 and 2009: a systematic analysis. The Lancet 2010; 376; 959–974.

Garg CC, Evans DB, Dmytraczenko T, Izazola-Licea J-A, Tangcharoensathien V, Ejeder TT-T. Study Raises Questions About Measurement Of "Additionality,"Or Maintaining Domestic Health Spending Amid Foreign Donations. Health Affairs 2012; 31; 417–425.

Glassman A. 2012. Ethiopia's AIDS Spending Cliff » Global Health Policy. Center for Global Development. 2012. Available at: http://blogs.cgdev.org/globalhealth/2012/09/ethiopias-aids-spending-cliff.php.

Institute for Health Metrics and Evaluation. 2013. Financing global health 2012: End of the golden age? Seattle, United States; January 2013.

Institute for Health Metrics and Evaluation. Financing Global Health 2009: Tracking Development Assistance for Health. IHME: Seattle, WA; 2009.

———. Financing Global Health 2011: Continued Growth as MDG Deadline Approaches. IHME: Seattle, WA; 2011.

James SL, Gubbins P, Murray CJ, Gakidou E. Developing a comprehensive time series of GDP per capita for 210 countries from 1950 to 2015. Population Health Metrics 2012; 10; 12.

Leach-Kemon K, Chou DP, Schneider MT, Tardif A, Dieleman JL, Brooks BPC, Hanlon M, Murray CJL. The Global Financial Crisis Has Led To A Slowdown In Growth Of Funding To Improve Health In Many Developing Countries. Health Affairs 2011. http://content.healthaffairs.org/cgi/doi/10.1377/hlthaff.2011.1154.

Lu C, Schneider MT, Gubbins P, Leach-Kemon K, Jamison D, Murray CJ. Public financing of health in developing countries: a cross-national systematic analysis. The Lancet 2010; 375; 1375–1387.

McGillvray M, Morrissey O. Aid Illusion and Public Sector Behaviour. Journal of Development Studies 2001; 37; 118–136.

Mehrhoff J. A Solution to the Problem of Too Many Insttruments in Dynamic Panal Data GMM. Discussion Paper Deutsche Bundesbank 2009. Series 1: Economic Studies. http://www.econstor.eu/bitstream/10419/19691/1/200712dkp.pdf.

Morrissey O. 2012. Aid and Government Fiscal Behaviour: What Does the Evidence Say? Working Paper UNU-WIDER Research Paper WP2012/01. World Institute for Development Economic Research (UNU-WIDER); 2012. Available at: http://ideas.repec.org/p/unu/wpaper/wp2012-01.html.

Nickell S. Biases in dynamic models with fixed effects. Econometrica: Journal of the Econometric Society 1981; 1417–1426.

Oates WE. An essay on fiscal federalism. Journal of Economic Literature 1999; 37; 1120–1149.

Ooms G, Decoster K, Miti K, Rens S, Van Leemput L, Vermeiren P, Van Damme W. Crowding out: are relations between international health aid and government health funding too complex to be captured in averages only? The Lancet 17; 375; 1403–1405.

Over M. How will the financial crisis affect aid to the health sector? Center for Global Development. Wash; June 12, 2009.

Pack H, Pack JR. Is foreign aid fungible? The case of Indonesia. The Economic Journal 1990; 100; 188–194.

———. Foreign Aid and the Question of Fungibility. The Review of Economics and Statistics 1993; 75; 258–265.

Paris Declaration. 2005. Paris Declaration. 2005.

Roodman D. A note on the theme of too many instruments. Center for Global Development, Washington, DC 2007.

Roodman D. How to do xtabond2: An introduction to difference and system GMM in Stata. The Stata Journal 2009; 9; 86–136.

———. Doubts about the evidence that foreign aid for health is displaced into non-health uses. The Lancet 2012; 380; 972–973.

Van de Sijpe N. Is Foreign Aid Fungible? Evidence from the Education and Health Sectors. The World Bank Economic Review 2012. http://wber.oxfordjournals.org/content/early/2012/10/30/wber.lhs023.

Sridhar D, Woods N. Are there simple conclusions on how to channel health funding? The Lancet 2010; 375; 1326–1328.

Swaroop V, Devarajan S, Rajkumar AS. What Does Aid to Africa Finance? The World Bank, Policy Research Working Paper Series: 2092; 1999.

Swaroop V, Jha S, Sunil Rajkumar A. Fiscal effects of foreign aid in a federal system of governance: The case of India. Journal of Public Economics 2000; 77; 307–330.

UNAIDS and WHO. Global report: UNAIDS report on the global AIDS epidemic 2010. 2010.

United Nations Population Division. World Population Prospects, the 2010 Revision. May 2011.

Uppsala Conflict Data Program - Uppsala University. UCDP Battle-Related Deathes Dataset v.5-2011. Available at: http://www.pcr.uu.se/research/ucdp/datasets/ucdp_battle-related_deaths_dataset/.

Wagstaff A. Fungibility and the impact of development assistance: Evidence from Vietnam's health sector. Journal of Development Economics 2011; 94; 62–73.

Van de Walle D, Mu R. Fungibility and the flypaper effect of project aid: Micro-evidence for Vietnam. Journal of Development Economics 2007; 84; 667–685.

Woods N. The shifting politics of foreign aid. International Affairs 2005; 81; 393–409.

Wooldridge JM. Econometric Analysis of Cross Section and Panel Data. 1st ed. The MIT Press; October 1, 2001.

World Health Organization. Guide to Producing National Health Accounts 2003.

———. World Health Statistics 2012. World Health Organization: Geneva; 2012.

Xu K, Saksena P, Holly A. The Determinants of Health Expenditure: A Country-Level Panel Data Analysis. Working Paper 2011.

Tables and Figures

Table 1: Definitions of variables

Abbreviation	Variables
$GHES_{it}$	Government health expenditure as source; measured as a percentage of the country's mean GDP for country i in year t
$DAHG_{it}$	Development assistance for health channeled to a government; measured as a percentage of the country's mean GDP for country i in year t
$DAHG_{it}$	Development assistance for health not channeled to a government; measured as a percentage of the country's mean GDP for country i in year t
GGE_{it}	General government expenditure, net of government health expenditure as source; measured as a percentage of the country's mean GDP for country i in year t
GDP_{it}	Gross domestic product; measured per capita and log transformed for country i in year t
POP_{it}	Population for country i in year t; log transformed
WAR_{it}	Combat deaths per 1,000 people; for country i in year t
$DISASTER_{it}$	Natural disaster deaths per 1,000 people; for country i in year t
α_i	Unobserved time-invariant country-specific characteristics (fixed effects)
α	A constant
ε_{it}, ν_{it}, υ_{it}	Error term for country i in year t

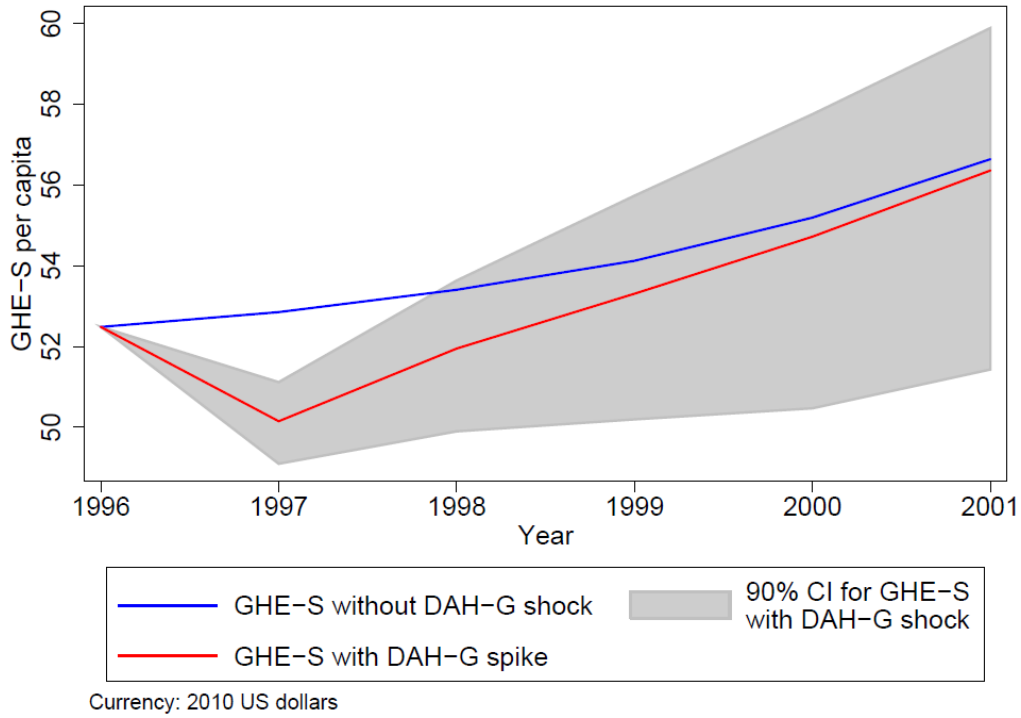
Table 2: Regression results

		Independent Variables																			
		LDV	p-value	DAHG	p-value	DAHG-	p-value	DAHNG	p-value	GGE	p-value	GDP	p-value	POP	p-value	War	p-value	Disaster	p-value	Hansen J Test p-value	AR(2) Test p-value
(1)	GHES	0.53	0.00	-0.65	0.02			0.39	0.00	0.02	0.02	0.01	0.03	0.00	0.95	-0.02	0.01	2.66	0.01	0.48	0.20
(2)	GHES	0.53	0.00	-0.62	0.02	0.34	0.10	0.22	0.15	0.02	0.04	0.01	0.05	0.00	0.95	-0.03	0.01	0.22	0.97	0.53	0.53

Model (1) and (2) are estimated using Difference GMM.

Coefficient estimates for time and region indicators reported in supplementary appendix.

Figure 1: One-year DAHG shock has ongoing effect on GHES



GHES is dynamic, meaning a one-year DAHG shock has an ongoing effect on GHES. The GHES lines are the fitted estimates of GHES per capita for the average country. The gap between the blue and red line shows the ongoing effect of a one-year DAHG shock. The blue line assumes that DAHG is constant across the 5 year period (at the sample's mean, roughly $6 per capita). The red line shows the consequence of a one-year 50% increase in DAHG in 1997 (and a subsequent return to the original level). For both lines, the other independent variables (non DAHG variables) are set to be the mean across the 132 countries. The 95% CI is generated taking the 1000 random draws from the estimated variance-covariance matrix.

Figure 2: The cost of displacement and replacement between 1995 and 2010

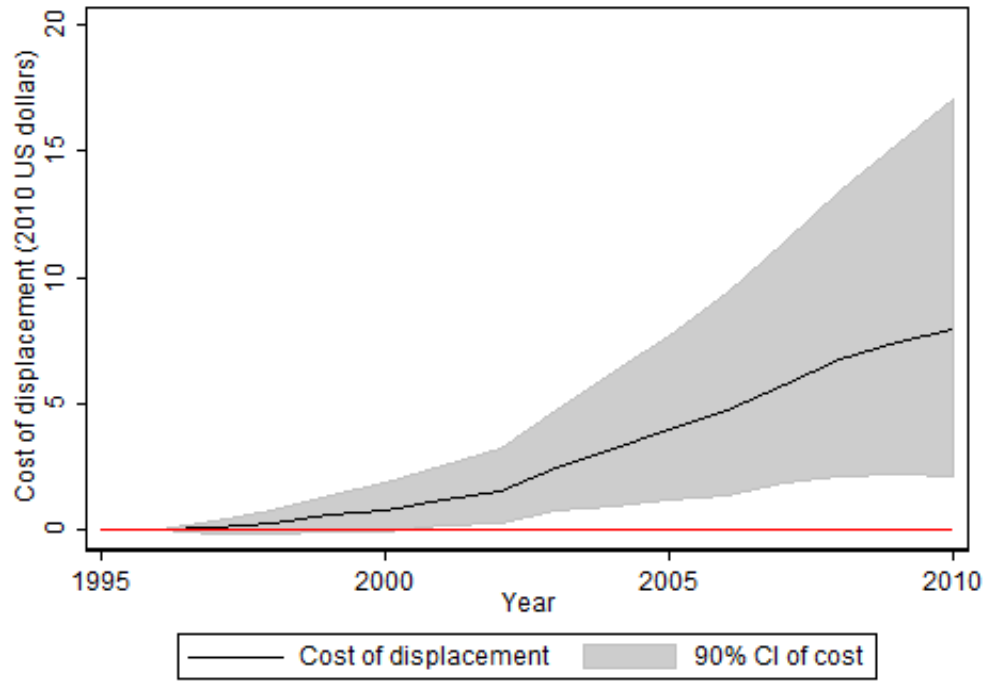


Figure 2 compares the effect of the actual scale-up of DAHG between 1995 and 2010 and a counterfactual. In the counterfactual scenario, recipient countries do not displace (or replace) any resources upon the receipt of DAHG. Cost is defined as the difference between GHE^{NoDis} (modeled assuming displacement = replacement = 0) and GHE^{Actual} (modeled assuming displacement and replacement rates estimated in model (2)). The cost is positive and significant showing had recipient countries not displaced resources more total government health resources would have been available. The 90% CI is generated taking the 1000 random draws from the estimated variance-covariance matrix.

Figure 3: The cost of sporadic disbursement of DAHG between 1995 and 2010

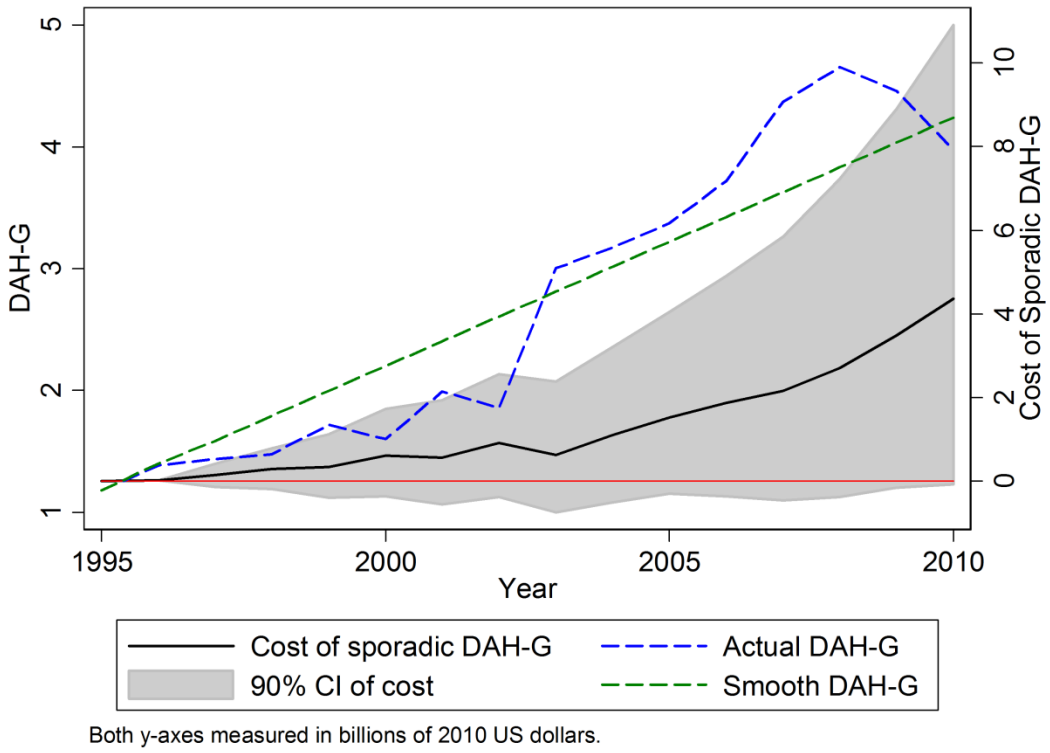
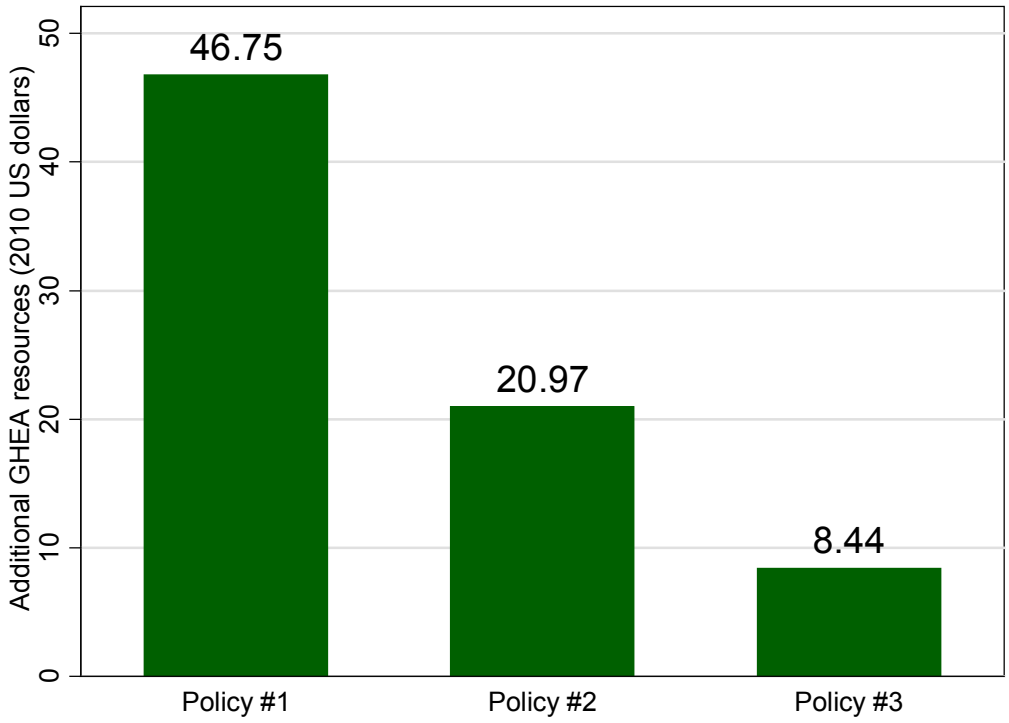


Figure 3 compares the effect of the actual scale-up of DAHG between 1995 and 2010 and a counterfactual. In the counterfactual scenario, the total amount of DAHG disbursed per country, but disbursement is only increases over time. (There are no country-level DAHG reductions.) The black solid line shows the cost of sporadically disbursed DAHG. Cost is defined as the difference between GHE^{Smooth} (caused by smooth DAHG disbursement) and GHE^{Actual} (caused by actual DAHG disbursement). The cost is positive and significant showing that a smooth DAHG scale would have led to more total government health resources. The 90% CI is generated taking the 1000 random draws from the estimated variance-covariance matrix.

Figure 4: Policy options to limit effects of displacement



Between 1995 and 2010, $43.8 billion of DAHG was disbursed. Figure 4 compares three possible policies that could have been used to insert more resources to the government health sector. Policy #1 is based on Figure 2 and assumes that recipient countries do not displace any DAHG. Policy #2 is based on Figure 3 and assumes that donor countries disburse DAHG at a non-decreasing rate. Policy #3 assumes that recipients displaced and replaced increases and decreases in DAHG at the same expected rate of displacement, where the expected rate of displacement is 1-(GHE/GGE).

Chapter 2: Appendix

1. Data

DAHG is reported for 156 low- and middle-income countries by the Institute for Health Metrics and Evaluation in Financing Global Health 2012: End of the Golden Age (Institute for Health Metrics and Evaluation, 2013). For each of these countries, DAHG is reported for 16 years, 1995 through 2010 in 2010 US dollars, resulting in 2496 observations.

The World Health Organization (WHO) reports general government health expenditure data (also known as government health expenditure as agent; GHE) in current local currency units (LCUs) (World Health Organization, 2012). 2,434 country-years of WHO data match the 2,496 country-years of DAHG data from IHME. I convert GHE reported in current LCUs to 2010 US dollars by first deflating to 2010 LCUs and then exchanging to 2010 US dollars. I use the IMF deflator series. For 18 countries in my sample, the IMF doesn't report any deflators for a country. In these cases, when available, I use deflators from the World Bank. (International Monetary Fund, 2012; The World Bank). I use the WHO 2010 exchange rate series to convert to US dollars (World Health Organization, 2012). Similarly, when WHO exchange rates aren't available for a country, I defer to World Bank exchange rates. Because of missing deflators or exchange rates, 49 observations can't be converted to 2010 US dollars. Only two years of GHE data for Zimbabwe are converted, which is not enough to enter my model using the primary estimation method, so both Zimbabwe estimates are dropped. This results in 2,383 observations across 132 countries, and makes up my primary dataset.

By definition, GHE ≡ GHES + DAHG (Lu et al., 2010; World Health Organization, 2003). Thus, GHES measured in 2010 US dollars is obtained by subtracting DAHG from GHE. All financial variables are measured as a share of the country-specific mean GDP, except GDP itself. GDP is measured in per capita terms and is log transformed. I measure each financial variable as a percentage of the country-specific mean GDP in order to standardize my data and prevent countries with large expenditure or population from dominating my analysis. Dividing by country-specific mean GDP (rather than contemporaneous GDP) avoids mislabeling changes over time. For example, suppose both GDP and GHES are increasing, but GDP is growing faster. In this scenario, GHES measured as a share of contemporaneous GDP will be decreasing, even though GHES was increasing in absolute terms. (Table S11 and S19 show that the estimated effects are robust to this decision.) Tables S1 through S4 include a list of countries included in the study, a complete list of data sources, and two tables of summary statistics (James et al., 2012; UNAIDS and WHO, 2010; EM-DAT Universite Catholique de Louvain, Brussels, Belgium; Uppsala Conflict Data Program - Uppsala University; Institute for Health Metrics and Evaluation, 2013; World Health Organization, 2012).

For the baseline regression estimates, I weight each observation by the total amount of DAHG received over the 16 year sample. Thus, countries that have received a great deal of DAHG, such as India, Tanzania, Ethiopia, and China, are heavily weighted. Using this weight-scheme affords a clearer interpretation of the DAHG effect. Financial effects now can be interpreted as total monies disbursed globally, rather than as a function of the country-years (Table S9 and S17 show that my estimates are not driven by this weighting scheme).

2. Estimating model (1) baseline

Model (1) assesses the effect that determinants have on GHES. The determinants used on the right-hand side of the model are quantities (levels), are common to this literature, and are described in Table 1 of the main text (Lu et al., 2010; Farag et al., 2009; Xu et al., 2011). Two additional variables, not used elsewhere in this literature, are included in my analysis. They are the number of deaths per thousand caused by natural disaster and the number of deaths per thousand caused by conflict. These variables are predictive of GHES at the 95% and 99% confidence interval, respectively. Model (1) also assumes that unobserved time-invariant country-level fixed-effects impact GHES, as there is a great deal of heterogeneity across my panel.

