Supplementary Appendix

Table of Contents

1. Statistical Appendix

Appendix S1: Examining the role of state capacity

Choice of state capacity indicator

To evaluate whether the relationships between tariff changes and health expenditure depend on a country's state capacity we use the government effectiveness (GE) component of the World Government Indicators (Kaufmann, Kraay and Mastruzzi, 2011). We use the GE as it is the most widely available proxy of the elements of state capacity that are of theoretical interest here. In particular, we aim to capture aspects of a state's bureaucracy and administration that influence the state's ability to effectively define, enforce, and administer non-trade taxes. This depends on whether states can reach their populations, collect and manage information, possess trustworthy agents to manage the revenue, and can ensure policies area adequately adhered to (Besley, 1995; Hanson and Sigman, 2013). These abilities are present where state bureaucracies are staffed by a well-trained civil service that are able to clearly formulate and credibly commit to a policy, effectively collect and administer reforms, distribute clearly formulated guidelines, and monitor compliance (Pomeranz and Vila-Belda, 2019).

There are additional dimensions of a state capacities that are not captured in the GE index and could also influence a government's ability to collect and administer tax revenue. For example, to achieve policy goals, including the collection of revenue, a state may need a monopoly on the legitimate use of the military within its territory and possess the force necessary to contain threats throughout its territory – or at least convince its rivals that this is the case (Besley and Persson, 2013). While these 'coercive' state capacities are not the only way to maintain order and evoke compliance from the population, they have historically represented a key aspect of the ability of states to survive and implement policies (Ardant, 1971). However, we proxy the bureaucratic and administrative aspects a state's capacity – rather than its ability to use force – as

examining the former can provide insights into what are arguably the most politically acceptable and actionable means of improving tax administration.

We also considered a number of alternative indicators of state capacity. One alternative is the International Country Risk Guide's (ICRG) Bureaucratic Quality rating. However, this is a comparatively narrow indicator of the quality of bureaucracy compared with the GE index as it measures fewer relevant components. Previous studies have also observed that the ICRG measure is more prone to measurement errors compared with the GE index due to analyst perceptions of economic or social outcomes, rather than bureaucratic quality per se (Rauch and Evans, 2000; Hendrix, 2010). In contrast, Arndt and Oman (2006) pointed out that the WGIs – including the GE index – are "probably the most carefully constructed indicators" (Arndt, 2006).

Another possible set of proxies for state capacity are the 'Varieties of Democracy', or 'V-Dem', measures. These are designed to capture whether electoral competition exists, whether elections are free and fair, and to what extent political and civic organizations can operate unrestrained (Lindberg *et al.*, 2014). Higher scores on the V-Dem index capture the presence of a well-functioning democracy. Democratic institutions may foster state capacity; for example, the process of carrying out elections may promote an expanded and deepened presence of the state over the national territory (Carbone and Memoli, 2015). The possible causal link between democratic institutions and state capacity suggests that indicators of democracy that are included in the V-Dem index may serve as suitable proxies for state capacity.

Yet, the very plausibility of a causal link between these two state characteristics also illustrates that they can (and arguably should) be seen as distinct concepts. Democratisation concerns the extent to which potentially unfettered political power is contestable and subject to checks and balances. That is what the V-Dem index is designed to measure, and relevant sub-indices are chosen on that basis. In contrast, state capacities concern the technical and administrative ability

of a government to construct and implement policies throughout a territory (Andersen, Møller and Skaaning, 2014). Hence, we did not use the V-Dem index as it captures aspects of a state's politics or governance that were distinct from my theoretical interests specifically in state capacities that influence domestic tax administration.

Another possible set of measures are the Quality of Government (QoG) indicators. The measure may seem related to aspects of state capacity, particularly where it captures competition over civil service appointments (Rothstein & Teorell, 2008; Teorell, 2016). However, documents describing the QoG indicators and codebook note that this indicator is primarily concerned with "the impartiality of institutions that exercise government authority" (Rothstein and Teorell, 2008; Teorell *et al.*, 2016). This is again distinct from state capacity, which concerns the ability to design and implement policies throughout a territory, rather than the extent to which this process is free from political interference (Andersen, Møller and Skaaning, 2014). Although the civil service recruitment is relevant to state capacity as government staff appointed via political processes may not be appropriately trained in order to effectively implement policy (Xu, 2018), the degree of impartiality of recruitment is not synonymous with these capacities. Indeed, staff selected by politicians can occasionally be effective in carrying out government decisions and implementing policy. This is because their closer and more amicable relationships with politicians can help create political stability and reduce disagreement (Arriola, 2009; Grindle, 2012).

Models

To statistically assess the role of state capacity we re-estimate the models specified in Equation 1 in the main paper with an additional interaction term to assess variation according to a country's GE score. We estimate Equation 2 below and then compute marginal effects using the margins command in R (Leeper, 2017). This calculates the Average Marginal Effect of a 1% tariff reduction on public health expenditure at a range of specified GE scores:

Equation 2. $HXP_{it} = \beta_0 + \beta_1 T_{it} + \beta_2 GE_{it} + \beta_3 T_{it} x GE_{it} + \beta_4 X_{it} + \alpha_i + \Upsilon_t + \varepsilon_{it}$

Here the variable GE_{it} is country i's score on the GE index in year t and $T_{it}xGE_{it}$ is an interaction of the GE score and a country's tariff rate. All other variables and coefficients are per Equation: the outcome variable, HXP_{it}, is health spending (either public or private, per capita or % GDP) in country *i* in year *t*. T_{it} is the weighted average tariff rate. β_0 is the intercept and α_i in Equation 1 is a vector of country fixed-effects which account for time-invariant, unmeasurable characteristics which may influence a country's tariff rate and health expenditure. We also incorporate year fixed effects, Υ_t, to control for common external shocks affecting tariff policies and health spending across all countries. X_{it} is a vector of time-varying controls with coefficients in the vector β_2. We control for GDP per capita, overseas development assistance (ODA), and the occurrence of war (see main text for rationale for these covariates). In robustness checks we incorporate additional possible predictors and covariates (see Appendix S3). ε_{it} is the error term. Robust standard errors were clustered by country.