(1) $$GHES_{it} = \alpha_i + \beta_1 DAHG_{it} + \beta_2 DAHNG_{it} + \beta_3 GGE_{it} + \beta_4 GDP_{it} + \beta_5 War_{it} + \beta_6 Disaster_{it} + \beta_7 Pop_{it} + \varepsilon_{it}$$

The Wooldridge test is used to check for the presence of autocorrelation in GHES. I find that the null hypothesis of no autocorrelation is soundly rejected ($p = 0.000$) (Drukker, 2003; Wooldridge, 2001). To deal with this autocorrelation, I include a one-year lag of the dependent variable (LDV). To avoid biasing the estimation with the Nickell bias, I use Difference General Method of Moments (GMM) estimation (Arellano and Bond, 1991; Arellano and Bover, 1995; Blundell and Bond, 1998; Nickell, 1981). I treat DAHG and the LDV as endogenous (although Table S10 shows that my results are not dependent on this choice). I use Difference GMM, rather than the more efficient system-GMM because (a) it seems plausible that the initial condition necessary for the valid application of system-GMM might not be met, and (b) in some specifications the GMM instruments for the levels equation are not exogenous. This second

concern can be tested using the Hansen J test and the difference –in-Hansen test (although the difference-in-Hansen test is only applicable when principle components analysis – PCA – is not used to factor the instruments). When PCA is not used and my baseline specification is estimated using system-GMM, the Hansen J test suggests that the null hypothesis of exogenous instruments not rejected ($p = 0.341$). Still, as I explored other specifications, this outcome was not consistent. For simplicity, I defer to Differencing GMM throughout this analysis.

I "collapse" the instrument set (Roodman, 2007, 2009) and use principal components analysis (PCA) to factor the instrument set (Bai and Ng, 2010; Roodman, 2009). I retain components with eigenvalue greater than one. My GMM instruments are the second through fifteenth year lags of the endogenous variables. I use first-differencing transformation (as opposed to forward orthogonal deviation) and traditional lags as instruments (as opposed to backward orthogonal deviation of the instruments). I use the two step estimation method and Windmeijer adjusted robust standard errors (Windmeijer, 2005; Roodman, 2009) (Table S9 shows that these choices do not drive my estimates.)

In addition to the internally derived instruments standard to Difference GMM, I include HIV prevalence rates and indicator variables marking the receipt of development assistance from the President's Emergency Plan for AIDS Relief and the Global Fund to Fight Aids, Tuberculosis and Malaria. These excluded instruments are predictive of DAHG, but when included directly in the model, none of these instruments affect changes in GHES. Therefore, they seem to affect GHES only via DAHG. Table S11 and S19 show excluding these instruments from the estimation does not substantively change my results.

A single time (year) indicator is included as right-hand-side variables in my baseline model. Generally speaking, GHES is not subject to idiosyncratic time shocks. Across my sample, GHES has increased over time, but this can be explained by other determinants. When the complete series of time indicators is included in my baseline regression none of the time indicators are statistically significant at any conventional confidence level. The coefficient estimate for the indicator for 2009 has a p-value of 0.12, while the next lowest p-value is 0.57. Further analysis shows that the 2009 indicator does have some predictive power across my sample, while the rest of the indicators do not. This is seen using simple fixed effects estimation or including the indicators individually into the baseline estimation. Thus, to be cautious, my baseline model includes the 2009 indicator. A Wald test assessing the joint significance of the time indicators cannot reject the null hypothesis that the indicators net of 2009 are jointly different than zero ($p = 0.928$). (Tables S11 and S19 show that my conclusions are not dependent on its inclusion.)

Table S5 reports my baseline estimation of model (1), repeating the estimates reported in Table 2 of the main text. Tables S6 through S12 and Figures S1 through S3 are used to test the sensitivity of those results to various datasets, estimation methods, and specifications. Most tables repeat the baseline estimation in column (1).

3. Estimating model (1) specification and sensitivity tests

Table S6 reports regressions that serve as pre-tests to establish the validity of the baseline regression. Column (2) uses system-GMM to estimate model (1), without using PCA on the instruments. Column (3) includes time indicators on the right-hand-side of the equation, and shows that none of the time indicators (except 2009) are significantly different from zero (at the 50% confidence level).

As an additional pretest, I evaluate test model (1) (and (2)) for multicolinearity. I calculate the variance inflation factor (VIF) for each of the eight right-hand-side variables. Unsurprisingly, the LDV has the largest VIF, although it is only 1.84. The mean VIF across all eight variables is 1.34. Since all eight VIF are well below ten, the rule-of-thumb marker for concern, I assume multicolinearity is not hampering my results (Kennedy, 2003).

Table S7 assesses the stability of the DAHG effect by looking at different subsamples of the data. Columns (2) through (9) assess subsamples where the subsamples are the full sample net of observations which may be considered outliers. For each column, an outlier is defined as an observation in the 10% largest modified z-score, where the modified z-score is the absolute value of the observation's value subtracted from the median and then divided by the maximum absolute deviation. The modified z-score is commonly used to identify outliers (Iglewicz and Hoaglin, 1993; 16:). For column (2) I exclude 10% of the observations based on GHES outliers (using the modified z-score based on GHES estimates). Column (3) repeats this exercise excluding DAHG outliers, column (4) uses DAHNG, and so forth across all eight right-hand-

side. Across the eight subsamples, the DAHG effect fluctuates, but in all estimates remain between -0.49 and -1.10, and all but two of the regressions the displacement effect remains statistically significant at the 95% confidence level.

Column (10) of Table S7 omits ten countries from my sample and re-estimates my baseline estimation. These ten countries were chosen by assessing the influence that each country has on the DAHG coefficient estimate. Following Van de Sijpe, I re-estimate my baseline regression 132 times, each time omitting a single country (Van de Sijpe, 2012). If the new DAHG coefficient estimate (found by omitting a given country) is significantly different from baseline beta coefficient (which includes all 132 countries) then the country is considered to have undue influence and is omitted. This procedure suggests that ten countries (Afghanistan, Angola, Ethiopia, Haiti, India, Jordan, Mozambique, Malawi, Rwanda, and Timor-Leste) have undue influence. Column (10) shows that omitting these ten countries does not affect my DAHG coefficient estimate.. Figure S1 shows the histogram for the 132 unique beta estimates. This figure shows that the distribution of DAHG betas is centered on the red vertical line that represents my baseline DAHG beta.

Table S8 assesses displacement across geographic regions. These columns suggest that some evidence of displacement exists in Sub-Saharan Africa and South and East Asia. No evidence of displacement in Latin America is found in this estimation, but when a similar exercise is conducted using model (2), evidence of displacement in Latin America that is comparable to the other regions exists. Although the sample sizes for both Latin America and South/East Asia are small.

Table S9 shows that the DAHG effect is consistent across a host of differencing- GMM specifications. Column (1) is the baseline regression, while columns (2) through (9) each adjust one or two inputs to the estimation. Here I consider not using PCA instrument factorization, not collapsing my instruments, using more and less lags of GMM-style instruments, using one-step estimation, using forward orthogonal deviations (instead of first differencing), using backward orthogonal instrument deviations (rather than traditional levels), and not using the country-level weighting scheme. The results show that the DAHG effect is stable across all these adjustments, ranging from -0.426 to -0.991. In all but one of the sensitivity analyses included on Table S9 the displacement effect is statistically different from zero at the 95% confidence interval, with the majority being statistically significant at the 99% confidence level.

Figures S2 through S3 assess the effect of changing the number of components retained by the PCA factorization. Figure S2 utilizes the maximum number of available instruments, by including instruments starting at the second lag and not collapsing the instruments. Figure S2 shows that when between 15 than 120 instruments are retained, the Hansen J test and AR(2) test are passed, with p-values greater than 0.15. For these regressions, the DAHG coefficient is stable and is always statistically significant at the 99% confidence level. As the number of retained instruments increases beyond 120, the p-value of the Hansen J test increases. At face value, this suggests the retained components are even less likely to be endogenous, although it is far more likely a function of a weakening test statistical caused to too many instruments (Roodman, 2007). Still, all the 200 regressions show statistically significant displacement at the 99% confidence level.

Figures S3 illustrates a similar exercise, estimating my model using a diverse set of retained instruments. The only difference is that in Figure S3 the instruments are collapsed prior to the PCA analysis. In most cases the DAHG coefficient estimate is statistically different from zero at the 95% confidence level. Figures S3 supports Figure S2 in reaffirming that significant displacement occurs, while in all cases relevant to my baseline model both the AR(2) and Hansen J tests are easily passed.

Table S10 makes three different assumptions regarding the exogeniety of right-hand-side variables. Recall that in my baseline model, I assume that DAHG and the LDV are endogenous. In column (2), I assume that only the LDV is endogenous. This would be the case if DAHG was measured perfectly. Column (3) assumes that the LDV, DAHG, DAHNG, and GGE are endogenous. This could be the case if GHES and GGE are determined simultaneously and donors (giving to non-government health sector) consider GHES when allocating their aid. Column (4) adds GDP to the list of variables considered endogenous in column (3). GDP could be endogenous if it is a function of GHES. All three sensitivity tests suggest displacement, although column (4) shows that the model is very poorly fit with no variables being statistically significant at any conventional level.

Table S11 assesses several more variants of my baseline estimation, focusing on what variables are included in my model and instrument set. Column (2) measures all financial variables as a share of contemporaneous GDP, rather than the country-specific mean GDP. Column (3) drops the 2009 time indicators from the right-hand-side of my model, while column (4) drops GGE, the

only variable not statistically significant at the 95% level in the baseline regression. Column (5) shows the effect of dropping the three excluded instruments. Column (6) includes debt relief (DR) as a determinant of GHES, column (7) includes GDP growth, and column (8) includes GDP growth. Some have hypothesized that debt relief, HIV, and GDP growth might affect GHES, although this does not seem to be the case in my sample (Lu et al., 2010; Van de Sijpe, 2012). Column (9) includes a democracy index as reported by Freedom House and column (10) includes capital formation measured as a share of GDP as reported by the World Bank (The World Bank). These variables provide interesting contextual analyses but none affect the coefficient of interest. Because neither of these have a clear mechnisim to effect GHES I leave them out of my baseline model. Finally, column (11) collapses my sample into five 3-year averages. The advantage to time averaging would be to smooth over business cycles and measurement error. While the effect is measured with littler precision, the correct sign is still identified even with only 287 observations. Excluding column (11), these sensitivity tests show remarkably similar results when compared to the baseline DAHG effect. The DAHG effect in these estimates range from -0.421 to -0.903, and all but two are statistically significant at the 95% confidence level.

Finally, Table S12 shows that the estimation of model (1) is stable across a variety of alternative estimation methods. Columns (2) through (4) employ system-GMM, first-differencing, and fixed effects estimation, respectively. System-GMM marginally passes the Hansen test, but does not estimate statistically significant displacement. The other two show significant displacement (at the 99% confidence level) although they are both most likely biased because they do not account for the endogeneity of the LDV or DAHG.

4. Estimating model (2) baseline

The feature which distinguishes model (2) from model (1) is that in model (2) DAHG is parsed into two distinct variables: DAHG and DAHG$^-$, where DAHG$^-$ is the interaction between DAHG and a binary indicator marking a year-over-year decrease in DAHG. (DAHG$^-$ is thus zero for all country-years where DAHG is the same or larger than the previous year.) By modeling increases and decreases in DAHG separately, model (2) measures the displacement and replacement of GHES without assuming these two rates are equal. $\hat{\beta}_1$ is the estimated displacement rate (which follows an increase in DAHG), while $\left(\hat{\beta}_1 + \hat{\beta}_{1spread}\right)$ is the estimated replacement rate (which follows a decrease in DAHG).

$$(2)\quad GHES_{it} = \alpha_i + \beta_1 DAHG_{it} + \beta_{1\,spread} DAHG^-_{it} + \beta_2 DAHNG_{it} + \beta_3 GGE_{it} + \beta_4 GDP_{it} + \beta_5 War_{it} + \beta_6 Disaster_{it} + \beta_7 Pop_{it} + \nu_{it}$$

For consistency with model (1), the 2009 year indicator is included in this model. It is shown to be statistically significant, while all other year indicators are not. When the Wooldridge test is applied to model (2), and again rejects the null hypothesis of no autocorrelation (p = 0.000). This implies changes in GHES are correlated over time. As a consequence I again turn to Difference GMM. Unless explicitly discussed, the estimation parameters used for model (1) baseline are applied to the model (2) estimation.

Table S13 reports the baseline model (2) estimates, which are also reported in Table 2 of the main text. This estimation shows that a $1 increase in DAHG leads to $0.63 decrease in GHES,

while a $1 decrease in DAHG leads to a $0.28 increase in GHES. The Wald test shows that estimates for displacement and replacement are statistically different (p-value = 0.098). Tables S14 through S20 test the sensitivity of model (2). In the majority of these tables, the baseline estimates are repeated in column (1).

5. Estimating model (2) sensitivity analyses

Table S14 shows the effects of estimating model (2) in two alternative manners. Using ordinary least squares these methods would be equivalent, but because of the instrumentation involved in Difference GMM, these methods result in different estimates. All three models are estimated using the same specification and parameter inputs into Differencing GMM. The only thing that changes is the underlying specification. Column (1) follows from equation (2), while column (2) estimates equation (3) and column (3) estimates equation (4). For space the coefficient estimates for the covariates are suppressed and the coefficient estimates are combined for each model so that the displacement and replacement rates are given.

$$(3)\quad GHES_{it} = \alpha_i + \beta_1 DAHG_{it} + \beta_{1\,spread} DAHG_{it}^{+} + \beta_2 DAHNG_{it} + \beta_3 GGE_{it} + \beta_4 GDP_{it} + \beta_5 War_{it} + \beta_6 Disaster_{it} + \beta_7 Pop_{it} + \nu_{it}$$

$$(4)\quad GHES_{it} = \alpha_i + \beta_{1\,up} DAHG_{it}^{+} + \beta_{1\,down} DAHG_{it}^{-} + \beta_2 DAHNG_{it} + \beta_3 GGE_{it} + \beta_4 GDP_{it} + \beta_5 War_{it} + \beta_6 Disaster_{it} + \beta_7 Pop_{it} + \nu_{it}$$

In theory, these three models should estimate identical rates, but because of complexities inserted by instrumentation, obvious differences persist. I choose to prioritize equation (2) (column (1))

over the other two because a) ordinary least squares estimation shows that first stage of the instrumentation is better for equation (2) and (3), relative to (4), b) equation (2) and (3) give more similar results, relative to (4), and c) equation (2) reflects equation (4) in that it shows that the displacement and replacement rates are statistically different.

Table S15 tests my results across subsamples that remove various outliers. In the eight additional analyses, the displacement rate varies between -0.43 and -1.06, and remains statistically significant at the 99% confidence level for seven of the eight estimates. For these same estimates, the replacement rate is never statistically significant at even the 90% confidence level. The Wald test varies more though. In three of the eight additional regressions the displacement rate and replacement rate are deemed statistically different at the 90% confidence level, while the other five p-values range from 0.13 to 0.643.

Table S16 tests model (2) across different geographic subsamples. In all three subsample analyses the displacement is larger (in absolute terms) than the replacement rate, although the two rates are only statistically different in Sub-Saharan Africa and South/East Asia (p-values = 0.067 for both subsamples).

Table S17 adjusts the input parameters for Difference GMM, in the same manner as table S9. Here, the displacement rate is statistically different and larger than the replacement rate (in absolute terms) for seven of the eight additional tests (at the 90% confidence level).

Table S18 adjusts my assumptions about which variables are endogenous. These sensitivity analyses support the baseline conclusions. The displacement rate varies between -0.62 and -0.81, and is always statistically significant. The replacement rates varies between 0.03 and -0.68 and is statistically different from the displacement rate in two of the three sensitivity analyses.

Table S19 adjusts the specification of model (2), including and excluding a variety of variables and instruments. Across the first nine of the sensitivity analyses (columns (2) through (10), the displacement rate varies from -0.53 and -0.758, and is statistically significant at the 90% confidence level in eight of the variants. The replacement rate is statistically different (at the 90% confidence level) and smaller (in absolute terms) than the displacement rate for six of the nine regressions. Column (11) collapses my data into five 3-year averages. This drastically reduces my sample size, but illustrates a relatively similar set of qualitative conclusions. The displacement rate is larger than the replacement rate, although the two only differ with p-value 0.319.

Finally, Table 20 applies different estimation methods to model (2). In this case all three variants show that the displacement rate is statistically different from zero and statistically different from the replacement rate (both at the 90% confidence level). While column (2), which applies system GMM should be consider valuable confirmation, some skepticism should be applied to the first-differenced and fixed-effects estimation because both are biased in finite samples.

6. Simulations

Two simulations are used to generate Figures 1 through 4 in the main text. Each simulation completes 1000 random draws from the variance-covariance matrix from each of baseline models. Figure 1 is based on the baseline estimation of model (1). In this figure the fitted GHES values are the sum of the products found by multiplying the randomly drawn coefficients by the mean of the independent variables (where the mean is across the 132 countries). Only DAHG, which distinguishes the two scenarios in Figure 1, is not based on the time-specific averages. For the blue-line scenario (constant DAHG), DAHG is the mean across the entire sample (mean across the 132 countries and 16 years, roughly $6 per capita) for each year. DAHG for the red-line scenario (spiked DAHG) is the same, except in 1997 I include a 50% increase in DAHG (to roughly $9 per capita). A 50% increase is roughly the median increase of DAHG year-over-year across my sample.

Figures 2 through 4 are based on the baseline estimation of model (2). For these figures, country-year level data enters the simulation. Each country is assigned its own unique set of beta coefficients drawn from the variance-covariance matrix, with 1000 rounds of drawing. Four distinct scenarios are used to create these figures. The first scenario estimates the actual fitted values of GHES using the random draws for DAHG and $DAHG^{-}$ based on the varianc-covariance matrix and the observed values of DAHG and $DAHG^{-}$. The second scenario uses these same estimated coefficients, but DAHG is structured so that it is monotonically increasing. More specifically, DAHG is set such that at in 1995 the smallest amount of aid is received (at the country-level), with DAHG increasing so that the actual amount of DAHG received over the

course of the 16 years is modeled. A key feature to this scenario is that $DAHG^-$ is always zero because DAHG is increasing monotonically. The third scenario set the DAHG and $DAHG^-$ coefficients to zero, while the fourth scenario sets both coefficients to the expected rate of displacement, where the expected rate is $1-\frac{GHEA}{GGE}$. After GHES is fitted using all four scenarios, I add accompanying DAHG to estimate GHE.

For Figures 2 and 3, I collapse these GHE estimates summing across countries to estimate a global total for the 132 countries for each year. The costs illustrated in Figure 2 are found by subtracting the GHE from the first scenario for the GHE from the third scenario. The costs illustrated in Figure 3 are found by subtracting the GHE from the first scenario from the GHE from the second scenario. For Figure 4, I collapse the GHE estimates summing across countries and time. The third policy is based on costs derived from subtracting GHE from the first scenario from GHE from the fourth scenario.

Tables and Figures

Table S1: Countries include in analysis

Afghanistan	Dominica	Macedonia, Former Yugoslav Rep of	Sao Tome and Principe
Albania	Dominican Republic	Madagascar	Senegal
Algeria	Ecuador	Malawi	Serbia
Angola	Egypt	Malaysia	Seychelles
Argentina	El Salvador	Maldives	Sierra Leone
Armenia	Eritrea	Mali	Solomon Islands
Azerbaijan	Ethiopia	Marshall Islands	South Africa
Bangladesh	Fiji	Mauritania	Sri Lanka
Belarus	Gabon	Mauritius	Sudan
Belize	Gambia	Mexico	Suriname
Benin	Georgia	Micronesia, Federated States of	Swaziland
Bhutan	Ghana	Moldova	Syrian Arab Republic
Bolivia	Grenada	Mongolia	Tajikistan
Bosnia and Herzegovina	Guatemala	Montenegro	Tanzania, United Rep of
Botswana	Guinea	Morocco	Thailand
Brazil	Guinea-Bissau	Mozambique	Timor-Leste
Bulgaria	Guyana	Myanmar	Togo
Burkina Faso	Haiti	Namibia	Tonga
Burundi	Honduras	Nepal	Tunisia
Cambodia	India	Nicaragua	Turkey
Cameroon	Indonesia	Niger	Turkmenistan
Cape Verde	Iran, Islamic Republic of	Nigeria	Uganda
Central African Republic	Iraq	Pakistan	Ukraine
Chad	Jamaica	Panama	Uruguay
Chile	Jordan	Papua New Guinea	Uzbekistan
China	Kazakhstan	Paraguay	Vanuatu
Colombia	Kenya	Peru	Venezuela
Comoros	Kiribati	Philippines	Viet Nam
Congo	Kyrgyzstan	Romania	Yemen
Congo, the DemRep of the	Lao People's Dem Rep	Russian Federation	Zambia
Costa Rica	Lebanon	Rwanda	
Cuba	Lesotho	Saint Lucia	
Côte d'Ivoire	Liberia	Saint Vincent and the Grenadines	
Djibouti	Libyan Arab Jamahiriya	Samoa	

Table S2: Data, units, and sources

Abbreviation	Variables	Units	Source
GHES	Government health expenditure as source	As a share of country's mean GDP	Calculated by authors
GHEA	Government health expenditure as agent	As a share of country's mean GDP	World Health Organization
DAHG	Development assistance for health channeled to government	As a share of country's mean GDP	Institute for Health Metrics and Evaluation
DAHNG	Development assistance for health channeled to non-governmental sector	As a share of country's mean GDP	Institute for Health Metrics and Evaluation
GGE	General government expenditure	As a share of country's mean GDP	World Health Organization
POP	Population	Prevalence	Institute for Health Metrics and Evaluation
GDP	Gross domestic product	Log Per capita (2010 US or international dollars)	James et al.
Disaster	Deaths attributable to conflict	Deaths per 10,000	EM-DAT Database
War	Deaths attributable to nautral disasters	Deaths per 10,000	Uppsala Conflict Data Program

Table S3: Descriptive statistics of key variables

Variable	Observation	Mean	Std. Dev.	Min	Max
GHEA / GDP	2070	0.0386	0.0264	0.0014	0.1865
GHES / GDP	2070	0.0349	0.0250	-0.0181	0.1861
DAHG / GDP	2070	0.0037	0.0083	-0.0001	0.0991
DAHNG / GDP	2070	0.0021	0.0068	0.0000	0.1052
GGE / GDP	2070	0.3542	0.2256	0.0419	2.3423
GDP (Log per capita)	2070	7.2558	1.0755	4.8067	9.5815
Disaster (Deaths per 10,000)	2070	0.0000	0.0001	0.0000	0.0022
War (Deaths per 10,000)	2070	0.0024	0.0234	0.0000	0.8566
Pop (Log)	2070	15.6633	2.0289	10.8404	21.0174
Year	2070	2003	5	1995	2010

Table S4: Correlation matrix of key variables

	GHEA	GHES	DAHG	DAHNG	GGE	GDP	HIV	Disaster	War	Year
GHEA	1.0000									
GHES	0.9498	1.0000								
DAHG	0.3236	0.0113	1.0000							
DAHNG	0.1958	0.0622	0.4376	1.0000						
GGE	0.6003	0.5804	0.1634	0.1579	1.0000					
GDP	0.2879	0.4064	-0.3090	-0.2862	0.1518	1.0000				
Disaster	-0.0267	-0.0270	-0.0037	0.0789	0.0039	-0.0308	1.0000			
War	-0.0299	-0.0367	0.0155	-0.0082	0.1686	-0.1028	-0.0032	1.0000		
Pop	-0.3156	-0.2755	-0.1753	-0.0020	-0.1085	-0.2125	0.0044	0.0144	1.0000	
Year	0.3953	0.3431	0.2260	0.3494	0.3845	0.1033	0.0319	-0.0471	0.0358	1.0000

Table S5: Baseline model (1) results

	(1)
VARIABLES	GHES / GDP
Lag GHES / GDP	0.530***
	(0.162)
DAHG / GDP	-0.653**
	(0.271)
DAHNG / GDP	0.388***
	(0.118)
GGE / GDP	0.024**
	(0.010)
Log GDP per capita	0.012**
	(0.006)
War deaths	-0.024***
	(0.009)
Disaster deaths	2.662**
	(1.064)
Log Population	-0.001
	(0.016)
Year 2009	0.003***
	(0.001)
Observations	1,804
Number of countries	132
Hansen p-value	0.488
AR(2) p-value	0.200
Instrument count	14
Robust standard errors in parentheses	
*** p<0.01, ** p<0.05, * p<0.1	

Table S6: Pre-tests to evaluate estimation method; model (1)

VARIABLES	(1)	(2)	(3)
	Dependent Variable: GHES / GDP		
Lag GHES / GDP	0.530***	0.890***	0.485*
	-(0.162)	-(0.059)	-(0.275)
DAHG / GDP	-0.653**	-0.205*	-1.028*
	-(0.271)	-(0.113)	-(0.572)
DAHNG / GDP	0.388***	0.216***	0.380**
	-(0.118)	-(0.080)	-(0.176)
GGE / GDP	0.024**	0.011**	0.028***
	-(0.010)	-(0.005)	-(0.009)
Log GDP per capita	0.012**	0.000	0.01
	-(0.006)	-(0.001)	-(0.007)
War deaths	-0.024***	-0.023**	-0.024**
	-(0.009)	-(0.011)	-(0.010)
Disaster deaths	2.662**	-0.816	2.063
	-(1.064)	-(1.118)	-(1.883)
Log Population	-0.001	-0.001	0.008
	-(0.016)	(0.000)	-(0.026)
Year 1996			-0.001
			-(0.007)
Year 1997			-0.001
			-(0.007)
Year 1998			-0.001
			-(0.007)
Year 1999			-0.002
			-(0.007)
Year 2000			-0.001
			-(0.007)
Year 2001			-0.001
			-(0.006)
Year 2002			-0.001
			-(0.006)
Year 2003			0.001
			-(0.006)
Year 2004			0.001
			-(0.006)
Year 2005			0.001
			-(0.005)
Year 2006			0.002
			-(0.004)
Year 2007			0.002
			-(0.003)
Year 2008			0.001
			-(0.003)
Year 2009	0.003***	0.002**	0.003
	-(0.001)	-(0.001)	-(0.002)
Constant		0.007	
		-(0.009)	
Estimation method	Diff GMM	Sys GMM	Diff GMM
Observations	1,804	1,937	1,804
Number of countries	132	133	132
Hansen p-value	0.488	0.205	0.463
AR(2) p-value	0.200	0.104	0.359
Difference-in-Hansen		0.341	
Instrument count	14	41	27

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table S7: Sensitivity excluding outliers; model (1)

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Dependent variable: GHES / GDP									
Lag GHES / GDP	0.530***	0.175	0.568***	0.638***	0.458***	0.478***	0.360**	0.630***	0.369***	0.570***
	-(0.162)	-(0.174)	-(0.122)	-(0.111)	-(0.162)	-(0.113)	-(0.173)	-(0.173)	-(0.131)	-(0.170)
DAHG / GDP	-0.653**	-0.728*	-1.101***	-0.568***	-0.794***	-0.488*	-0.818***	-0.657***	-0.883**	-0.616**
	-(0.271)	-(0.411)	-(0.158)	-(0.157)	-(0.227)	-(0.293)	-(0.309)	-(0.229)	-(0.422)	-(0.247)
DAHNG / GDP	0.388***	0.353***	0.111	0.104	0.382***	0.362**	0.322**	0.373***	0.301**	0.282*
	-(0.118)	-(0.117)	-(0.242)	-(0.368)	-(0.115)	-(0.145)	-(0.139)	-(0.116)	-(0.129)	-(0.151)
GGE / GDP	0.024**	0.015**	0.032**	0.028**	0.021***	0.029**	0.035***	0.026**	0.023**	0.020
	-(0.010)	-(0.007)	-(0.014)	-(0.014)	-(0.008)	-(0.011)	-(0.011)	-(0.012)	-(0.010)	-(0.013)
Log GDP per capita	0.012**	0.019***	0.013*	0.012**	0.016***	0.01	0.013**	0.009	0.015***	0.016***
	-(0.006)	-(0.005)	-(0.007)	-(0.005)	-(0.005)	-(0.007)	-(0.005)	-(0.005)	-(0.005)	-(0.006)
War deaths	-0.024***	-0.017***	-0.017**	-0.026**	-0.004	-0.032***	-0.025***	-0.029***	-0.022**	-0.024**
	-(0.009)	-(0.004)	-(0.008)	-(0.011)	-(0.005)	-(0.011)	-(0.008)	-(0.011)	-(0.009)	-(0.010)
Disaster deaths	2.662**	2.409**	0.229	6.794	2.244**	1.829	1.523	9.368	2.097*	8.716
	-(1.064)	-(0.971)	-(1.478)	-(10.383)	-(1.013)	-(2.068)	-(1.247)	-(9.062)	-(1.185)	-(10.689)
Log Population	-0.001	0.019	0.004	-0.004	0.011	-0.005	0.008	-0.008	0.018	-0.002
	-(0.016)	-(0.019)	-(0.005)	-(0.007)	-(0.017)	-(0.013)	-(0.023)	-(0.016)	-(0.019)	-(0.010)
Year 2009	0.003***	0.002***	0.002**	0.003*	0.002***	0.003***	0.003***	0.004***	0.003***	0.001***
	-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)
Excluding:	None	based on GHE-S	based on DAHG	based on DAHNG	based on GGE	based on GDP	based on war	based on disaster	based on pop	10 countries
Observations	1,804	1,623	1,623	1,623	1,623	1,623	1,623	1,623	1,623	1,676
Number of countries	132	128	132	132	128	125	129	132	120	122
Hansen p-value	0.488	0.528	0.136	0.226	0.805	0.581	0.595	0.511	0.588	0.261
AR(2) p-value	0.2	0.0964	0.979	0.836	0.0672	0.361	0.042	0.301	0.308	0.684
Instrument count	14	14	15	15	14	15	14	14	14	15

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table S8: Sensitivity across geographic subsamples; model (1)

	(1)	(2)	(3)	(4)
VARIABLES	Dependent variable: GHES / GDP			
Lag GHES / GDP	0.530***	0.081	0.699***	0.681***
	-(0.162)	-(0.191)	-(0.176)	-(0.177)
DAHG / GDP	-0.653**	-1.084***	-0.755***	0.581
	-(0.271)	-(0.265)	-(0.225)	-(1.375)
DAHNG / GDP	0.388***	0.255**	0.292**	0.073
	-(0.118)	-(0.125)	-(0.137)	-(0.330)
GGE / GDP	0.024**	0.028***	0.018	-0.016
	-(0.010)	-(0.010)	-(0.012)	-(0.010)
Log GDP per capita	0.012**	0.016**	0.01	0.046***
	-(0.006)	-(0.007)	-(0.006)	-(0.009)
War deaths	-0.024***	-0.023***	-0.091	-0.149
	-(0.009)	-(0.006)	-(0.099)	-(0.735)
Disaster deaths	2.662**	-122.121*	-2.181	3.856*
	-(1.064)	-(73.352)	-(10.344)	-(2.176)
Log Population	-0.001	0.045**	-0.021	0.004
	-(0.016)	-(0.021)	-(0.018)	-(0.017)
Year 2009	0.003***	0.002	0.002**	0.002
	-(0.001)	-(0.001)	-(0.001)	-(0.001)
Geographic subsamples:	All	Sub-Saharn Africa	South and East Asia	Latin America
Observations	1,804	596	392	350
Number of countries	132	43	29	25
Hansen p-value	0.488	0.713	0.269	0.548
AR(2) p-value	0.2	0.148	0.756	0.615
Instrument count	14	14	15	14

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table S9: Sensitivity across Difference GMM specifications; model (1)

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Dependent variable: GHES / GDP								
Lag GHES / GDP	0.530***	0.504***	0.560***	0.521***	0.672***	0.490***	0.689***	0.427**	0.572***
	-(0.162)	-(0.101)	-(0.087)	-(0.160)	-(0.173)	-(0.166)	-(0.186)	-(0.204)	-(0.108)
DAHG / GDP	-0.653**	-0.856***	-0.890***	-0.769***	-0.991**	-0.694***	-0.426*	-0.807***	-0.684***
	-(0.271)	-(0.158)	-(0.154)	-(0.251)	-(0.390)	-(0.197)	-(0.219)	-(0.231)	-(0.226)
DAHNG / GDP	0.388***	0.325***	0.253***	0.326***	0.373***	0.347***	0.225**	0.320**	0.319***
	-(0.118)	-(0.101)	-(0.082)	-(0.123)	-(0.133)	-(0.116)	-(0.107)	-(0.161)	-(0.082)
GGE / GDP	0.024**	0.018**	0.030***	0.026**	0.032***	0.030***	0.016*	0.033**	0.018***
	-(0.010)	-(0.009)	-(0.011)	-(0.010)	-(0.012)	-(0.011)	-(0.010)	-(0.013)	-(0.006)
Log GDP per capita	0.012**	0.014***	0.01	0.012**	0.007	0.011**	0.008*	0.008	0.012***
	-(0.006)	-(0.005)	-(0.007)	-(0.005)	-(0.006)	-(0.006)	-(0.004)	-(0.006)	-(0.004)
War deaths	-0.024***	-0.020***	-0.030***	-0.027***	-0.027***	-0.029***	-0.031*	-0.046*	-0.022***
	-(0.009)	-(0.007)	-(0.010)	-(0.009)	-(0.009)	-(0.010)	-(0.019)	-(0.026)	-(0.005)
Disaster deaths	2.662**	-1.736	1.883	2.263**	2.028*	1.782	-2.997*	-7.522***	1.868
	-(1.064)	-(4.632)	-(1.163)	-(1.133)	-(1.205)	-(1.114)	-(1.779)	-(2.644)	-(2.258)
Log Population	-0.001	0.009	0.016	0.004	0	0.004	0	0.009	0.009
	-(0.016)	-(0.010)	-(0.010)	-(0.016)	-(0.016)	-(0.015)	-(0.007)	-(0.013)	-(0.011)
Year 2009	0.003***	0.002***	0.003***	0.003***	0.003***	0.003***	0.003***	0.003**	0.003***
	-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)
PCA factorization of instruments	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Collapse instruments	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Instrument lag limits	(1 15)	(1 15)	(1 15)	(2 15)	(1 3)	(1 15)	(1 15)	(1 15)	(1 15)
Two-step method	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Forward orthogonal transformation	No	No	No	No	No	No	Yes	Yes	No
Backward orthogonal deviated instruments	No	No	No	No	No	No	No	Yes	No
Weighted observations	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Observations	1,804	1,804	1,804	1,804	1,804	1,804	1,804	1,804	1,804
Number of countries	132	132	132	132	132	132	132	132	132
Hansen p-value	0.488	0.788	0.278	0.457	0.853	0.488	0.253	0.267	0.824
AR(2) p-value	0.2	0.223	0.18	0.22	0.183	0.193	0.123	0.282	0.102
Instrument Count	14	38	37	14	12	14	14	17	15

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table S10: Sensitivity across instrument specifications; model (1)

VARIABLES	(1)	(2)	(3)	(4)
	Dependent variable: GHES / GDP			
Lag GHES / GDP	0.530***	0.434***	0.831***	0.402
	-(0.162)	-(0.122)	-(0.144)	-(0.487)
DAHG / GDP	-0.653**	-0.908***	-0.849***	-0.704
	-(0.271)	-(0.088)	-(0.192)	-(0.461)
DAHNG / GDP	0.388***	0.300***	0.239**	0.291
	-(0.118)	-(0.086)	-(0.122)	-(0.184)
GGE / GDP	0.024**	0.024**	-0.014	0.02
	-(0.010)	-(0.010)	-(0.018)	-(0.063)
Log GDP per capita	0.012**	0.015***	0.020***	0.023
	-(0.006)	-(0.005)	-(0.006)	-(0.017)
War deaths	-0.024***	-0.024***	-0.004	-0.025
	-(0.009)	-(0.009)	-(0.009)	-(0.040)
Disaster deaths	2.662**	2.079*	4.593***	2.221
	-(1.064)	-(1.073)	-(1.733)	-(4.546)
Log Population	-0.001	0.014	0.017	0.004
	-(0.016)	-(0.009)	-(0.011)	-(0.022)
Year 2009	0.003***	0.003***	0.003***	0.002
	-(0.001)	-(0.001)	-(0.001)	-(0.001)
Endogenous variables	LDV, DAH-G	LDV	LDV, DAH-G, DAH-NG, GGE	LDV, DAH-G, DAH-NG, GGE, GDP
Observations	1,804	1,804	1,804	1,804
Number of countries	132	132	132	132
Hansen p-value	0.488	0.476	0.609	0.632
AR(2) p-value	0.2	0.269	0.156	0.351
Instrument count	14	15	16	18

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table S11: Sensitivity across variable included; model (1)

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
						Dependent variable: GHES / GDP					
Lag GHES / GDP	0.530***	0.298	0.608***	0.618***	0.530***	0.535***	0.521***	0.501***	0.458	0.507***	1.532***
	-(0.162)	-(0.272)	-(0.153)	-(0.138)	-(0.162)	-(0.162)	-(0.151)	-(0.190)	-(0.316)	-(0.150)	-(0.539)
DAHG / GDP	-0.653**	-0.903***	-0.421*	-0.545**	-0.653**	-0.647**	-0.638**	-0.685**	-0.782**	-0.767	-0.186
	-(0.271)	-(0.269)	-(0.227)	-(0.235)	-(0.271)	-(0.273)	-(0.261)	-(0.283)	-(0.307)	-(0.526)	-(0.686)
DAHNG / GDP	0.388***	0.216	0.455***	0.451***	0.388***	0.396***	0.412***	0.378***	0.476*	0.311	-0.137
	-(0.118)	-(0.208)	-(0.110)	-(0.116)	-(0.118)	-(0.118)	-(0.120)	-(0.121)	-(0.250)	-(0.192)	-(0.309)
GGE / GDP	0.024**	0.022**	0.029***		0.024**	0.023**	0.022*	0.023**	0.014	0.025**	-0.001
	-(0.010)	-(0.009)	-(0.010)		-(0.010)	-(0.011)	-(0.011)	-(0.010)	-(0.018)	-(0.011)	-(0.023)
Log GDP per capita	0.012**	-0.001	0.006	0.016***	0.012**	0.012**	0.012**	0.015**	0.017***	0.010*	0.017***
	-(0.006)	-(0.002)	-(0.006)	-(0.004)	-(0.006)	-(0.006)	-(0.005)	-(0.007)	-(0.005)	-(0.006)	-(0.006)
War deaths	-0.024***	-0.013***	-0.029***	-0.010*	-0.024***	-0.024***	-0.022**	-0.024***	-0.017	-0.025***	-0.001
	-(0.009)	-(0.004)	-(0.009)	-(0.006)	-(0.009)	-(0.009)	-(0.010)	-(0.009)	-(0.015)	-(0.007)	-(0.030)
Disaster deaths	2.662**	1.116	1.162	4.583***	2.662**	2.652**	2.851***	2.488**	6.069	2.382	4.047
	-(1.064)	-(1.306)	-(1.028)	-(0.743)	-(1.064)	-(1.093)	-(1.102)	-(1.149)	-(10.291)	-(1.461)	-(8.501)
Log Population	-0.001	0.009	-0.012	0.002	-0.001	-0.002	0.000	0.000	0.008	0.005	-0.022
	-(0.016)	-(0.011)	-(0.013)	-(0.013)	-(0.016)	-(0.015)	-(0.015)	-(0.017)	-(0.021)	-(0.020)	-(0.026)
Year 2009	0.003***	0.001***		0.003***	0.003***	0.003***	0.003***	0.003***	0.002**	0.003***	
	-(0.001)	(0.000)		-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)	
Extra covariate						0.013	0	-0.005	0.000*	0.000**	
						-(0.042)	(0.000)	-(0.009)	(0.000)	(0.000)	
Notes:	Standard set	Normalize with current GDP	Drop time indicator	Drop generally insignificant covariate	Drop excluded instruments	Add debt relief	Add HIV prevalence	Add GDP growth	Add democracy index	Add capita formation	Collapse by 3-yr time averages
Observations	1,804	1,804	1,804	1,804	1,804	1,784	1,804	1,804	1,660	1,633	387
Number of countries	132	132	132	132	132	132	132	132	132	124	131
Hansen p-value	0.488	0.718	0.639	0.332	0.488	0.491	0.451	0.49	0.071	0.13	0.595
AR(2) p-value	0.200	0.090	0.166	0.177	0.200	0.197	0.202	0.207	0.651	0.0453	0.499
Instrument count	14	15	13	12	14	15	14	15	15	15	15

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table S12: Sensitivity across estimation methods; model (1)

VARIABLES	(1)	(2)	(3)	(4)
	Dependent variable: GHES / GDP			
Lag GHE-S / GDP	0.530***	0.858***	0.032	
	-(0.162)	-(0.096)	-(0.063)	
DAH-G / GDP	-0.653**	-0.213	-0.957***	-0.443***
	-(0.271)	-(0.156)	-(0.087)	-(0.097)
DAH-NG / GDP	0.388***	0.266***	0.148	0.144
	-(0.118)	-(0.100)	-(0.109)	-(0.122)
GGE / GDP	0.024**	0.01	0.027**	0.034***
	-(0.010)	-(0.009)	-(0.011)	-(0.012)
Log GDP per capita	0.012**	0.001	0.015***	0.015**
	-(0.006)	-(0.001)	-(0.005)	-(0.006)
HIV prevalence	-0.024***	-0.015	0	-0.057
	-(0.009)	-(0.011)	-(0.001)	-(0.037)
War deaths	2.662**	-0.534	-0.027**	-6.834***
	-(1.064)	-(1.243)	-(0.011)	-(2.022)
Disaster deaths	-0.001	-0.001	-0.296	0.029***
	-(0.016)	-(0.001)	-(0.852)	-(0.010)
Year 2009	0.003***	0.003***	0.003***	0.003***
	-(0.001)	-(0.001)	-(0.001)	-(0.001)
Constant		0.006	0.027	-0.583***
		-(0.010)	-(0.019)	-(0.161)
Estimation method	Differencing-GMM	System-GMM	First differencing	Fixed effects
Observations	1,804	1,937	1,804	2,070
R-squared			0.352	0.516
Number of countries	132	133		133
Hansen p-value	0.488	0.21	.	.
AR(2) p-value	0.2	0.1	.	.
Instrument count	14	16	.	.
Robust standard errors in parentheses				
*** p<0.01, ** p<0.05, * p<0.1				

Table S13: Baseline model (2) results

VARIABLES	(1) GHES / GDP
Lag GHES / GDP	0.540***
	-(0.161)
DAHG / GDP	-0.623**
	-(0.270)
$DAHG^-$ / GDP	0.347*
	-(0.210)
DAHNG / GDP	0.219
	-(0.154)
GGE / GDP	0.025**
	-(0.012)
Log GDP per capita	0.011**
	-(0.005)
War deaths	-0.026***
	-(0.010)
Disaster deaths	0.224
	-(5.355)
Log Population	0.001
	-(0.013)
Year 2009	0.002***
	-(0.001)
Observations	1,804
Number of countries	132
DAHG + $DAHG^-$	-0.276
se(DAHG + $DAHG^-$)	(0.350)
Test up = down p-value	0.098
Hansen p-value	0.531
AR(2) p-value	0.533
Instrument count	20
Robust standard errors in parentheses	
*** p<0.01, ** p<0.05, * p<0.1	

Table S14: Alternative models; model (2)

	(1)	(2)	(3)
VARIABLES	Dependent variable: GHES / GDP		
$DAHG^{+}$ / GDP	-0.623**	-0.667***	-0.484**
	(0.270)	(0.259)	-(0.246)
$DAHG^{-}$ / GDP	-0.276	-0.494	-0.056
	(0.350)	(0.330)	(0.356)
Test up = down p-value	0.098	0.457	0.016
Hansen p-value	0.531	0.525	0.864
AR(2) p-value	0.533	0.361	0.54
Instrument count	20	17	21
All other coeficient estimates surpressed.			
Robust standard errors in parentheses			
*** p<0.01, ** p<0.05, * p<0.1			

Table S15: Omitting outliers; model (2)

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Dependent variable: GHES / GDP								
Lag GHES / GDP	0.553***	0.344**	0.541***	0.644***	0.718***	0.479***	0.500***	0.598***	0.487***
	(0.082)	(0.166)	(0.100)	(0.112)	(0.105)	(0.099)	(0.091)	(0.151)	(0.096)
DAHG / GDP	-0.941***	-0.425***	-0.947***	-0.912***	-0.974**	-1.063***	-0.961***	-0.887***	-0.560*
	(0.169)	(0.120)	(0.320)	(0.151)	(0.437)	(0.137)	(0.149)	(0.161)	(0.331)
$DAHG^-$ / GDP	0.463*	0.318	0.638	0.661**	0.401	0.545	0.573**	0.673**	0.702
	(0.275)	(0.265)	(1.375)	(0.293)	(0.516)	(0.364)	(0.287)	(0.341)	(0.814)
DAHNG / GDP	0.145	0.174**	0.114	-0.487	0.339**	0.110	0.009	0.125	0.128
	(0.106)	(0.086)	(0.216)	(0.489)	(0.155)	(0.129)	(0.101)	(0.110)	(0.145)
GGE / GDP	0.019**	0.010**	0.030***	0.020**	0.010	0.019*	0.011**	0.015	0.025***
	(0.008)	(0.005)	(0.008)	(0.010)	(0.006)	(0.010)	(0.005)	(0.010)	(0.009)
Log GDP per capita	0.015***	0.017***	0.013**	0.013***	0.014***	0.016***	0.017***	0.014***	0.012***
	(0.005)	(0.003)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
War deaths	-0.019***	-0.013***	-0.018**	-0.021***	-0.010	-0.022***	-0.002	-0.021**	-0.023***
	(0.006)	(0.003)	(0.008)	(0.008)	(0.007)	(0.008)	(0.010)	(0.009)	(0.008)
Disaster deaths	-0.667	1.509	-2.003	9.005	0.035	-0.775	-0.872	-13.074	-0.581
	(3.906)	(1.400)	(4.924)	(10.649)	(4.050)	(3.708)	(4.243)	(18.887)	(3.272)
Log Population	0.016***	0.014	0.009	0.015*	0.004	0.021**	0.021**	0.009	0.008
	(0.006)	(0.010)	(0.009)	(0.008)	(0.010)	(0.010)	(0.009)	(0.013)	(0.010)
Year 2009	0.002***	0.002***	0.002*	0.002*	0.002**	0.003***	0.002***	0.003***	0.003**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	1,804	1,623	1,623	1,623	1,623	1,623	1,623	1,623	1,623
Number of countries	132	128	132	132	128	125	129	132	120
Outliers excluded	None	GHES	DAHG	DAHNG	GGE	GDP	WAR	Disaster	Pop
DAHG + $DAHG^-$	-0.478	-0.107	-0.310	-0.250	-0.572	-0.518	-0.387	-0.214	0.142
se(DAHG + $DAHG^-$)	0.341	0.304	1.461	0.364	0.650	0.421	0.366	0.440	1.043
Test up = down p-value	0.0916	0.229	0.643	0.0239	0.437	0.134	0.0462	0.0487	0.388
Hansen p-value	0.586	0.785	0.296	0.338	0.148	0.935	0.897	0.664	0.549
AR(2) p-value	0.0538	0.439	0.214	0.0697	0.376	0.0440	0.133	0.0209	0.161
Instrument count	33	20	18	19	20	20	20	20	19

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table S16: Geographic subsamples; model (2)

VARIABLES	(1)	(2)	(3)	(4)
	Dependent variable: GHES / GDP			
Lag GHES / GDP	0.540***	0.470***	0.646***	0.628*
	-(0.161)	-(0.129)	-(0.114)	-(0.340)
DAHG / GDP	-0.623**	-0.751**	-0.972***	-0.658
	-(0.270)	-(0.315)	-(0.147)	-(0.925)
$DAHG^{-}$ / GDP	0.347*	0.336*	0.343*	0.216
	-(0.210)	-(0.183)	-(0.187)	-(0.864)
DAHNG / GDP	0.219	0.264	0.188	0.099
	-(0.154)	-(0.193)	-(0.237)	-(0.420)
GGE / GDP	0.025**	0.016	0.019*	-0.004
	-(0.012)	-(0.010)	-(0.011)	-(0.019)
Log GDP per capita	0.011**	0.017**	0.014**	0.044***
	-(0.005)	-(0.008)	-(0.006)	-(0.015)
War deaths	-0.026***	-0.019***	-0.09	-0.196
	-(0.010)	-(0.007)	-(0.075)	-(1.162)
Disaster deaths	0.224	-122.117	1.416	3.269
	-(5.355)	-(102.007)	-(11.807)	-(2.035)
Log Population	0.001	0.011	-0.023*	0.001
	-(0.013)	-(0.013)	-(0.012)	-(0.020)
Year 2009	0.002***	0.001	0.003***	0.002
	-(0.001)	-(0.002)	-(0.001)	-(0.002)
Observations	1,804	596	392	350
Number of countries	132	43	29	25
Geographic region	All	Sub-Saharan Africa	South and East Asia	Latin America / Caribbean
DAHG + $DAHG^{-}$	-0.276	-0.415	-0.629	-0.441
se(DAHG + $DAHG^{-}$)	0.35	0.392	0.251	1.491
Test up = down p-value	0.0981	0.0669	0.0667	0.802
Hansen p-value	0.531	0.427	0.36	0.14
AR(2) p-value	0.533	0.0662	0.488	0.75
Instrument count	20	19	20	19

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table S17: Sensitivity across Difference GMM specifications;; model (2)

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Dependent variable: GHES / GDP								
Lag GHES / GDP	0.540***	0.508***	0.580***	0.582***	0.751***	0.492***	0.899***	0.657***	0.478***
	-(0.161)	-(0.080)	-(0.061)	-(0.147)	-(0.223)	-(0.151)	-(0.103)	-(0.239)	-(0.112)
DAHG / GDP	-0.623**	-0.849***	-0.660***	-0.598**	-0.959**	-0.771***	-0.109	-0.631**	-1.014***
	-(0.270)	-(0.132)	-(0.093)	-(0.289)	-(0.380)	-(0.174)	-(0.157)	-(0.281)	-(0.177)
DAHG$^-$ / GDP	0.347*	0.281**	0.357***	0.411**	0.131	0.317**	0.702***	0.607**	0.590*
	-(0.210)	-(0.111)	-(0.122)	-(0.206)	-(0.390)	-(0.127)	-(0.163)	-(0.302)	-(0.327)
DAHNG / GDP	0.219	0.211*	0.125	0.224	0.361**	0.213	-0.055	0.071	0.13
	-(0.154)	-(0.122)	-(0.102)	-(0.153)	-(0.170)	-(0.133)	-(0.078)	-(0.222)	-(0.105)
GGE / GDP	0.025**	0.024**	0.025*	0.025*	0.031**	0.027**	0.01	0.014	0.014*
	-(0.012)	-(0.011)	-(0.013)	-(0.013)	-(0.013)	-(0.012)	-(0.008)	-(0.011)	-(0.008)
Log GDP per capita	0.011**	0.012***	0.012**	0.010*	0.006	0.013**	0.004	0.008	0.015***
	-(0.005)	-(0.004)	-(0.006)	-(0.005)	-(0.006)	-(0.006)	-(0.003)	-(0.005)	-(0.004)
War deaths	-0.026***	-0.023**	-0.026***	-0.025**	-0.027***	-0.027***	-0.021	-0.021	-0.017**
	-(0.010)	-(0.009)	-(0.010)	-(0.011)	-(0.010)	-(0.010)	-(0.022)	-(0.032)	-(0.007)
Disaster deaths	0.224	-2.721	-0.13	0.365	1.778	-4.108	0.242	-2.699	-1.408
	-(5.355)	-(4.055)	-(2.639)	-(4.342)	-(1.823)	-(4.760)	-(1.256)	-(3.488)	-(4.730)
Log Population	0.001	0.01	0.011	-0.002	-0.006	0.014	-0.007*	0.005	0.023**
	-(0.013)	-(0.009)	-(0.009)	-(0.013)	-(0.017)	-(0.016)	-(0.004)	-(0.011)	-(0.011)
Year 2009	0.002***	0.002**	0.002**	0.002**	0.003***	0.002**	0.003***	0.002**	0.002***
	-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)
PCA factorization of instruments	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Collapse instruments	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Instrument lag limits	(1 15)	(1 15)	(1 15)	(2 15)	(1 3)	(1 15)	(1 15)	(1 15)	(1 15)
Two-step method	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Forward orthogonal transformation	No	No	No	No	No	No	Yes	Yes	No
Backward orthogonal deviated instruments	No	No	No	No	No	No	No	Yes	No
Weighted observations	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Observations	1,804	1,804	1,804	1,804	1,804	1,804	1,804	1,804	1,804
Number of countries	132	132	132	132	132	132	132	132	132
DAHG + DAHG$^-$	-0.276	-0.568	-0.302	-0.187	-0.828	-0.454	0.592	-0.0242	-0.423
se(DAHG + DAHG$^-$)	0.35	0.168	0.147	0.397	0.608	0.229	0.263	0.529	0.416
Test up = down p-value	0.098	0.011	0.003	0.047	0.737	0.013	0.000	0.0446	0.071
Hansen p-value	0.531	0.784	0.317	0.474	0.649	0.531	0.603	0.306	0.771
AR(2) p-value	0.533	0.519	0.485	0.602	0.285	0.59	0.74	0.784	0.0477
Instrument count	20	51	67	19	12	20	20	22	20

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table S18:Sensitivity across instrument specifications; model (2)

VARIABLES	(1)	(2)	(3)	(4)
	Dependent variable: GHES / GDP			
Lag GHES / GDP	0.540***	0.528***	0.559***	0.795***
	-(0.161)	-(0.118)	-(0.144)	-(0.256)
DAHG / GDP	-0.623**	-0.805***	-0.810***	-0.633***
	-(0.270)	-(0.109)	-(0.178)	-(0.231)
$DAHG^-$ / GDP	0.347*	0.237***	0.134	0.659***
	-(0.210)	-(0.080)	-(0.316)	-(0.229)
DAHNG / GDP	0.219	0.247**	0.232	0.075
	-(0.154)	-(0.107)	-(0.231)	-(0.209)
GGE / GDP	0.025**	0.026**	0.013	-0.014
	-(0.012)	-(0.010)	-(0.021)	-(0.054)
Log GDP per capita	0.011**	0.012***	0.015**	0.018
	-(0.005)	-(0.005)	-(0.007)	-(0.018)
War deaths	-0.026***	-0.024***	-0.02	-0.002
	-(0.010)	-(0.009)	-(0.014)	-(0.031)
Disaster deaths	0.224	1.373	2.093	2.314
	-(5.355)	-(1.040)	-(1.799)	-(4.345)
Log Population	0.001	0.004	0.011	0.015
	-(0.013)	-(0.009)	-(0.013)	-(0.014)
Year 2009	0.002***	0.003***	0.002**	0.002***
	-(0.001)	-(0.001)	-(0.001)	-(0.001)
Endogenous variables	LDV, DAH-G	LDV	LDV, DAH-G, DAH-NG, GGE	LDV, DAH-G, DAH-NG, GGE, GDP
Observations	1,804	1,804	1,804	1,804
Number of countries	132	132	132	132
DAHG + $DAHG^-$	-0.276	-0.568	-0.676	0.0264
se(DAHG + $DAHG^-$)	0.35	0.126	0.35	0.392
Test up = down p-value	0.0981	0.0981	0.672	0.004
Hansen p-value	0.531	0.561	0.313	0.457
AR(2) p-value	0.533	0.422	0.32	0.619
Instrument count	20	15	22	23

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table S19: Sensitivity across specifications; model (2)

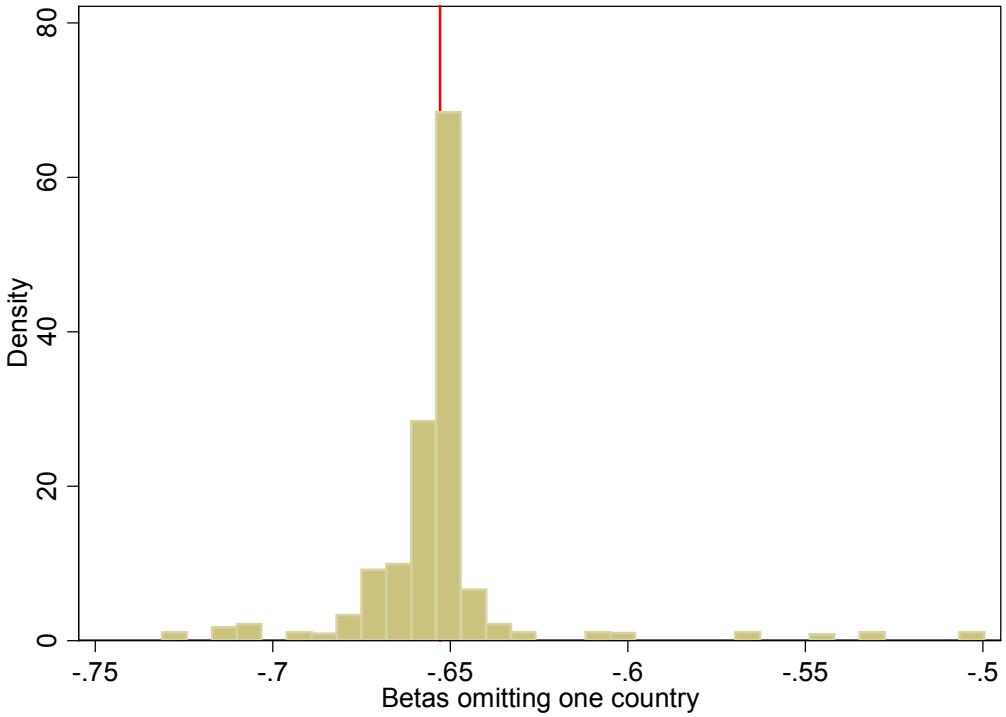
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Dependent variable: GHES / GDP										
Lag GHES / GDP	0.540***	0.548*	0.559***	0.629***	0.540***	0.548***	0.535***	0.543***	0.639***	0.643***	-0.038
	-(0.161)	-(0.286)	-(0.156)	-(0.139)	-(0.157)	-(0.155)	-(0.160)	-(0.176)	-(0.203)	-(0.190)	-(0.873)
DAHG / GDP	-0.623**	-0.758*	-0.596**	-0.568**	-0.634**	-0.617**	-0.620**	-0.627**	-0.517**	-0.532	-1.975**
	-(0.270)	-(0.402)	-(0.239)	-(0.232)	-(0.258)	-(0.271)	-(0.265)	-(0.270)	-(0.246)	-(0.364)	-(0.850)
DAHG$^-$ / GDP	0.347*	0.347	0.387**	0.415**	0.383***	0.357*	0.344	0.354*	0.545**	0.4	0.646
	-(0.210)	-(0.367)	-(0.191)	-(0.206)	-(0.141)	-(0.205)	-(0.211)	-(0.215)	-(0.251)	-(0.391)	-(0.648)
DAHNG / GDP	0.219	0.26	0.242	0.264*	0.213	0.223	0.217	0.219	0.115	0.197	0.396
	-(0.154)	-(0.245)	-(0.158)	-(0.136)	-(0.139)	-(0.155)	-(0.151)	-(0.156)	-(0.207)	-(0.174)	-(0.355)
GGE / GDP	0.025**	0.019*	0.026**		0.023*	0.025**	0.026**	0.024**	0.016	0.015	0.042
	-(0.012)	-(0.010)	-(0.013)		-(0.012)	-(0.012)	-(0.012)	-(0.011)	-(0.012)	-(0.013)	-(0.028)
Log GDP per capita	0.011**	0.032***	0.009	0.015***	0.011**	0.010*	0.010*	0.012*	0.013***	0.010**	0.017**
	-(0.005)	-(0.007)	-(0.006)	-(0.004)	-(0.006)	-(0.005)	-(0.005)	-(0.007)	-(0.005)	-(0.005)	-(0.007)
War deaths	-0.026***	-0.018*	-0.028**	-0.012**	-0.024**	-0.026***	-0.027***	-0.026***	-0.020*	-0.022***	-0.022
	-(0.010)	-(0.009)	-(0.011)	-(0.006)	-(0.010)	-(0.010)	-(0.010)	-(0.010)	-(0.012)	-(0.008)	-(0.028)
Disaster deaths	0.224	2.37	-1.130	2.304	-0.152	0.117	0.121	0.103	8.541	1.291	-14.386
	-(5.355)	-(2.361)	-(5.722)	-(4.859)	-(5.438)	-(5.351)	-(5.377)	-(5.426)	-(9.565)	-(2.311)	-(9.276)
Log Population	0.001	0.026	0.002	0.007	0.002	0.000	0.001	0.000	-0.003	0.000	0.068**
	-(0.013)	-(0.020)	-(0.014)	-(0.011)	-(0.013)	-(0.013)	-(0.013)	-(0.012)	-(0.013)	-0.014	-(0.034)
Year 2009	0.002***	0.003***		0.003***	0.002***	0.002***	0.002***	0.002***	0.003***	0.003***	
	-(0.001)	-(0.001)		-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)	-(0.001)	
Additional covariates						0.039	0.000	-0.002	0.000	0.000**	
						-(0.030)	(0.000)	-(0.010)	(0.000)	(0.000)	
Notes:	Standard set	Normalize with current GDP	Drop time indicator	Drop generally insignificant covariate	Drop excluded instruments	Add debt relief	Add HIV prevalence	Add GDP growth	Add democracy index	Add capita formation	Collapse by 3-yr time averages
Observations	1,804	1,804	1,804	1,804	1,804	1,784	1,804	1,804	1,660	1,633	387
Number of countries	132	132	132	132	132	132	132	132	132	124	131
DAHG + DAHG$^-$	-0.276	-0.411	-0.209	-0.153	-0.252	-0.260	-0.277	-0.273	0.028	-0.133	-1.329
se(DAHG + DAHG$^-$)	0.35	0.698	0.308	0.316	0.353	0.344	0.346	0.36	0.384	0.691	1.333
Test up = down p-value	0.098	0.345	0.043	0.044	0.007	0.082	0.103	0.099	0.030	0.306	0.319
Hansen p-value	0.531	0.007	0.292	0.49	0.416	0.539	0.449	0.521	0.239	0.173	0.0559
AR(2) p-value	0.533	0.524	0.6	0.488	0.587	0.536	0.531	0.541	0.943	0.221	0.555
Instrument count	20	20	19	18	19	21	20	21	20	20	12
Robust standard errors in parenthes											
*** p<0.01, ** p<0.05, * p<0.1											

Table S20: Sensitivity across estimation methods; model (2)

VARIABLES	(1)	(2)	(3)	(4)
	Dependent variable: GHES / GDP			
Lag GHES / GDP	0.540***	0.939***	0.051	
	-(0.161)	-(0.092)	-(0.066)	
DAHG / GDP	-0.623**	-0.182*	-0.916***	-0.437***
	-(0.270)	-(0.097)	-(0.092)	-(0.101)
DAHG$^-$ / GDP	0.347*	0.745***	0.080*	0.128**
	-(0.210)	-(0.280)	-(0.048)	-(0.056)
DAHNG / GDP	0.219	0.027	0.137	0.108
	-(0.154)	-(0.126)	-(0.104)	-(0.134)
GGE / GDP	0.025**	0.005	0.027**	0.035***
	-(0.012)	-(0.005)	-(0.011)	-(0.012)
Log GDP per capita	0.011**	0.001	0.015***	0.015**
	-(0.005)	-(0.001)	-(0.005)	-(0.006)
War deaths	-0.026***	-0.011	-0.027**	-0.057
	-(0.010)	-(0.007)	-(0.011)	-(0.038)
Disaster deaths	0.224	-0.355	-0.454	-6.622***
	-(5.355)	-(0.812)	-(0.833)	-(2.063)
Log Population	0.001	0.000	0.025	0.029***
	-(0.013)	(0.000)	-(0.019)	-(0.010)
Year 2009	0.002***	0.002**	0.003***	0.003**
	-(0.001)	-(0.001)	-(0.001)	-(0.001)
Constant		-0.003	0.000	-0.580***
		-(0.008)	(0.000)	-(0.162)
Estimation method	Diff GMM	Sys GMM	First Difference	Fixed Effects
Observations	1,804	1,937	1,804	2,070
R-squared			0.357	0.518
Number of countries	132	133	133	133
DAHG + DAHG$^-$	-0.276	0.563	-0.836	-0.309
se(DAHG + DAHG$^-$)	0.35	0.324	0.118	0.113
Test up = down p-value	0.098	0.008	0.099	0.025
Hansen p-value	0.531	0.271	.	.
AR(2) p-value	0.533	0.708	.	.
Instrument count	20	21	.	.

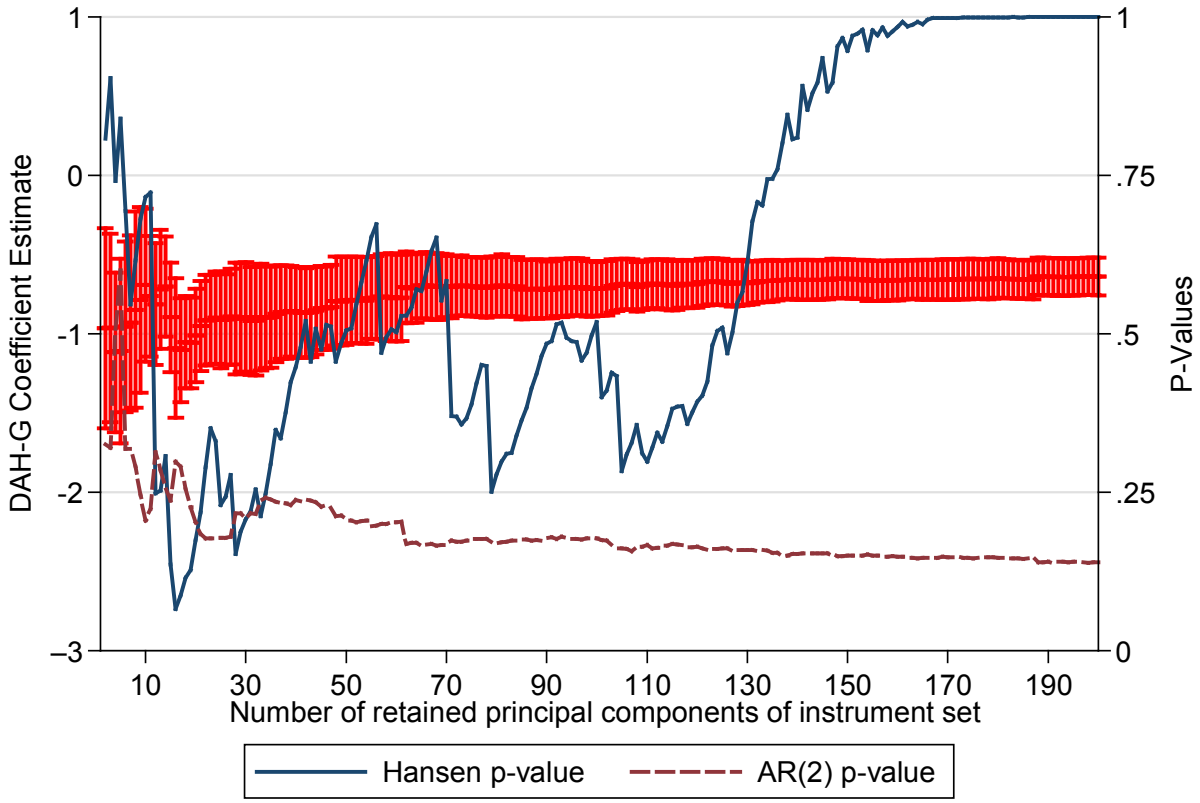
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure S1: Testing the effect of each country



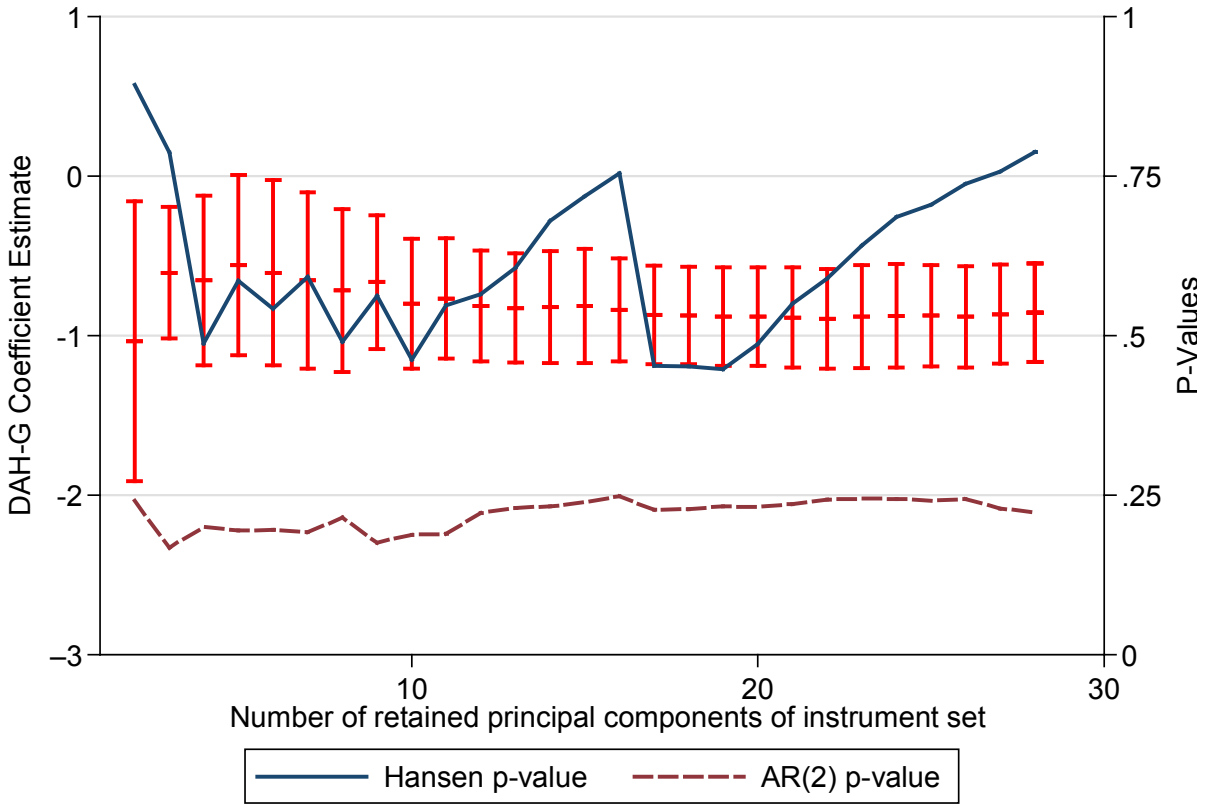
I rerun our baseline regression 132 times, each time omitting one country. The coefficients on DAHG are graphed in this histogram to measure undue pressure from a single country.

Figure S2: Stable DAH-G coeficient estimates across a range of retained components



The p-value from the Hansen J test statistic and AR(2) test statistics, and the DAHG coefficient are stable across a wide range of retained components. Estimates derived using full, un-collapsed instrument set.

Figure S3: Stable DAH-G coeficient estimates across a range of retained components



The p-value from the Hansen J test statistic and AR(2) test statistics, and the DAHG coefficient are stable across a wide range of retained components. Estimates derived using full, un-collapsed instrument set.

References

Arellano M, Bond S. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. The Review of Economic Studies 1991; 58; 277.

Arellano M, Bover O. Another look at the instrumental variable estimation of error-components models* 1. Journal of Econometrics 1995; 68; 29–51.

Bai J, Ng S. Instrumental variable estimation in a data rich environment. Econometric Theory 2010; 26; 1577–1606.

Blundell R, Bond S. Initial conditions and moment restrictions in dynamic panel data models. Journal of Econometrics 1998; 87; 115–143.

Drukker D. Testing for serieal correlation in linear panel-data models. The Stata 2003; 3; 168–77.

EM-DAT Universite Catholique de Louvain, Brussels, Belgium. The OFDA/CRED International Disaster Databse. Available at: http://www.emdat.be/database.

Farag M, Nandakumar AK, Wallack SS, Gaumer G, Hodgkin D. Does funding from donors displace government spending for health in developing countries? Health Affairs 2009; 28; 1045.

Iglewicz B, Hoaglin DC. How to Detect and Handle Outliers. 1st ed. vol. 16. ASQ Quality Press; January 1993.

Institute for Health Metrics and Evaluation. 2013. Financing global health 2012: End of the golden age? Seattle, United States; January 2013.

International Monetary Fund. 2012. World Economic Outlook, October 2012: Coping with High Debt and Sluggish Growth. Washington D.C.; 2012.

James SL, Gubbins P, Murray CJ, Gakidou E. Developing a comprehensive time series of GDP per capita for 210 countries from 1950 to 2015. Population Health Metrics 2012; 10; 12.

Kennedy P. A Guide to Econometrics. 5th ed. The MIT Press; August 1, 2003.

Lu C, Schneider MT, Gubbins P, Leach-Kemon K, Jamison D, Murray CJ. Public financing of health in developing countries: a cross-national systematic analysis. The Lancet 2010; 375; 1375–1387.

Nickell S. Biases in dynamic models with fixed effects. Econometrica: Journal of the Econometric Society 1981; 1417–1426.

Roodman D. A note on the theme of too many instruments. Center for Global Development, Washington, DC 2007.

Roodman D. How to do xtabond2: An introduction to difference and system GMM in Stata. The Stata Journal 2009; 9; 86–136.

Van de Sijpe N. Is Foreign Aid Fungible? Evidence from the Education and Health Sectors. The World Bank Economic Review The World Bank Economic Review 2012.

The World Bank. World Development Indicators.

UNAIDS and WHO. Global report: UNAIDS report on the global AIDS epidemic 2010. 2010.

Uppsala Conflict Data Program - Uppsala University. UCDP Battle-Related Deathes Dataset v.5-2011. Available at: http://www.pcr.uu.se/research/ucdp/datasets/ucdp_battle-related_deaths_dataset/.

Windmeijer F. A finite sample correction for the variance of linear efficient two-step GMM estimators. Journal of Econometrics 2005; 126; 25–51.

Wooldridge JM. Econometric Analysis of Cross Section and Panel Data. 1st ed. The MIT Press; October 1, 2001.

World Health Organization. Guide to Producing National Health Accounts 2003.

———. World Health Statistics 2012. World Health Organization: Geneva; 2012.

Xu K, Saksena P, Holly A. The Determinants of Health Expenditure: A Country-Level Panel Data Analysis. Working Paper 2011.

Chapter 3: Effects of effects: using random-effects, fixed-effects and the within-between specification for clustered data in observational health studies

Running Header: Effects of effects

Version: June 5, 2013

Author: Joseph L. Dieleman, M.A.
Department of Economics, University of Washington
Institute for Health Metrics and Evaluation, University of Washington

Correspondence: Joseph L. Dieleman
dieleman@uw.edu
Institute for Health Metrics and Evaluation
2301 5th Avenue, #600
Seattle, WA 98121 USA
+ 1 (206) 897-2800 (telephone)
+ 1 (206) 897-2899 (fax)

Abstract

When unaccounted-for group-level characteristics affect an outcome variable, traditional linear regression is inefficient and can be biased. The random- and fixed-effects estimators (RE and FE, respectively) are two competing methods that address these problems. Evidence from PubMed shows that health researchers tend to overwhelmingly favor RE estimation, while other disciplines favor FE estimation. While each estimator controls for unobserved effects, the two estimators require different assumptions. I create a diverse set of 16,200 simulated scenarios to compare the circumstances when RE and FE estimation are each most appropriate. I also evaluate an augmented version of the RE estimator that was suggested by Mundlak in 1978 and is asymptotically equivalent to the FE estimator. The simulations vary the number of groups, size of the groups, within-group variation, goodness-of-fit of the model, and the degree to which the model is correctly specified. I use mean squared error of the estimated marginal effect and root means squared error of fitted-values to indicate when each estimator is preferred. I also evaluate the conventional tool used to choose between the estimators, the Hausman test. Simulation shows that there are cases when each estimator is most appropriate, although the cases in which RE estimation is MSE-preferred are less common and reserved for a few special cases that have only a few, small groups. FE estimation is generally preferred over traditional RE, although in finite samples the augmented RE estimator out-preforms both traditional estimators. The Hausman test guides the practitioner to the estimator with the smallest absolute error only 61% of the time, and in many sample sizes simply applying the Mundlak's specification and using the RE estimator produces smaller absolute errors than following the suggestion of the test.

Keywords: random effects, fixed effects, within-between estimation, Mundlak-approach, Hausman test, observational studies

1. Background

Observational studies can be fraught with a diverse set of complications. One type of complication, which I consider here, is grouped or clustered data. When observations are clustered into groups, common group-level characteristics can affect outcomes. If *all* these unique characteristics are observed and measured it would be possible to include them in a model, but in most cases this is unrealistic. For example, health facilities in a single geographic region may face common budgets, policies, attitudes towards treatment, population demographics, disease patterns, and supplies of medications. If multiple facilities from multiple regions are considered in a single analysis, facilities will implicitly be clustered by region. Although these group-level commonalities affect the facilities' ability to supply services, it would be unrealistic to measure and include all of them in a model. However, not addressing the unifying group-level characteristics violates assumptions needed to prevent many regression methods from being biased.

Fortunately, there are two relatively common methods for improving the estimation of clustered data: random- and fixed-effects estimation (RE and FE, respectively) (Kennedy, 2003). RE and FE estimation each address unobserved group-level characteristics, and are both common throughout the social sciences. Applying these methods can eliminate bias and improve efficiency. There are substantial literatures describing the theory underlying these two estimators (Greene, 2007; Wooldridge, 2001; Snijders and Bosker, 2011; Gelman and Hill, 2006; Mundlak, 1978).

Although the theory supporting the RE and FE estimators is well established, there remains little consensus across disciplines about when each is most appropriate. Figure 1 shows health researchers have disproportionally preferred RE estimation, relative to other disciplines (National Library of Medicine (US); American Economics Association; CSA Illumina). Moreover, controversy regarding which estimator to use remains within the health sciences as well (Kravdal, 2010; Leyland, 2010; Setodji CM and Shwartz M, 2013). While it is possible that the disparate choices shown in Figure 1 are derived from fundamentally different data that requires different estimators, the research that follows shows that in practice it can be difficult to confidently choose between estimators (using standard specifications), especially in small samples. Naively following a discipline-specific norm can lead to biased or inefficient estimates.

In addition to choosing which estimator to use, correctly specifying the variables used for each estimator is also important. Traditional RE and FE specifications call for a single outcome variable (y) to be regressed on observed explanatory variables (x). In many cases, each estimation method transforms the data (discussed below) and ordinary least squares (OLS) is applied to estimate the model. Still, a number of variants of this traditional specification have proven useful. Mundlack argued persuasively in 1978 that when correctly specified, the RE and FE estimators are effectively the same estimator (Mundlak, 1978). Thus, properly specifying a model proves as important as selecting which estimator to use.

With the disparate perspectives on estimation selection and the importance of specification in mind, the objective of this paper is to explain the traditional RE and FE estimators, and to

illustrate when each estimator is most appropriate. In addition to this, I consider an alternative specification that uses the RE estimator to achieve estimates asymptotically equivalent to those from FE estimation. This hybrid, called the within-between (WB) approach, is a slightly augmented version of the specification proposed by Mundlak, and retains the best characteristics of traditional RE and FE estimation (Bell and Jones, 2012; Snijders and Bosker, 2011).

I follow Clark and Linzer and use simulation to explain and illustrate the differences between these estimators (Clark and Linzer, 2013). Clark and Linzer use simulation to evaluate the Hausman test and compare RE and FE estimation. The Hausman test is the primary tool used to assist researchers in choosing between the two estimators (Hausman, 1978). Clark and Linzer show that the Hausman test is neither a "necessary nor sufficient statistic" for choosing between the traditional estimators. Furthermore, they use a series of well-constructed graphs to illustrate that the test does not adequately assess the bias-precision tradeoff that should inform the practitioners' choice. I build upon Clark and Linzer by framing the FE versus RE choice using language and examples from population health, including a number of additional simulation dimensions, assessing each estimators ability to predict outcomes (in addition to evaluating each estimators ability to estimate marginal effects), and evaluating a third estimator – the WB estimator. The inclusion of these additional features and a different interpretation of the overlapping results lead me to a unique set of recommendations.

In the simulation, I create 16,200 unique scenarios. Each scenario is a unique combination of important dimensions, varying the number of groups, size of the groups, within-group variation,

between-group variation, amount of measurement error, amount of autocorrelation, and amount of the variation explained by the covariates. On each simulated dataset I apply the three estimators and evaluate which estimator has the least biased marginal effect estimates (coefficient estimates) and predictions (fitted-outcomes). I also apply the Hausman test to each data set.

This paper closes by offering suggestions of when each estimator and specification is most appropriate. In preparation for these suggestions, I begin here with an explanation of the underlying clustered data model and the traditionally specified RE and FE estimators. I also outline and explain the WB approach. A simulation, described subsequently, illustrates there is a time and place for each of the three estimators. Unfortunately, the simulation also shows that no single rule can offer infallible guidance when selecting an estimator, although in many cases the WB approach retains the best characteristics of the two traditional estimators. While the Hausman test can be marginally informative, it is only reliable in large samples.

1.1 Clustered data model

Observations within a sample are clustered when there are underlying characteristics, otherwise unaccounted-for, that connect observations. For simplicity, the clustered data model includes a single group-level effect which is the aggregated effect of all the group-specific characteristics not included elsewhere in the model. These unaccounted-for group-level characteristics that affect the outcome variable are often unknown or unmeasured. While group membership is

observed, the actual determinants of the outcome variable are considered unobserved, as they are not, and in most cases cannot, be included in the model.

In practice, individuals are routinely clustered into households, treatment facilities, and health status groups. Similarly, households and facilities are clustered into service platforms, states and regions; and even countries can be clustered into disease-endemic regions or income-groups. In longitudinal (panel) data, individual units (individuals, facilities, countries, etcetera) can be clustered with themselves, as each is sampled multiple instances across time.[1] Clustered data becomes problematic when unobserved group-level characteristics affect the outcome. In these cases the conditional group-level means of the outcome vary across groups. This characteristic, called unobserved heterogeneity, violates an assumption necessary for OLS to be the best linear unbiased estimator, leading to inefficient estimation and biased inference (heteroskedastic residuals) and potentially biases estimates (Kennedy, 2003; Bartels, 2008).

Equation (1) is an observation from such a model. Here y is the outcome variable of interest, x is explanatory variable, β is the marginal effect, ε is the residual, and μ is the single, aggregated, unobserved group-level effect. (Without loss of generality a constant should be included, and if appropriate the model can be extended to include many explanatory variables. Here I abstract to

[1] In longitudinal data, time points can also be clustered together, such that at a specific time a change influences all observations. This would be a second dimension of clustering. Without loss of generality, everything discussed in this paper can be generalized to a multiple dimensions of clustering, although exploring this in simulation is beyond the scope of this paper.

the simplest case.) ε is assumed to be independently and identically distributed across the sample. Subscripts *j* and *n*, where $j \in (1 \ldots J)$ and $n \in (1 \ldots N)$, indicate the group and observation identification within each group, respectively. If the data is longitudinal then *n* would indicate the time at which each observation is sampled, while *j* would indicate the unit being observed.

$$(1) \quad y_{jn} = \beta x_{jn} + \mu_j + \varepsilon_{jn}$$

For estimation, it is possible to ignore the unobserved group-level effect, rewrite equation (1) as equation (2), and apply OLS. In equation (2) the group-level effect is contained within the residual because it is not formally included in the estimation. Applying OLS to equation (2) is often called the pooled or population average estimator (Greene, 2007). There are two central concerns regarding the pooled estimator. The first concern is related to bias. Because μ_j is part of the true data generating process of y_{jn} but is ignored in estimation, the pooled estimator can suffer omitted variable bias (OVB). If I rewrite $\mu_j = z_{jn}\delta$ such that *z* is a set of *J-1* binary variables indicating group-membership and γ is the vector of effects measuring how group-membership effects y_{jn}, then it can be shown that $OVB = \frac{\text{cov}(x,z)\delta}{\text{var}(x)}$ (Greene, 2007). Thus, the pooled estimator will be biased unless the included explanatory variables and the (excluded) group-level effects are independent ($\text{cov}(x,z) = 0$). The second common concern that theorists have for the pooled estimator is related to heteroskedasticity. Heteroskedastic residuals are residual that are not distributed with the same variance across a sample. As a consequence, an

estimator suffers from biased inference (the estimated standard errors are too small) and inefficiency (it is less precise that other linear estimators) (Greene, 2007; Kennedy, 2003).[2]

$$(2) \qquad y_{jn} = \beta x_{jn} + \omega_{jn} \text{ where } \omega_{jn} = \mu_j + \varepsilon_{jn} \; \omega_{jn} = \delta z_{jn} + \varepsilon_{jn}$$

1.2 Random-effects estimation

If $cor(x_{jn}, \mu_j) = 0$ then the pooled estimator is not biased (as OVB = 0) , and the only concern regarding the pooled estimator is the heteroskedastic residuals. An alternative to the pooled estimator that controls for heteroskedasticity is the RE estimator. Like pooled estimation, RE estimation does not explicitly model the unobserved group-level effects, and thus, it is only unbiased if the group-level effect is independent from the included explanatory variables (Greene, 2007). RE estimation assumes μ_j is normally distributed with mean zero and employs feasible generalized least squares (FGLS). FGLS applies OLS to equation (3), and is an efficient method for dealing with heteroskedasticity (Wooldridge, 2001).[3]

$$(3) \qquad (y_{jn} - \theta \overline{y}_j) = \beta(x_{jn} - \theta \overline{x}_j) + (\varepsilon_{jn} - \theta \overline{\varepsilon}_j) \text{ where } \theta = 1 - \frac{\text{var}(\varepsilon_{jn})}{\sqrt{\text{var}(\varepsilon_{jn}) + N \, \text{var}(\mu_j)}}$$

[2] These problems occur if the group-level effects do varies across the sample ($\mu_j \neq \mu$ for at least some *j*). In clustered-data it is unlikely that the group-level effect does not vary across groups.
[3] In practice, maximum likelihood estimation often replaces FGLS. It is asymptotically equivalent.

1.3 Fixed-effects estimation[4]

Unlike the pooled and RE estimators, the FE estimator explicitly models the group-level effect. To do this, the FE estimator includes a set of *J-1* binary variables indicating group membership. Each group has its own indicator. When OLS is applied to equation (4), where *z* is a matrix of *J-1* indicator variables and δ is a vector of marginal effects, it is called the least squares dummy variable (LSDV) version of FE estimation (Greene, 2007).

$$(4) \qquad y_{jn} = \beta x_{jn} + \delta z_{jn} + \varepsilon_{jn}$$

In practice the LSDV version of FE is often replaced by a numerically equivalent estimator that is less computationally taxing. This estimator regresses *y*, net of the group-specific *y* mean, on *x,* net of the group-specific *x* mean. That is, OLS is applied to equation (5). Equation (5) illustrates why using this transformation accounts for the unobserved group-level effects even though μ_j is not explicitly included. FE estimation differences away all variation between groups and relies completely on variation within groups. This is why the FE estimator is sometimes called the within estimator. The FE estimator literally assesses how changes in *y*, *within* each group, are associated with changes in *x*, *within* each group.

[4] In some disciplines the term "fixed-effects" is used to mean a marginal effect that is constant across the sample. Using that terminology all the right-hand-side variables from equations (1) - (6) would be considered fixed, because β is assumed to be homogenous. Using this alternative terminology, fixed-effects are contrasted with random parameters (or effects), which allow for group-specific marginal effects. For this paper, FE estimation refers exclusively to unobserved group-specific effects that set the groups-specific intercept (the constant). Here, RE estimation refers to estimation with a group-specific intercept.

$$(5) \quad (y_{jn} - \overline{y}_j) = \beta(x_{jn} - \overline{x}_j) + (\mu_j - \mu_j) + (\varepsilon_{jn} - \overline{\varepsilon}_j)$$

1.4 Comparing random- and fixed-effects estimation

Both RE and FE estimation rely on the assumptions of OLS. The estimated models (equations (3) and (5), respectively) must be correctly specified, each variable of *x* must be strictly exogenous and linearly independent, and the residual must be independently and identically distributed. When these conditions are met, theory states that FE estimation is unbiased and consistent. RE estimation requires the addition assumption; the group-level effect and the included explanatory variables must be independent in order to avoid OVB. When this assumption is met, RE estimation is unbiased, consistent, and, because it utilized both the within- and between-group variation, it is efficient. In these circumstances, FE estimation is not efficient because it only utilizes the within variation (Wooldridge, 2001).

Thus, it is the correlation between the explanatory variable(s) and the group-level effect that distinguishes which of these two estimators to utilize. There has been great confusion on this matter (Bartels, 2008; Wooldridge, 2001). Determining if group-level effects are *random*, meaning they are representative of random draws from broader population is of marginal importance, as a (large) number of draws from any cross-section will most likely appear random (Wooldridge, 2001; Mundlak, 1978). More importantly, if $cor(x_{jn}, \mu_j) = 0$ then RE estimation is unambiguously superior to FE estimation, regardless of if group-level effect is determined to be random. In these cases, both estimators are unbiased and consistent, but only RE is efficient.

In many situations, even if the sample is randomly drawn from a larger sample, the unobserved effect and included explanatory variables are not independent, $cor(x_{jn}, \mu_j) = \rho \neq 0$. In these cases, the OVB of RE estimation increases with ρ. There are many reasons why the unobserved group-level effect and the included explanatory variables might be correlated. For example, when health facilities are clustered into regions, the unobserved group-level effect controls for policies, supply of medicines, disease patterns, and budgets common across each region. It is highly probable that these characteristics that make up the unobserved effect are correlated with variables included in the estimation such as population density or number of physicians. Similarly, in cross-country analyses of population health, it common for researchers to cluster countries into geographic regions to control for unobserved disease patterns and demographic characteristics. Undoubtedly these unobserved group-level characteristics are correlated with even the simplest covariates included in the model such as gross domestic product, baseline population health status, or education levels. In all of these cases, RE estimation is biased.

Unfortunately, knowing RE estimation is biased does not unambiguously guide a researcher to the FE estimator. Even when $cor(x_{jn}, \mu_j) = \rho \neq 0$, RE estimation remains a more precise estimator than the FE estimator, although, as ρ moves away from zero, RE estimation precisely estimates a biased marginal effect. For this reason choosing between the two estimators, even if ρ were to be known, is not always simple and revolves around a bias-precision tradeoff. Clark and Linzer state "The appropriate question to ask… is *how much* bias results – and whether the resulting bias can be justified by the gain in efficiency (Clark and Linzer, 2013)."

Figure 2 illustrates why assessing the bias-precision tradeoff can be less than straight forward. In each of the three panels, two distributions are shown. For each simulated dataset, the RE estimator (red) and FE estimator (blue) imperfectly estimates the marginal effect, leading to one error $(\hat{\beta} - \beta)$ per estimator per simulated dataset. This is repeated 1,000 times to create a distribution of1,000 errors for each estimator. The accompanying dashed vertical lines show the mean of each distribution of errors. In Figure 2, bias is illustrated when the mean error (the dashed vertical line) is different from zero. Precision is illustrated by a tight distribution of errors.

Panel (a) of Figure 2 shows cases when $\rho = 0$, while panel (b) shows cases when $\rho = 0.3$, and panel (c) shows cases when $\rho = 0.6$. In panel (a) RE is clearly a superior estimator because it is unbiased and efficient. On the other hand, panel (c) shows cases where it is clear that FE is superior (because RE is so biased). More ambiguity is present in panel (b). The vertical line right of zero shows that RE is biased upward, but the variance around the RE errors' mean is much smaller. It is possible that despite the bias, the absolute error from RE estimation is still sometimes smaller than the corresponding FE error. For panel (b) theory is less instructive and the practitioner needs to make a judgment between accepting bias or imprecision.

In addition to weighing the costs of bias and imprecision, there are several other practical differences between the traditional RE and FE estimators. One difference exists because the FE estimator removes all between-group variation and only uses within-group variation. This means that group-level variables (that are constant across the entire group) cannot be included in the

estimated model. For longitudinal data, this means that time-invariant variables cannot be included. Thus, when the objective of an analysis is to measure the effect of group-level variables, FE estimation is not a viable method. On the other hand, RE estimation capitalizes on both within- and between-group variation, and thus allows for the inclusion of variables that are constant within a group.

A second distinction between the two traditional estimators is that RE estimation tends to be more flexible and fits easily into a hierarchical framework. Using this framework, groups are easily nested within one another, variables that effect different levels of a hierarchy are more easily explored, and heterogeneous marginal effects can be explored in a random coefficient context (Greene, 2007; Snijders and Bosker, 2011). While greater flexibility is a noteworthy advantage favoring RE estimation, it does not obviate the assumption regarding the independence between the group-level effects and explanatory variable. When this assumption is not met, the traditional RE estimation will be biased.

1.5 The within-between approach

In the traditional specifications, RE is precise and quite flexible, but is also likely to be biased. Alternatively, FE estimation is unbiased, but less flexible, less precise, and can't be used to explore the effect of group-level characteristics. In this subsection I explore one estimation variant meant to marry the two traditional estimators and take advantage of the best characteristics of each. The hybrid I consider has several names, and is the same or vary similar to the "within-between" estimator by some (Snijders and Bosker, 2011; Bell and Jones, 2012),

the "Mundlack-approach" by others (Greene, 2007; Debarsy, 2012), or the "Mundlak correlated random effects" model by another (Wooldridge, 2010). It applies RE estimation to a non-standard FE specification. Quite simply, this estimator applies RE estimation to equation (6), where $\overline{x}_j$ is the group-level mean of each of the included explanatory variables.[5]

$$(6) \qquad y_{jn} = \beta\left(x_{jn} - \overline{x}_j\right) + \gamma\overline{x}_j + \omega_{jn} \text{ where } \omega_{jn} = \mu_j + \varepsilon_{jn}$$

Some consider the WB approach to be a compromise between RE and FE. While the mechanics applied and versatility are based on RE estimation, the group-demeaning of *x* essentially applies an alternate version of FE estimation. The group-demeaning ensures that $\hat{\beta}$ from equation (4), (5), and (6) are asymptotically equivalent (Wooldridge, 2010), causing Mundlak to declare, "Thus there is only one estimator (Mundlak, 1978)." $\hat{\beta}$ can be considered the "within" effect, which assess changes within a group (as the FE estimator does), while $\hat{\gamma}$ measures the effect of *x* between groups (Snijders and Bosker, 2011). Most importantly, the explanatory variable and the unaccounted-for group-level effects of equation (6) will be fully independent when the group mean is also included as an explanatory variable. Thus, this specification of RE is unbiased. "The whole literature which has been based on an imaginary difference between the two estimators… is based on an incorrect specification [of RE] which ignores the correlation between the [group-level] effects and the explanatory variables (Mundlak, 1978)."

[5] Mundlak suggested estimating $y_{jn} = \beta x_{jn} + \gamma\overline{x}_j + \omega_{jn}$ where $\omega_{jn} = \mu_j + \varepsilon_{jn}$, although Bell and Jones (2012) point out three reasons why equation (6) should be preferred. For space, I only consider equation (6) in simulation.

While the WB approach and FE estimation are equivalent asymptotically, in finite samples this is not the case. Furthermore, using the RE estimator on an augmented specification to derive the standard FE estimates does not solve the bias-precision trade-off. It is still plausible that in some scenarios a biased RE estimate might be preferred to the unbiased estimate from FE estimation or the WB approach. This possibility is even more likely when samples are small. To explore these concerns, I turn to simulation.

2. Methods

To garner insights on the RE and FE estimators and the WB approach, I simulate and draw conclusions from over 16 million datasets. The small sample properties for the WB approach are not well known. Moreover, theory offers limited guidance for adequately addressing the bias-precision tradeoff. While it may sound ideal to never use a biased estimator, Figure 2 (shown above) illustrated that at times using an imprecise estimator might be worse than using a slightly biased estimator. This section includes an explanation of the dimensions that define each data generating process, an explanation of how each dataset is generated, and an explanation of the metrics used to compare RE, FE and WB estimators.

2.1 Input dimensions

I consider 16,200 unique combinations of input parameters, simulating 1,000 datasets for each combination. The dimensions adjusted are listed in Table 1. I allow the number of clusters (J) to

range from ten to 100 and number of observations per cluster (*N*) to range from five to 50. The group-level effect (μ_j) is normally distributed with mean zero and variance one. The single explanatory variable (x_{jn}) is normally distributed with mean zero and variance (σ_x^2). σ_x^2 is set to be the same, half, or double the variance of the group-level effect. The variance of the explanatory variable can be disaggregated into the variance that is within groups (σ_W^2) and variance between groups (σ_B^2), such that $\sigma_W^2 + \sigma_B^2 = \sigma_x^2$. My simulation includes cases where 10% to 90% of σ_x^2 variance is from within groups. The residual (ε_{jn}) is also normally distributed with mean zero and variance (σ_ε^2), such that 10% to 90% of the variation of *y* is explained by residual. The marginal effect (β) of the single explanatory variable is set to one. Most importantly, the correlation (ρ) between group-level effect and the explanatory variable is set to range between zero and 0.7.[6]

To include a wide-set of plausible scenarios, I also consider two types of model misspecification. The first type of model misspecification is that I set the correlation between the explanatory variable and residual to be zero or 0.2 ($cor(x_{jn}, \varepsilon_{jn}) = \psi$). When $\psi > 0$ the explanatory variable is endogenous and OLS is biased. In observed data, an endogenous variable can be caused by measurement error, reverse causation (simultaneity), serial correlation, or an omitted variable. The second type of model misspecification is that I induce serial correlation directly into the

[6] In many cases it is impossible to generate $\rho > 0.7$ and maintain the desired values for the within, between, and total explanatory variable variance. In these cases the covariance matrix needed to generate the group-level mean of the explanatory variable that is properly correlated with the group-level effect is not positive semi-definite.

residual. In these cases, the correlation between residual and the pervious within group residual (assuming longitudinal data) is zero or 0.2.

2.2 Simulating the datasets

To generate the simulated datasets I follow a four-step process. First, I use *J* and *N* to set the total number of observations, such that *j* and *n* uniquely identify each of the *J***N* observations. Second, I randomly draw two *J*-length vectors from a multivariate normal distribution, such that:

$$(7) \qquad \begin{bmatrix} \mu_j \\ x_j \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \kappa \\ \kappa & \sigma_B^2 \end{bmatrix} \right)$$

where $\kappa = \rho * \sigma_x$, $\bar{x}_j$ is the group mean explanatory variable, $\sigma_B^2 = (1-\tau)\sigma_x^2$, and $\sigma_W^2 = \tau\sigma_x^2$. Third, I randomly draw two *J***N*-length vectors from a multivariate normal distribution such that:

$$(8) \qquad \begin{bmatrix} x_{jn} \\ \varepsilon_{jn} \end{bmatrix} \sim MVN\left(\begin{bmatrix} \overline{x_j} \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_W^2 & \psi \\ \psi & \sigma_\varepsilon^2 \end{bmatrix} \right)$$

and (9)

$$\sigma_\varepsilon = \frac{\pi}{2(\pi-1)} \left[\frac{-2\sigma_x \psi \sqrt{1+\sigma_x^2+2\rho\sigma_x}}{\sqrt{1+\sigma_x^2}} \pm \sqrt{\frac{4\sigma_x^2\psi^2\left(1+\sigma_x^2+2\rho\sigma_x\right)}{1+\sigma_x^2} - 4\left(1-\frac{1}{\pi}\right)\left(1+\sigma_x^2+2\rho\sigma_x\right)} \right].$$

Equation (9) ensures that the variance of the residual accounts for π of the variance of the outcome variable ($\sigma_\varepsilon^2 = \pi\sigma_{\hat{y}}^2$). Fourth, I generate my outcome variable y_{jn} following equation (1).

2.3 Evaluating RE and FE estimation

For each simulated dataset I apply three estimators. I complete traditional RE and FE estimation and apply the WB approach. I measure the error of the estimated marginal effect $(\hat{\beta} - \beta)$, and for each estimator and each combination of input parameters I calculate the distributions of the 1,000 errors (mean, 2.5th percentile, and 97.5th percentile) and the mean squared error (MSE). MSE is a valuable summary statistic because it penalizes an estimate for both bias and inefficiency, such that $MSE(\hat{\beta}) = \text{var}(\hat{\beta}) + (bias(\hat{\beta}, \beta))^2$

I also evaluate how good each estimator is at predicting fitted values of the outcome variable ($\hat{y}_{jn}$). I calculate the root means squared error (RMSE) for predicted values for each simulation. For each estimator and combination of dimensions, I calculate the distribution of the 1,000 RMSE (mean, 2.5th percentile, and 97.5th percentile).

Finally, for each simulation I also conduct the Hausman test (Hausman, 1978). The Hausman test is the conventional tool used to guide practitioners towards or away from the RE estimator. The test is based on the intuition that if the estimated marginal effects of RE and FE are not statistically different then both estimators must be unbiased and consistent. This conclusion makes sense because FE is unbiased and consistent, so if the estimated marginal effects are not statistically different it is reasonable that the omitted variable bias that could corrupt RE is negligible. Thus, the null of the Hausman test is that both RE and FE are consistent. If the null cannot be rejected, then conventional wisdom suggests RE estimation should be applied because

it is efficient. On the other hand, if the null is rejected, conventional wisdom suggests FE estimation.

3. Results

3.1 Baseline results

I start by comparing the RE, FE, and WB approach estimators using the baseline simulation setup, and altering only the correlation between the group-level effects (ρ), number of clusters (J), and number of observations per group (N).[7] The baseline inputs are listed in Table 1. At the baseline, the group-level effect's variance (σ_{μ}^2) and explanatory variable's variance (σ_x^2) are one. I disaggregated σ_x^2 evenly, such that the between-group variance (σ_B^2) and within-group variances (σ_W^2) are each set to 0.5 ($\tau = 0.5$). The baseline setup also assumes that the model is correctly specified (no endogeneity or serial correlation is induced, $\psi = \upsilon = 0$) and 50% of the variance of the outcome variable can be explained by the group-level and explanatory variable ($\pi = 0.5$).

[7] Clark and Linzer employ a series of graphs that illustrate bias and MSE at various sample sizes, across ρ (Clark and Linzer, 2013). These graphs are an effective and efficient method to displace a host of results. For comparability, Figures 3 through 11 of this paper modeled after Clark and Linzer's original graphs.

Figure 3 assesses the distribution of errors for the baseline setup. In Figure 3, and all the subsequent figures, the x-axis measures ρ, red is used for the RE estimation, blue is used for FE estimation, and green is used for WB approach. The y-axis shows the error of the marginal effect estimate $(\hat{\beta} - \beta)$. The solid line is the mean of the 1,000 errors, while the space within the dashed lines shows the 95% spread of the errors resulting from each estimator. Each of the nine graphs has a different number of groups (J) and observations per groups (N). In Figure 3 bias can be seen when the mean of the distribution of errors is not zero. Precision can be seen by examining the range of the errors, or the width (up and down) of the band surrounding the mean error.

Three important insights from theory are confirmed in Figure 3. First, the solid blue and green lines are always at zero, meaning the FE estimator and WB approach are unbiased. Second, RE estimation is more precise than the two competing estimators. The variance of errors around the mean is smaller for the RE estimator. Third, as ρ increases, RE estimation becomes biased.[8] These points are confirmed at all sample sizes, but are clearest and most extreme in the smallest samples.

Figure 3 suggests that at some points it might be advantageous for a practitioner to use the more precise RE estimator even if it is biased. For example, the $J = 10$, $N = 5$ and $\rho = 0.1$ case illustrates some scenarios where the bias might be small enough that the practitioner would

[8] The direction of the bias is the same as the direction of the correlation between the group-level effects and the explanatory variable, and symmetric set of graphs could be made with $\rho < 0$, resulting in downward bias of the RE estimator.

prefer accepting it rather than deferring to the less precise FE estimator. Figure 4 makes this point more clear. This figure illustrates the MSE empirically generated from the 1,000 simulations, for each estimator and J, N, and ρ combination. Figure 4 uses the same baseline simulation inputs and displays a similar set of nine graphs, with J, N, and ρ varying. The MSE provides one potential guide for dealing with the bias-precision tradeoff because it combines bias and precision into a single metric. In Figure 4, when the red line is less than the blue and green lines, RE estimation is MSE-preferred, even if it is slightly biased. Based on this metric, Figure 4 shows that in most baseline cases there is either negligible difference between RE, FE, and WB approach or FE and the WB approach are clearly MSE-preferred.

Figure 5 evaluates if the Hausman test is a valuable guide for determining which estimator to select when ρ is unknown (as it is in most cases). The backgrounds of the graphs in Figure 5 are based on the MSE estimates graphed in Figure 4. If RE is MSE-preferred the background of Figure 5 is set to red. On the other hand, if FE estimation is MSE-preferred, then the background of Figure 5 is set to blue. If the MSE of the two estimators are within 0.005 of each other, I consider this a trivial difference and leave the background white.

In Figure 5, the *solid* orange lines shows the share of the 1,000 simulations for which the Hausman test did not reject the null hypothesis, using $\alpha = 0.1$. In other words, the solid orange line shows the share of the 1,000 simulations that the Hausman test recommended the standard specification of RE estimation. If the Hausman test is an effective test, the orange line will be near 1 (100%) when the background is red, and near zero when the background is blue. The

dashed orange lines of Figure 5 show the share of the 1,000 simulations where the Hausman test recommended estimator with the smallest absolute error ($\left|\hat{\beta} - \beta\right|$), comparing only the traditional specifications of RE and FE. This metric assumes that the practitioner prefers to minimize absolute error and has a binary choice between traditional RE and FE estimation. When the dashed orange line is near one it indicates that the Hausman test recommended the "better" estimator nearly 100% of the time.

Figure 5 confirms that the Hausman test is most effective in large samples. In small samples, the test frequently fails to reject the null hypothesis even at relatively large ρ. As a result, the Hausman test suggests the "better" estimator roughly 50% of the time. As the sample size increases, especially in *J*, the estimator is more effective – the Hausman test is more apt to reject the null as ρ increases and the frequency the test recommends the "better" estimator moves towards 100%.

Figure 6 shows that FE estimation is better at predicting outcomes. The basic setup of Figure 6 is the same as the earlier figures, although in this case the y-axis measures the distribution of 1,000 RMSE, where a RMSE statistic is derived for each estimator and each simulation based on the error of the *J***N* predicted values ($\hat{y}_{jn} - y_{jn}$). The solid line shows the mean of the 1,000 RMSE, while the pairs of dashed lines show the 95% range of RMSE, for each estimator. No matter what the combination of *J*, *N*, and ρ, FE estimation is a better predictor. The mean RMSE

and variance of the RMSE are both smallest for FE estimation (relative to RE estimation and the WB approach), meaning that the FE estimator is a less biased and more precise predictor.

3.2 Varying between-group and within-group variance

Figure 7 evaluates the effect of adjusting the explanatory variable's variance such that the majority of σ_x^2 is caused by between-group variation. This adjustment leaves only a small portion of σ_x^2 within groups. For Figure 7, I leave $\sigma_x^2 = 1$ but increase σ_B^2 to 0.9 and decrease σ_W^2 to 0.1 (by setting $\tau = 0.1$). This is characteristic of data with a great deal of (observed) heterogeneity between groups and little change within groups. Examples of such data could include antenatal care within a region, health facilities' expenditure over time, or countries' rate of maternal mortality over time. If the data is longitudinal, this could be considered "sluggish" data (Clark and Linzer, 2013).

Considering three (instead of nine) combinations of *J* and *N*, Figure 7 illustrates the distribution of the errors of the marginal effect estimates (row 1), the MSE of the estimated marginal effects (row 2), and effectiveness of each estimator at predicting the outcome variable (row 3). Row 1 shows that with small sample size, RE estimation is much more precise than the FE estimation and WB approach. This is to be expected, because RE is using both between- and within- group variation, whereas FE only uses within-group variation and the coefficient that is being evaluated from the WB approach only measures the within-group effect. At the smallest sample size, row 2 shows that RE is always MSE-preferred, despite being biased (when $\rho > 0$). On the other hand,

the largest sample size ($J = 100$, $N = 50$) shows that the difference between RE, FE, and the Mundlack-approach is either negligible or FE estimation and the WB approach are MSE-preferred. In-between these two extremes the medium sized sample ($J = 50$, $N = 10$) exhibits more ambiguity. Row 3 confirms that FE remains marginally, but unambiguously, a better predictor of outcome variables regardless of J, N, and ρ.

While not shown here, the Hausman test offers only marginal insight for these cases, especially when the sample size is small to moderate. This has been recently document elsewhere in this literature (Hahn et al., 2011). In these simulations, the test recommends the "better" estimator 55%, 54%, and 74% of the time, for the three sample sizes illustrated in Figure 7, respectively. These estimates confirm a trend I see across all of my simulation. At small to moderate sample size scenarios (with observations less than or equal to 500), the Hausman test offers limited guidance, except when ρ is exceptionally large. As the sample size increases past 1,000 observations the Hausman test can be relied on more heavily.

Figure 8 illustrates the baseline setup adjusted so that 75% of the variation of the explanatory variable is found within groups, while only 25% is between groups. This type of data is characteristic of health outcomes grouped by facility, national mortality rates grouped by region, or HIV prevalence over time. In these settings there is a great deal of heterogeneity within the group. Figure 8 shows that except for the smallest samples with ρ at or near 0, FE estimation and the WB approach preform as good or better than RE estimation.

3.3 Varying the amount of variance explained by the model

Figures 9 and 10 vary the share of the variance of the outcome variable that is explained by model (π). Returning all other inputs back to the baseline level, I set $\pi = 0.9$ which sets the variance of the residual so that the estimated model is equipped to measure only 10% of the variation of y. This data would be characteristic of any model that is poorly fit, with a relatively small R^2 (coefficient of variation). Figure 9 shows that in small samples, RE estimation clearly outperforms FE estimation and the WB approach, and the RE estimator is unambiguously MSE-preferred, regardless of ρ. As the sample size grows, especially in J, FE and the WB approach become more tenable estimators. In the largest sample it is clear that FE and Mundlak estimation are either equivalent to RE estimation or MSE-preferred.

Figure 10 considers a model that is exceptionally well fit with the error only explaining 10% of the variance of y. In these cases, FE and the WB approach are clearly equivalent or MSE-preferred, relative to RE estimation, even in the smallest samples. Row 3 shows that while FE remains a better predictor of outcome variables, it is only at the smallest margin.

3.4 Varying all the other dimensions

Figure 11 examines simulations with the baseline setup, but ψ is adjusted to 0.2 so that the explanatory variable is endogenous. This is characteristic of data suffering from measurement error, reverse causation, serial correlation, or an omitted explanatory variable that is correlated with an included explanatory variable. What is unique about Figure 11 is that we see for the first

time that the FE estimator and WB approach are biased, as the necessary assumptions of OLS are not met. Still, the endogeneity of *x* does not change our primary baseline findings—except for very small samples, FE estimation and the WB approach are effectively equivalent or MSE-preferred over RE estimation. Simulation shows that varying the variance of the explanatory variable or inducing serial correlation into the residual has trivial effects on the choice between the three estimators. Across all these simulations, the baseline results hold. The effects of these simulation parameters are reported in the web appendix, but not discussed further here.

3.5 Comparing traditional fixed effects estimation and the WB approach

Asymptotically the traditional FE estimator and WB approach are equivalent; both are consistent. Still, this might not be the case in finite samples. Figures 5 through 11 show that FE estimation is an unambiguously superior predictor. The mean RMSE for each scenario is smaller for the FE estimator than the mean RMSE for the WB approach. Furthermore, the distribution of RMSE is tighter for the FE estimator, meaning it is a more precise predictor as well.

The same cannot be said regarding the estimated (within-group) marginal effects ($\hat{\beta}$). Figures 3 and 6 through 11 show that the distribution of the errors appear to be identical, but comparing the MSE suggests that in small samples the two estimators are not the same. To quantify this difference, I subtract the FE approach's MSE from the WB estimation's MSE, for each of the 16,200 scenarios. As the observations increase to infinity, the estimates from the two estimators should converge to each other and towards the true marginal effect. Thus, the two MSE should

be identical and the difference between two MSE should converge to zero. In small samples, I see that the difference between the MSE is not zero, nor centered at zero. 40% of the MSE differences across all scenarios with less than 500 observations are less than -0.01, while 0% are greater than 0.01. This means that in a disproportionally large amount of small sample scenarios, the WB approach garnered a smaller MSE.

To explore this further I look at the pairwise correlation between the simulation input parameters and difference found by subtracting the FE approach's MSE from the WB estimation's MSE, assessing only scenarios with less than 500 observations. Table 2 shows that several input parameters are correlated with the difference in MSE, statistically significant at the 99% confidence level. The most important predictor of a smaller WB appoach's MSE is the number of groups and number of observations per group – as either increases, the difference between the Mundlak MSE and FE MSE moves towards zero. The next two determinates most correlated with the difference in MSE are the share of variation within versus between groups and how well the model is fit. As the share of the variation in the explanatory variable shifts to be more within-groups, the difference between the two estimators moves towards zero. Similarly, as the model fit gets worse the two estimators become more equal as well. To summarize these points: the FE and WB approach are generally very similar, although in small samples the WB approach is generally superior at estimating within-group marginal effects. This superiority is especially true the smaller the sample size, the larger the share of the variation is between groups, and the better the model fit. In small samples (less than 500 observations), the FE estimator never has a smaller MSE (with a difference greater than 0.01).

As sample size increases, the differences between the WB approach and FE estimates diminish. For samples with between 500 and 1,000 observations (inclusive), the WB approach is MSE-preferred beyond a non-trivial amount (0.01) for only 19% of the scenarios that have. As the number of observations climbs above 1,000, none of the scenarios prefer FE or WB approach above a 0.01 margin. Thus it seems clear that as the number of observations increases, the two methods generate indistinguishable estimates.

4. Conclusion

When analyzing clustered data, researchers make practical decisions regarding which empirical methods to use and how to specify their model. Both RE and FE estimators control for unobserved group-level effects. Theory offers one significant piece of guidance: the bias of RE is directly related to $cor(x_{jn}, \mu_j) = \rho$. Thus, when it is known with certitude that $\rho = 0$, RE estimation is the obvious choice. In most observational studies, though, it seems likely that $\rho \neq 0$. Often it is not hard to imagine why some correlation between the group-level effects and the explanatory variables will bias the marginal effect estimation.

When it is impossible to conclude with confidence that $\rho = 0$, choosing between the two estimators can be difficult. Researchers are forced to choose between a potentially *precise biased* estimator and an *imprecise unbiased* estimator. Little convention exists about how to balance this bias-precision tradeoff and properly specify estimation. Fortunately, simulation can

offer some guidance. Five simple and practical rules-of-thumb can be derived from this set of simulations.

First, if the purpose of an analysis is prediction (as opposed to inference on marginal effects) then FE is unambiguously the preferred estimator. Figure 6 and row 3 of Figures 7-11 show that the FE predictions always have a smaller mean RMSE. Because the FE estimator cannot predict outcomes for groups that are not included in the original estimation, the WB approach is clear second-best estimator in these situations.

Second, when comparing the WB approach and traditional FE estimation and the purpose of an analysis is inference on marginal effects, this simulation suggests the WB approach is always preferred. The simulation confirms that the within-group marginal effects estimates are asymptotically identical, while the WB approach provides researcher with additional information regarding the between group marginal effects ($\hat{\gamma}$). In the small samples in which I have concentrated here, the WB approach consistently has (essentially) equivalent or smaller MSE. Outside of variation that can be attributed to noise, there are no scenarios when FE is MSE-preferred when compared to the WB approach. Furthermore, since the WB approach operates using RE estimation it has a flexible environment, with options extending to nested groups, hierarchical models, and random coefficients. Moreover, there is some evidence that the WB approach is also appropriate for non-linear estimation as well (Wooldridge, 2010).

Third, as a general rule, the larger the sample size the more a practitioner should avoid traditional RE estimation. Applying FE estimation on all simulated samples with greater than 500 led to a median absolute error of 4% of the true marginal effect. RE estimation led to a median absolute error of 8% of the true marginal effect. Furthermore, in simulations with more 1,000 observations, RE estimation was only MSE-preferred beyond a trivial threshold (0.005) in a very few cases where 90% of variation of *y* could not be explained by the model ($\pi = .9$). It is important to note that this rule contradicts the tendency researchers have to use RE estimation for problems with large *J*. Presumably researchers are hesitant to "waste" degrees-of-freedom for each of the *J* groups when *J* is quite large. While it is true that more degrees of freedom are used by choosing FE in large-*J* scenarios, simulation shows that using that FE estimation is MSE-preferred (relative to RE) in most large-*J* scenarios. Furthermore, this is again a case where the WB approach is a valuable compromise between the two traditional estimators. The WB approach generates practically equivalent estimates as FE but uses only a fraction of the degrees-of-freedom.

Fourth, small samples mark the rare occasion when a practitioner might consistently choose precision over bias. My simulation show that, when combined with small samples (of less than or equal to several hundred observations), two observable data characteristics make it especially likely that RE estimation would be MSE-preferred. One scenario is when the estimated model explains a very small portion of the variation in the outcome measurement. When small sample is combined with a poorly fit model, the imprecision of FE and Mundlak estimation tends to mislead the researcher more than the bias of RE estimation, even at large ρ. The goodness-of-fit of the clustered model can be explored by looking at the R^2 statistic associated with LSDV

estimation. Considering only simulations with $R^2 < 0.5$ and less than 500 observations, the traditional RE estimator had a smaller absolute error than the FE estimator 57% of the time. In these same cases, the RE estimator had a median absolute error of 33% of the true marginal effect, relative to 37% from the FE estimator.

Another unique scenario when the RE estimator is consistently MSE-preferred and should be considered is small samples that have within-group variation for the variable of interest that is relatively small. Again, in these cases the imprecision of the FE and Mundlak estimators might be more caustic than the RE estimator's bias. Within-group variation can be measured by taking the ratio of the average variation within each group relative to the total variance of the same variable. In simulation cases with less than 500 observations and within variation less than 20% of the total variation, RE estimation led to a smaller absolute error 53% of the time.

The fifth rule-of-thumb garnered from this simulation is that the Hausman test is not tremendously insightful. The test preforms best in large samples. Unfortunately, these are the same cases when the test is needed the least. Looking only at simulations with greater than or equal to 1,000 observations, the Hausman test recommends the estimator with the smaller error 67% of the time. Following the tests' suggestions (of using FE when p-value < 0.1) yields a median absolute error of 4% the true marginal effect. On the other hand, when the FE estimator is blindly applied to all these same cases, it generates the same median absolute error and is the "better" estimator (when compared to traditional RE) in 70% of the simulations. In small samples (with fewer than 500 observations), the Hausman test correctly suggests the estimator

with smaller absolute error only 55% of the time, meaning that is was only a marginal improvement upon randomly selecting an estimator. In these cases, the median absolute error derived by following the Hausman test is similar to those found by always applying the RE or FE estimator.

In population health analyses the distinction between RE and FE estimation receives little attention. When discussed, attention seems to have focused on the relatively irrelevant distributional assumptions or the Hausman test, and too often neglects the bias-precision trade-off. Furthermore, the importance of correctly specifying RE estimation seems lost on many disciplines. When correctly specified, the RE estimator is the WB estimator and achieves unbiased estimates asymptotically equivalent to FE estimation. This research shows the value of understanding each estimator and critically considering which estimator and specification is most appropriate. Fully understanding the assumptions and consequences associated with each estimator and specification will likely yield preferred results and improve the science of our discipline.

Acknowledgements

In addition to those thanked in the prefix, I would like to thank Andrew Bell, Drew Linzer, Marie Ng, Anirban Basu, Miriam Alvarado, and Benjamin Brooks for their valuable feedback regarding this chapter. Any remaining errors are of course my own.

Abbreviations

FE	Fixed effects
FGLS	Feasible generalized least squares
LSDV	Least squares dummy variables
MSE	Mean squared error
MVN	Multivariate normal distribution
OLS	Ordinary least squares
OVB	Omitted variable bias
R^2	R-squared; coefficient of variation
RE	Random effects
RMSE	Root mean squared error

References

American Economics Association. EconLit [internet]. Available at: http://web.ebscohost.com/ehost/search.

Bartels B. Beyond "fixed versus random effects": A framework for improving substantive and statistical analysis for panel, time-series cross-sectional, and multilevel data. Working Paper 2008.

Bell A, Jones K. Explaining fixed effects: random effects modelling of time-series crosss-sectional and panel data. Working Paper 2012. http://polmeth.wustl.edu/media/Paper/FixedversusRandom_1_2.pdf.

Clark TS, Linzer DA. Should I use fixed or random effects? Working Paper 2013. http://polmeth.wustl.edu/media/Paper/ClarkLinzerREFEMar2012.pdf.

CSA Illumina. PAIS International [internet]. Available at: http://search.proquest.com/pais/advanced?accountid=14784.

Debarsy N. The Mundlak approach in the spatial Durbin panel data model. Working Paper 2012. http://www.uclouvain.be/cps/ucl/doc/core/documents/Debarsy.pdf.

Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models. 1st ed. Cambridge University Press; December 18, 2006.

Greene WH. Econometric Analysis. 6th ed. Prentice Hall; August 17, 2007.

Hahn J, Ham J, Moon HR. Test of random versus fixed effects with small within variation. Economics Letters 2011; 112; 293–297.

Hausman J. SPECIFICATION TESTS IN ECONOMETRICS. Econometrica 1978; 46.

Kennedy P. A Guide to Econometrics. 5th ed. The MIT Press; August 1, 2003.

Kravdal Ø. The importance of community education for individual mortality: a fixed-effects analysis of longitudinal multilevel data on 1.7 million Norwegian women and men. Journal of Epidemiology and Community Health 2010; 64; 1029–1035.

Leyland AH. No quick fix: understanding the difference between fixed and random effect models. Journal of Epidemiology and Community Health 2010; 64; 1027–1028.

Mundlak Y. On the pooling of time series and cross section data. Econometrica 1978; 46.

National Library of Medicine (US). PubMed health [internet]. Bethesda, MD. Available at: http://www.ncbi.nlm.nih.gov/pubmed.

Setodji CM, Shwartz M. Fixed-effect or random-effect models: what are the key inference issues? Medical Care 2013; 51; 25–7.

Snijders TAB, Bosker R. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. Second Edition. SAGE Publications Ltd; December 6, 2011.

Wooldridge JM. Econometric Analysis of Cross Section and Panel Data. 1st ed. The MIT Press; October 1, 2001.

Wooldridge JM. Correlated random effects models with unbalanced panels. Working Paper 2010. http://econ.msu.edu/faculty/wooldridge/docs/cre1_r4.pdf.

Tables and Figures

Table 1: Dimensions adjusted for simulation

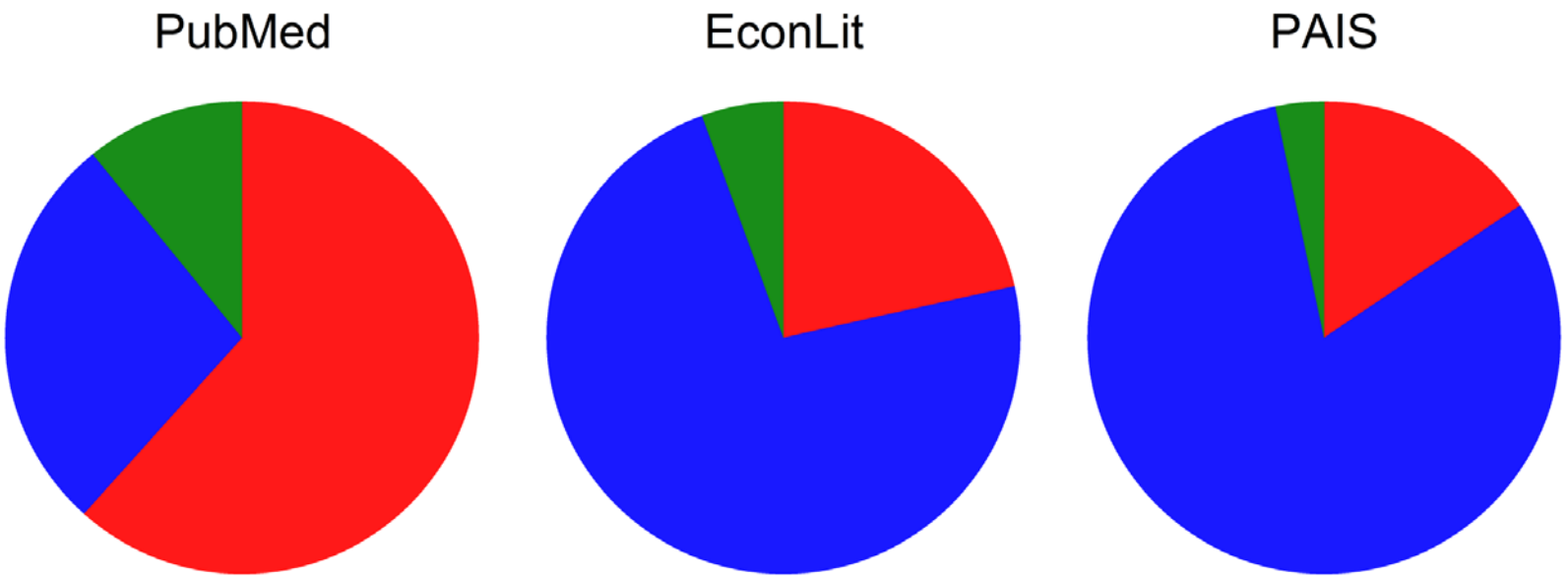
Input	Baseline value	Adjusted values
Number of groups:	$J \in (10, 50, 100)$	$J \in (10, 50, 100)$
Number of observations per unit:	$N \in (5, 10, 50)$	$N \in (5, 10, 50)$
Correlation between fixed effect and explanatory variable:	$\rho \in (0.00, 0.10, \ldots 0.70)$	$\rho \in (0.00, 0.10, \ldots 0.70)$
Correlation between explanatory variable and residual:	$\psi = 0$	$\psi \in (0.0, 0.2)$
Variation in explanatory variable:	$\sigma_x^2 = 1$	$\sigma_x^2 \in (0.5, 1.0, 2.0)$
Share of the variation in explanatory variable that is within unit:	$\tau = 0.50$	$\tau \in (0.10, 0.25, 0.50, 0.75, 0.90)$
Share of the variation in outcome that is due to residual:	$\pi = 0.50$	$\pi \in (0.10, 0.25, 0.50, 0.75, 0.90)$
Coefficient on independent variable:	$\beta = 1$	$\beta = 1$
Autocorrelation between residual and previous residual:	$\upsilon = 0$	$\upsilon \in (0,0.2)$

Table 2: Correlation between WB MSE minus FE MSE and simulation input parameters

	J	N	ρ	σ_x^2	τ	ψ	π
Correlation	0.1716*	0.1682*	-0.0661*	0.034	0.1738*	-0.0613*	-0.2554*
p-value	0.0000	0.0000	0.0000	0.0128	0.0000	0.0000	0.0000

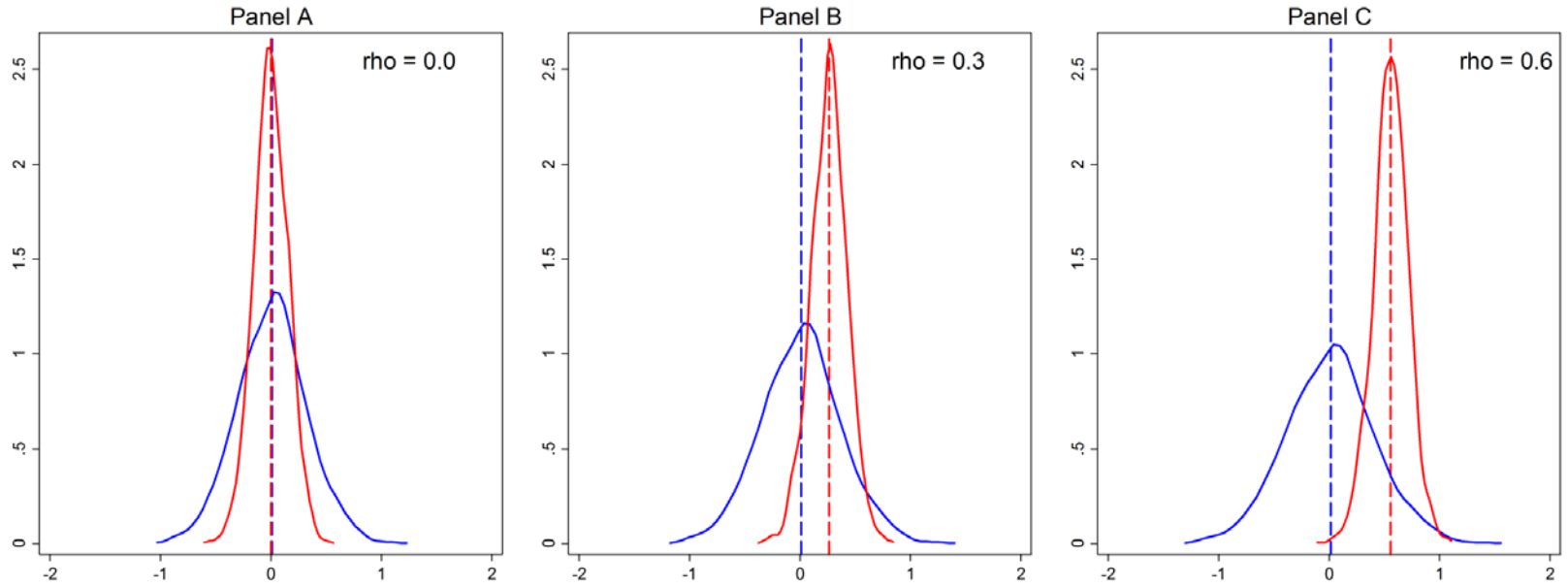
Pairwise correlation. Asterisk suggests the correlation is significant at the 99% confidence level.

Figure 1: Prevalence of random- and fixed-effects in health, economics, and political science literature



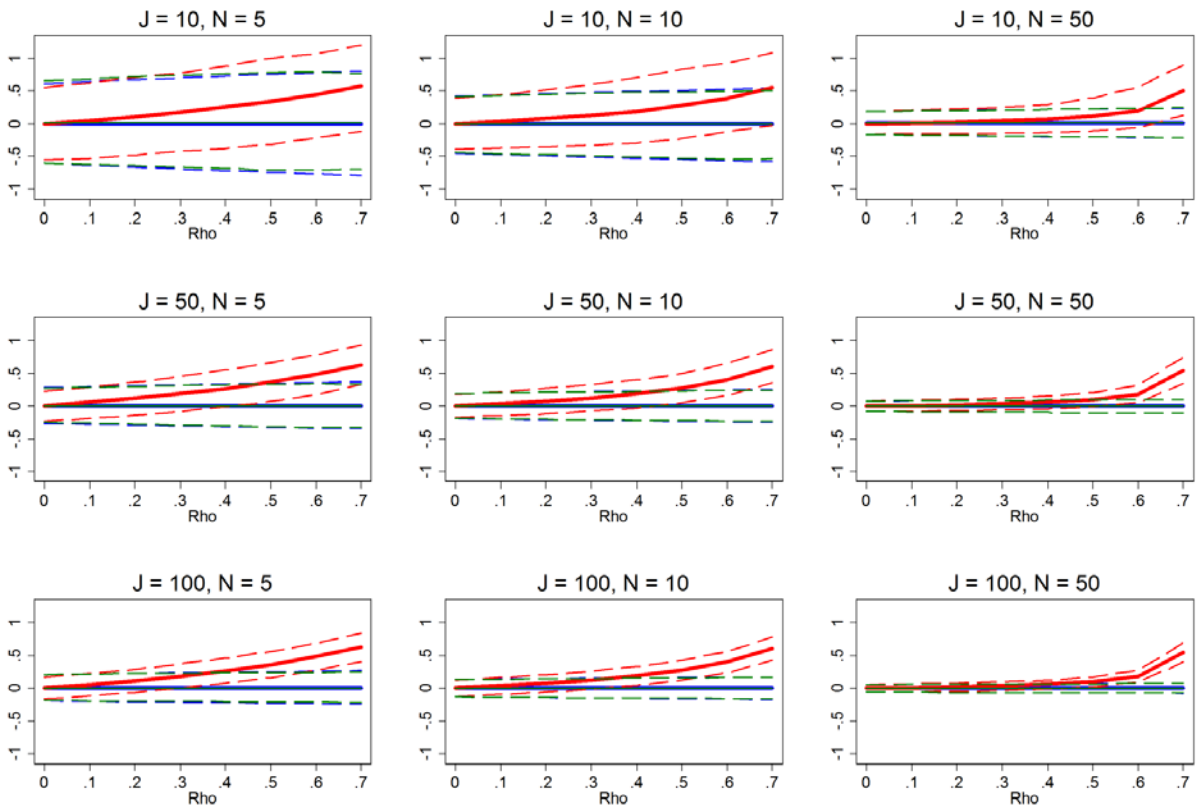
Each archive was searched for the terms "random effects" or "random effect" and "fixed effects" or "fixed effect" present in abstracts. Papers that also used the term "meta" in the abstract were not included in to avoid including meta-analyses which is a very specific use of RE and FE estimation. PubMed is database archiving life science and biomedical abstracts and references, and is primarily drawn from the MEDLINE database. EconLit is also an archiving database, is published by the American Economic Association and focuses on economics literature. PAIS is the Public Affairs Information Service International database and archives references focusing on public affairs.

Figure 2: Distribution of errors



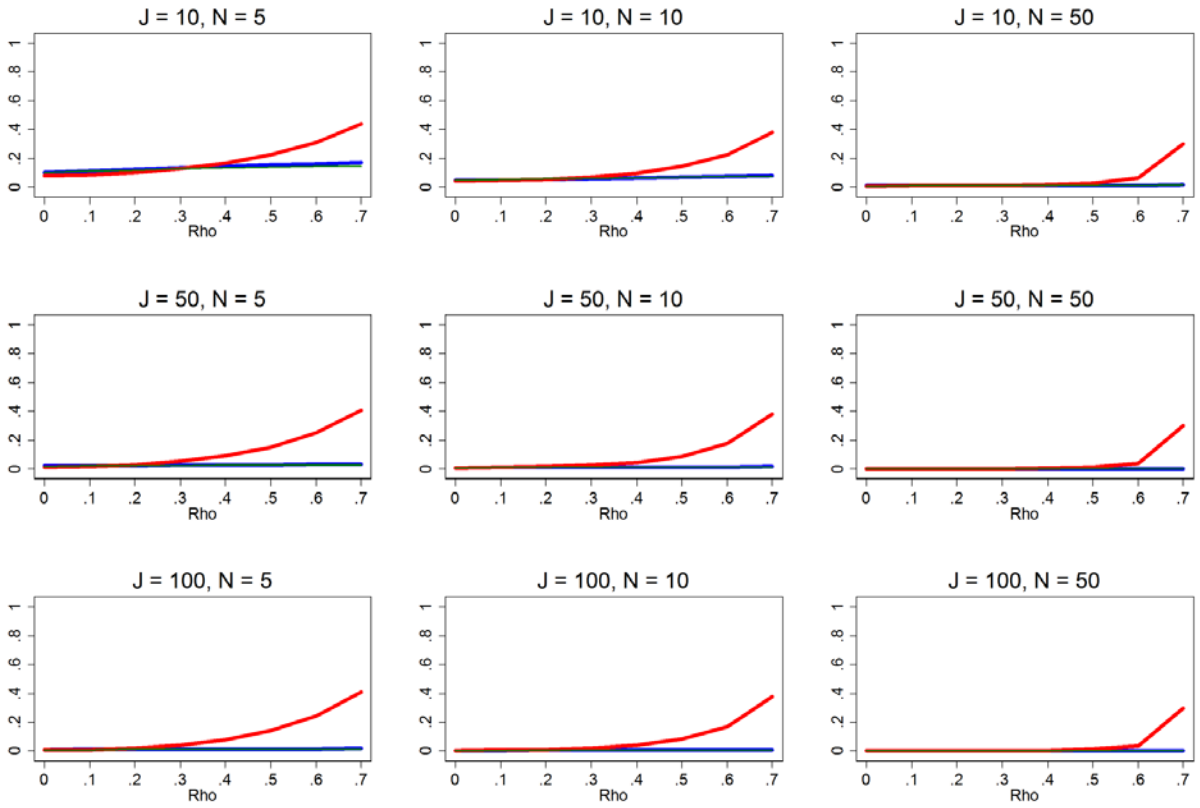
Red lines show the distribution of errors from RE estimation, while the blue lines show the distribution of errors from FE estimation. Each panel shows the correlation between the explanatory variable and the group-level effect set to a different value (0.0, 0.3, 0.6), increasing from left to right. Simulation based on the correct specification of a model with 50 groups and 10 observations per group. 50% of the variation of the outcome variable is explained by residual, while only 10% of the variation in the explanatory variable is within groups.

Figure 3: Distribution of errors of estimated marginal effects at baseline specification



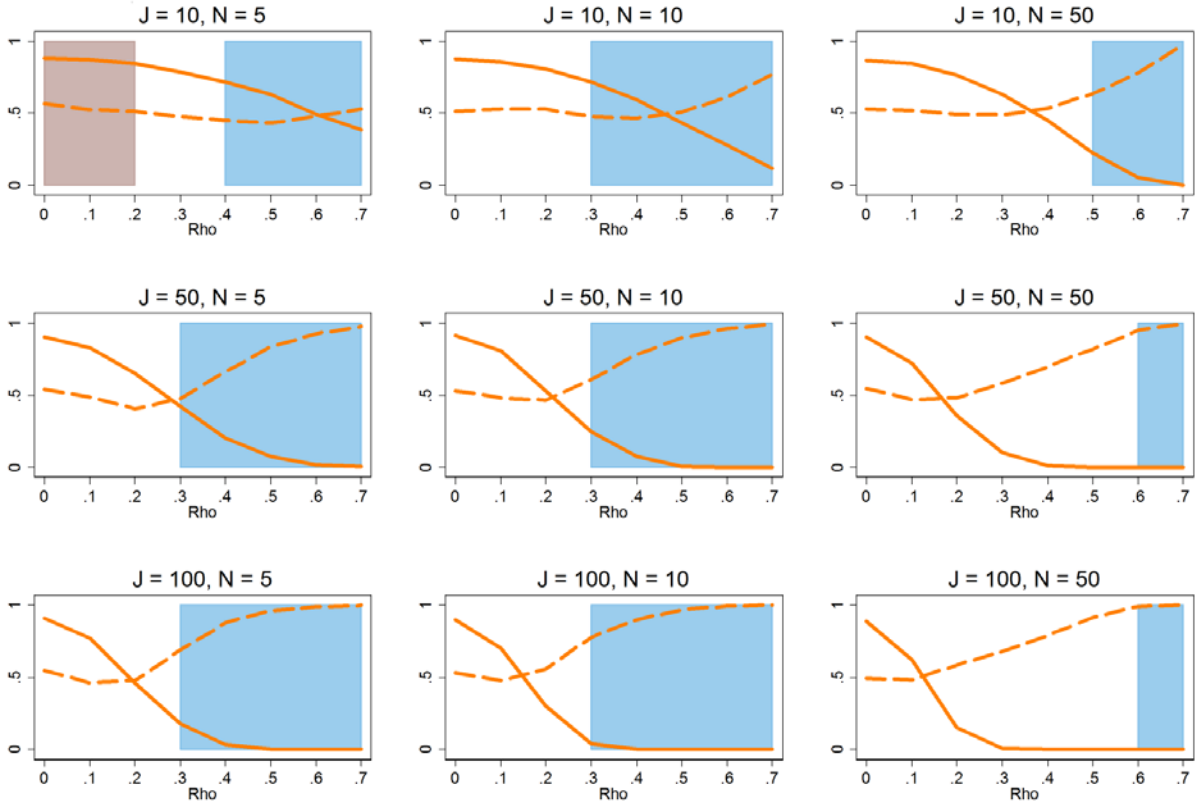
The solid red line show the mean error in the marginal effect estimates from RE estimation, while the dashed red lines show the 95% range of the RE estimation errors. The solid blue line and dashed blue lines show that mean and 95% range of the errors from FE estimation. The solid green line and dashed green lines show that mean and 95% range of the errors from the WB approach. All simulation inputs are baseline.

Figure 4: MSE of marginal effect estimates at baseline



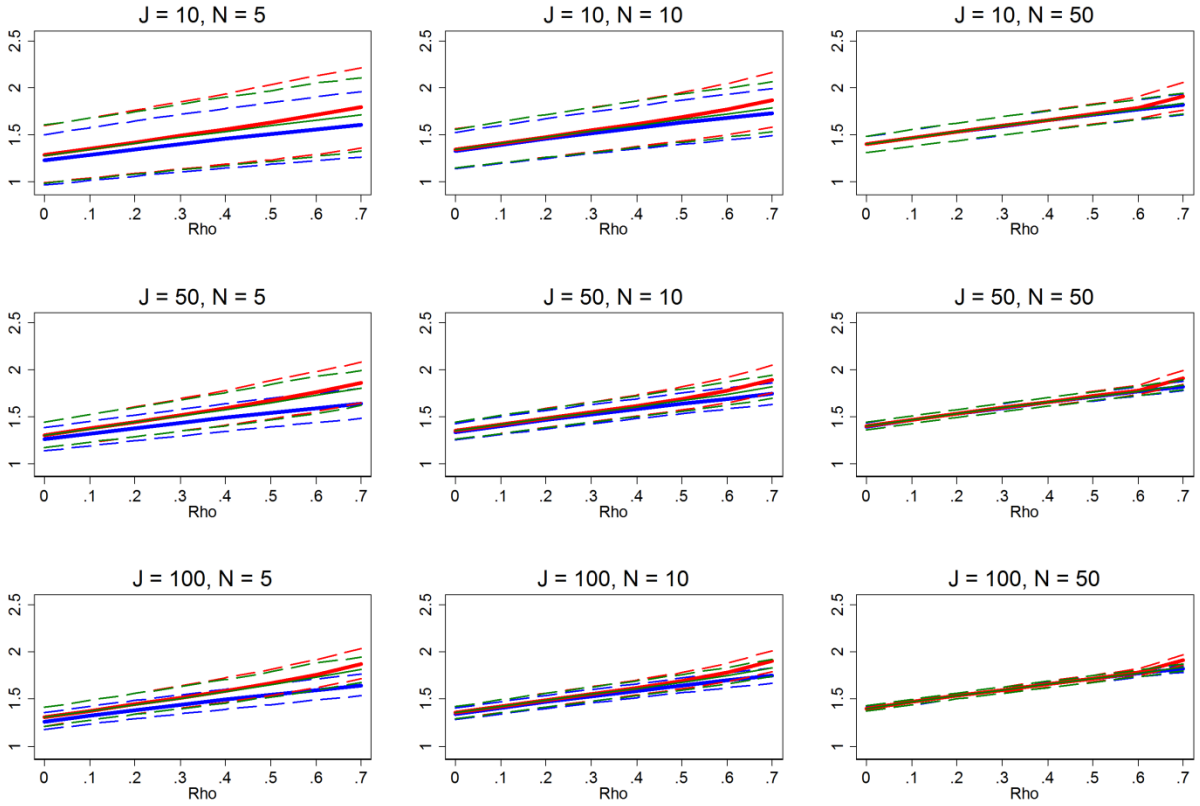
The red line shows the MSE from the errors in the marginal effect estimates from RE estimation. The blue and green lines shows the same for FE estimation and WB approach, respectively. All simulation inputs are baseline.

Figure 5: Hausman test



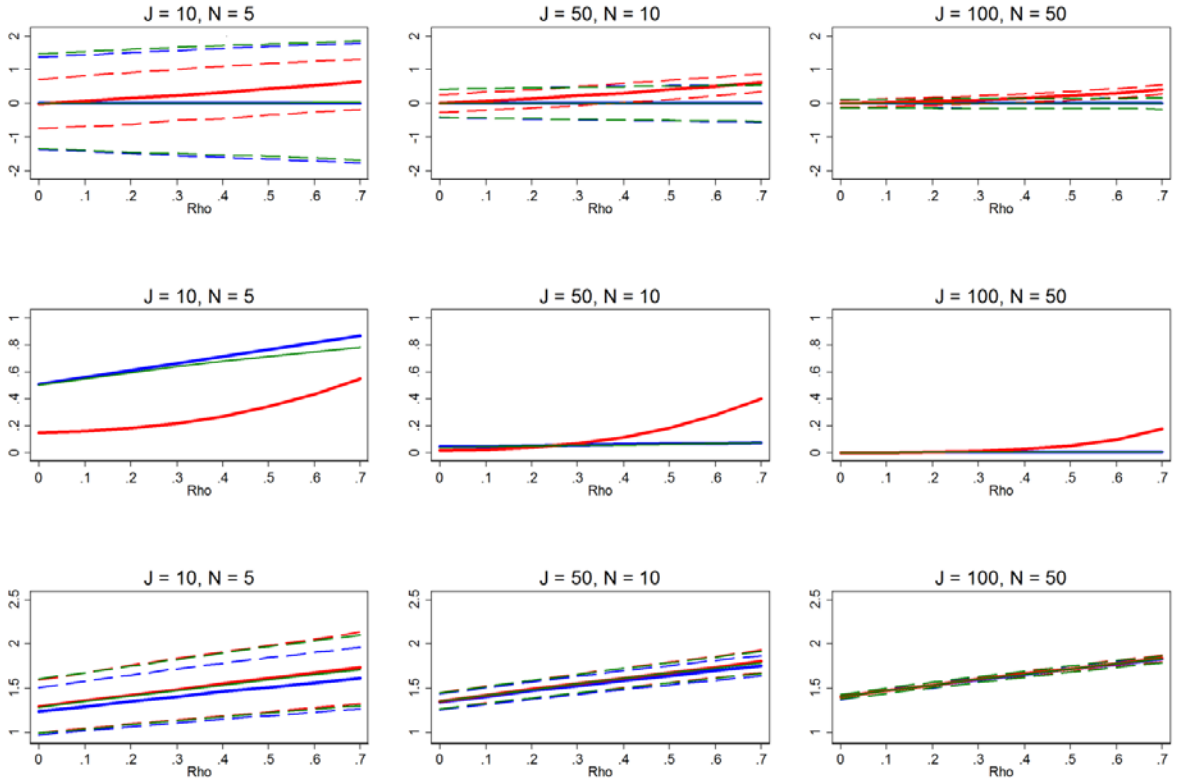
The solid green lines show the share of the simulations that the Hausman test did not reject at the 90% confidence level the null hypothesis that both RE and FE estimation are consistent. Conventional wisdom is that this is a suggestion that the researcher should use RE estimation as it is efficient. The dashed green line shows the share of the simulations that the Hausman test suggested the estimator with smaller absolute error. The red background indicates where the RE estimator is MSE-preferred, while the blue background indicates regions where the FE estimator is MSE-preferred. The white regions indicate that the difference between the MSE of the two estimators is trivial. All simulation inputs are baseline.

Figure 6: Distribution of RMSE from predicted outcomes



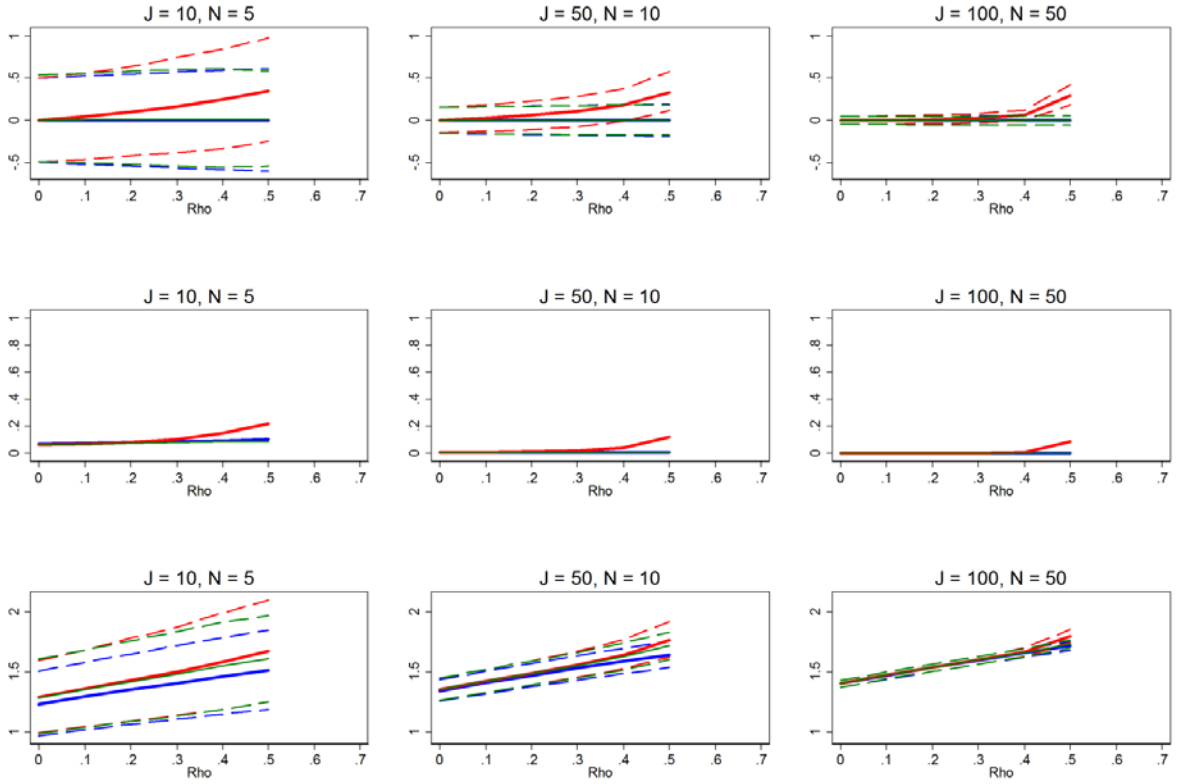
The red lines show the mean and 95% confidence interval of the RMSE derived from the fitted values using RE estimation. Each combination of inputs has 1,000 simulations completed and each receives its own RMSE based on the errors of the fitted values. The blue lines show the mean and 95% range of the RMSE acquired using the FE estimator. The green lines show the mean and 95% range of the RMSE from the WB approah. All simulation inputs are baseline.

Figure 7: Significant between-group variation relative to within-group variation



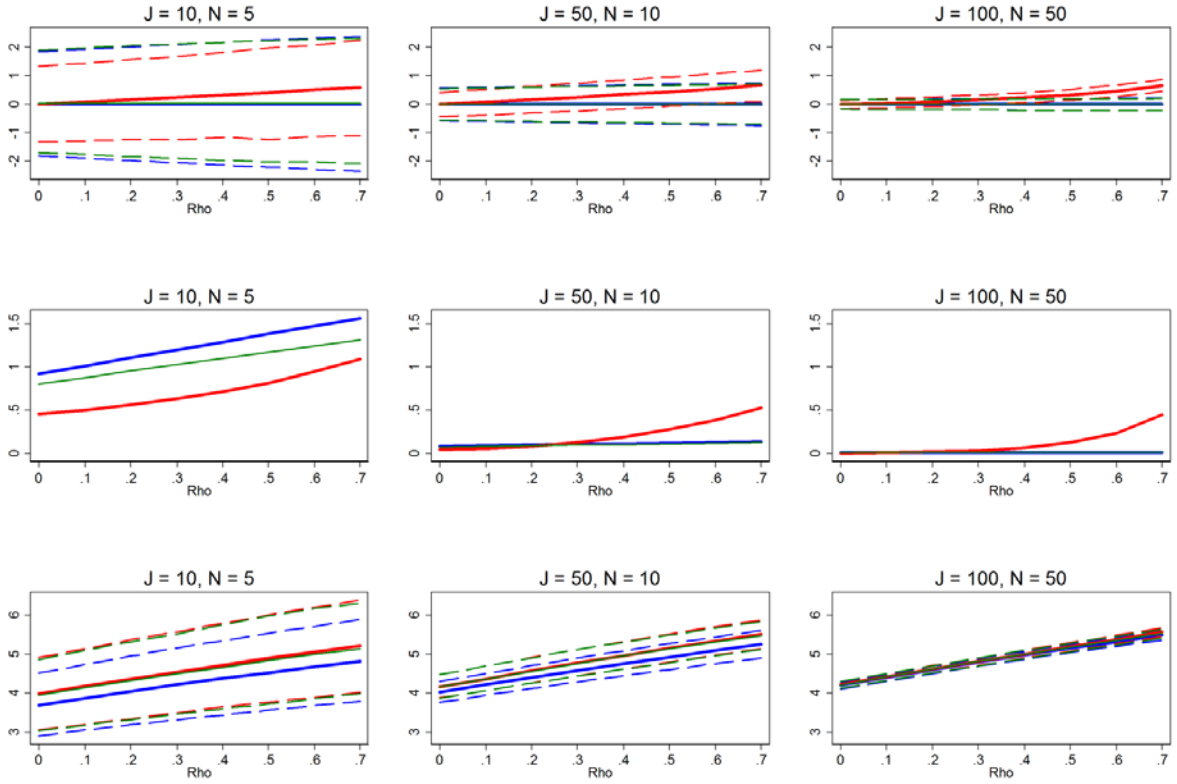
Row 1 can be interpreted like Figure 3, and shows the distribution of the errors in marginal effects estimates from the RE estimation (red), FE estimation (blue), and WB approach (green). Row 2 can be interpreted like Figure 4, and shows MSE associated with the RE estimation (red), FE estimation (blue), and WB approach (green) errors. Row 3 can be interpreted like Figure 6, and shows the distribution of the RMSE from the fitted values estimated using RE estimation (red), FE estimation (blue), and WB approach (green). The between-group variation is set to 0.9, while the within-group variation is 0.1. All other simulation input parameters are set to baseline.

Figure 8: Significant within-group variation relative to between-group variation



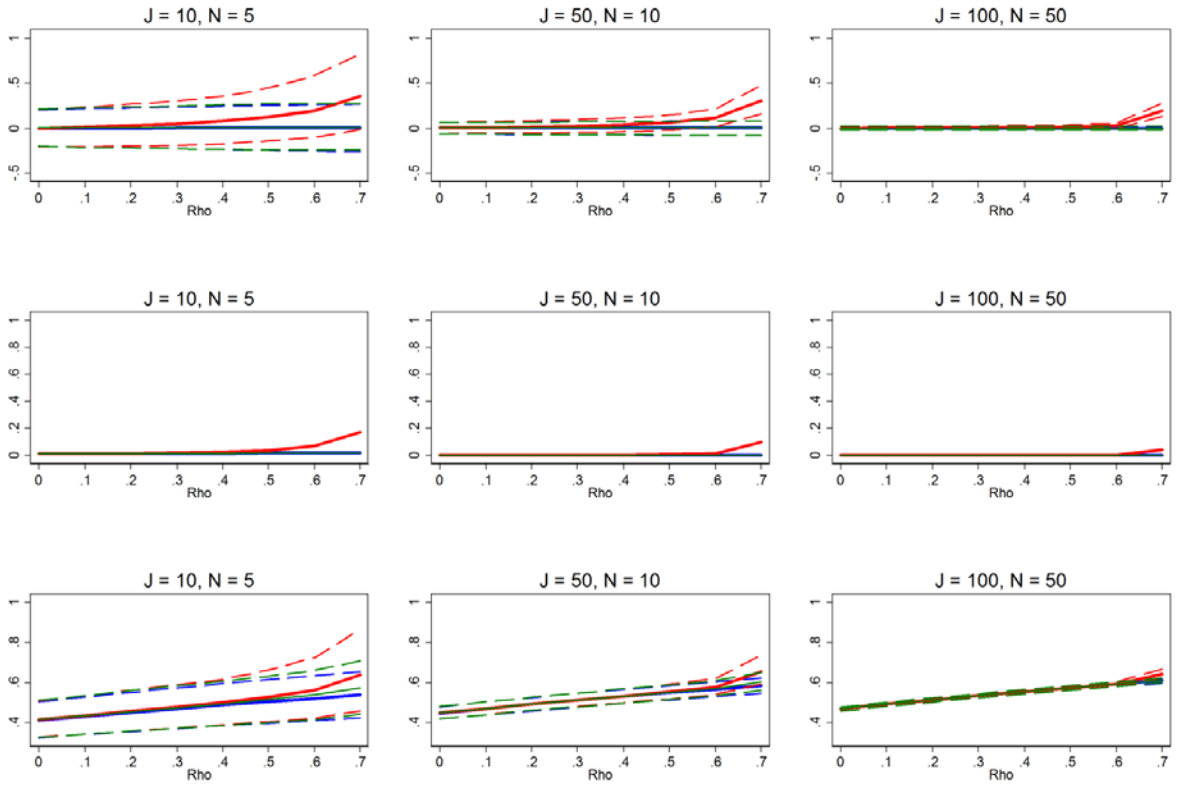
Row 1 can be interpreted like Figure 3, and shows the distribution of the errors in marginal effects estimates from the RE estimation (red), FE estimation (blue), and WB approach (green). Row 2 can be interpreted like Figure 4, and shows MSE associated with the RE estimation (red), FE estimation (blue), and WB approach (green) errors. Row 3 can be interpreted like Figure 6, and shows the distribution of the RMSE from the fitted values estimated using RE estimation (red), FE estimation (blue), and WB approach (green). The within-group variation is set to 0.75, while the between-group varition is 0.25. All other simulation input parameters are set to baseline.

Figure 9: Poorly fit model that explains only a small portion of the outcome variable's variance



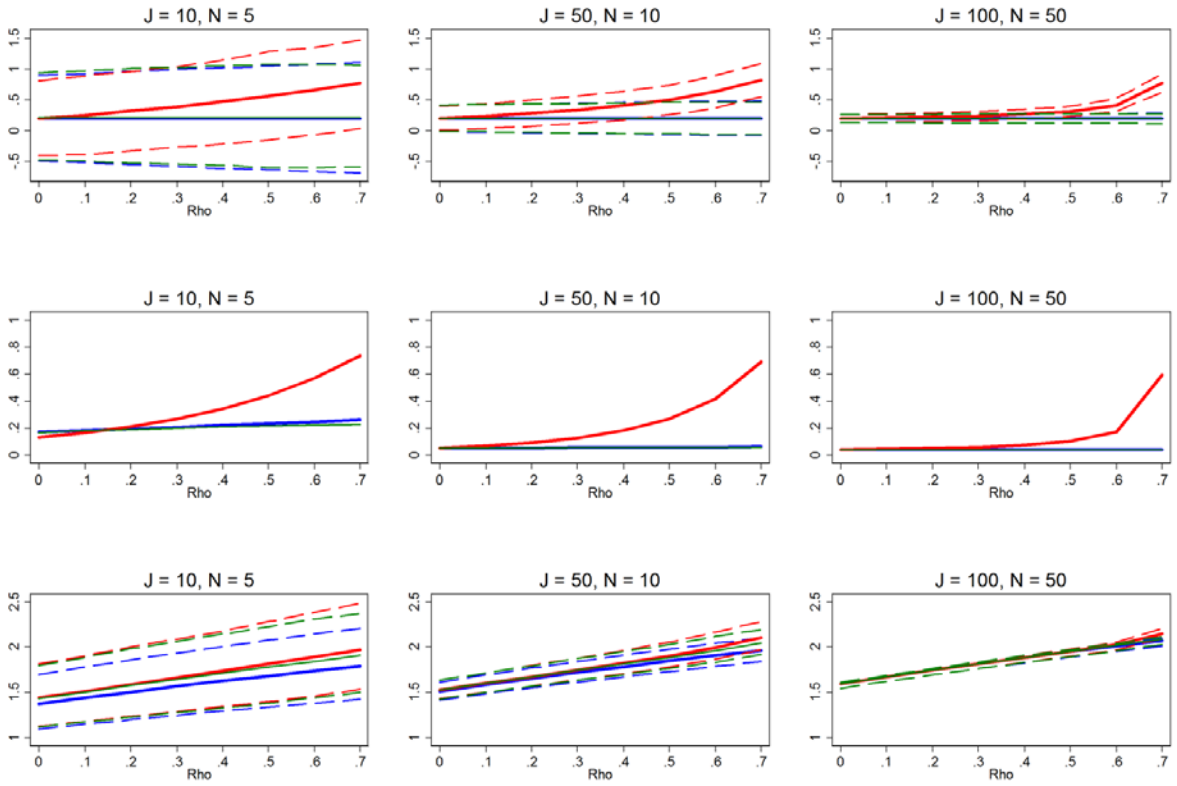
Row 1 can be interpreted like Figure 3, and shows the distribution of the errors in marginal effects estimates from the RE estimation (red), FE estimation (blue), and WB approach (green). Row 2 can be interpreted like Figure 4, and shows MSE associated with the RE estimation (red), FE estimation (blue), and WB approach (green) errors. Row 3 can be interpreted like Figure 6, and shows the distribution of the RMSE from the fitted values estimated using RE estimation (red), FE estimation (blue), and WB approach (green). The variance of the residual is set such that it explains 90% of the variation of the outcome variable. All other simulation input parameters are set to baseline.

Figure 10: Well fit model that explains a significant portion of the outcome variable's variance



Row 1 can be interpreted like Figure 3, and shows the distribution of the errors in marginal effects estimates from the RE estimation (red), FE estimation (blue), and WB approach (green). Row 2 can be interpreted like Figure 4, and shows MSE associated with the RE estimation (red), FE estimation (blue), and WB approach (green) errors. Row 3 can be interpreted like Figure 6, and shows the distribution of the RMSE from the fitted values estimated using RE estimation (red), FE estimation (blue), and WB approach (green). The variance of the residual is set such that it explains only 10% of the variation of the outcome variable. All other simulation input parameters are set to baseline.

Figure 11: Misspecified model



Row 1 can be interpreted like Figure 3, and shows the distribution of the errors in marginal effects estimates from the RE estimation (red), FE estimation (blue), and WB approach (green). Row 2 can be interpreted like Figure 4, and shows MSE associated with the RE estimation (red), FE estimation (blue), and WB approach (green) errors. Row 3 can be interpreted like Figure 6, and shows the distribution of the RMSE from the fitted values estimated using RE estimation (red), FE estimation (blue), and WB approach (green). The correlation between the explanatory variable and the residual is set to 0.2. All other simulation input parameters are set to baseline.