Appendix S2: Mediation analysis

We used a Sobel test and Baron and Kenny's four-step procedure to examine whether any heterogeneous associations between tariffs and health expenditure according to a country's GE score could be explained by corresponding differences in government tax revenues (Baron and Kenny, 1986). The steps are as follows:

> Step 1: Test for an association between the explanatory variable, the tariff rate, and the outcome variable, public health expenditure, and test whether this association varies according to a country's GE score.
>
> Step 2. Test for an association between the explanatory variable, the tariff rate, and the potential mediator, per capita tax revenue, and test whether this association varies according to a country's GE score.
>
> Step 3. Test for an association between the potential mediator, per capita tax revenue, and the outcome variable, public health expenditure.
>
> Step 4. Test for an association per Step 1 (between the explanatory variable, the tariff rate, the outcome variable, public health expenditure, and whether this association varies according to a country's GE score) whilst also controlling for the potential mediator.

The goal of the first three steps is to establish whether relationships between the explanatory variable, outcome variable, and mediator actually exist. The logic of these steps is that it would be difficult to claim that the observed associations between were mediated by differences in tax revenue and corresponding GE scores if any of these relationships were insignificant. In step 4, if the tariff rate variable is attenuated or no longer significant when tax revenue is controlled, the finding supports partial or full mediation via tax revenue.

The main results of the paper confirm Step 1, i.e. there is a statistically significant association between the treatment (tariff *x* GE scores interaction) and outcome (public health expenditure). Panel A in Figure S5 visualises the results from Step 2. We find that every 1% reduction in tariffs is associated with a reduction in per capita tax revenue among countries with a GE score below the 10th percentile (Average Marginal Effect of a 1% tariff reduction, AME_{tariff} = $-10.7; 95% CI: -20.3 to -0.95), no change in tax revenue among countries with scores in the 10th-30th percentile (AME_{tariff} = $-3.23; 95% CI: -10.6 to 4.15), and a rise in tax revenue among countries with scores above the 30th percentile (AME_{tariff} = 12.1; 95% CI: 5.08 to 19.21).

In Step 3 we further find that every $1 increase in per capita government tax revenue was associated with a $0.1 (95% CI: 0.08 to 0.13) increase in per capita government spending on health after adjusting for possible covariates. Panel B in Figure S5 visualises the results from Step 4. In this model the tariff *x* government effectiveness interaction was substantially attenuated and no longer statistically significant at the 1%, 5% or 10% significance thresholds. Taken together, these results suggest partial or full mediation via tax revenue.

We also conducted a Sobel test to determine whether there is a statistically significant effect of the Tariff x GE score variable on government health spending as mediated through tax revenues (Mustillo, Lizardo and McVeigh, 2018). The result of this test rejects the null hypothesis of no mediation ($z = -2.47$, $p = 0.007$), suggesting that tax revenues at least partially mediate the relationships between tariffs, GE scores, and government health spending.

Appendix S3: Additional analyses

We conducted a series of additional tests to evaluate the consistency of our results in alternative sample and model specifications. These are summarised in Figure 3 in the manuscript and described briefly in the accompanying text. Here we first provide details of these procedures, including the rationale for each robustness check. These are grouped according to the type of bias or modelling sensitivity that each test addresses. In the subsequent section we summarise the results from all tests (see also Figure 3 in the main text). Table S1 provides the data sources and measurement of additional variables included in these analyses.

Non-random assignment

First, a country's tariff rate or 'treatment' level is non-randomly assigned. Our statistical models implicitly account for variables that predict a country's 'treatment' level and might bias the estimated coefficients by adjusting for variables that are confounders of the treatment-outcome association (Morgan and Winship, 2007). However, the characteristics of countries with different tariff rate 'treatment' levels may differ in ways that undermine their validity as a counterfactual when estimating the tariff effect. One way of addressing this issue is to re-weight observations in order to reduce differences in the characteristics which predict a country's tariff level across countries with different tariff rates (Abadie and Cattaneo, 2018). I therefore estimated an additional model that uses the non-parametric Covariate Balancing Generalised Propensity Score (npCBGPS) weighting procedure developed by Fong et al. (main text Figure 3, Model 1) (Fong, Hazlett and Imai, 2018). This uses an algorithm to search for a set of country-weights that, when applied to the data, minimise the correlation between tariff rate predictors and the tariff rate. These weights were then incorporated into the fixed-effects regression model. Figure S6 shows that this weighting procedure reduces the correlation between covariates and the tariff rate 'treatment' level. Model 1 in Figure 3 (main text) shows the results from this test.

Alternative specifications

Next, we incorporated a series of additional possible predictors of public health spending as controls in our models (main text Figure 3, Models 2-8). In the first set of models we evaluate whether our results are robust when accounting for aspects of a country's governance that may be correlated with government effectiveness and alternatively influence whether tariff reductions translate into a rise or fall in public health expenditure. This includes electoral accountability and other democratic indicators – which can influence demand for health expenditure – as well as the degree of corruption in a country, which can influence the efficacy tax collection and may divert resources away from health-systems (Franco, Álvarez-Dardet and Ruiz, 2004; Delavallade, 2006). To evaluate this we estimated additional models in which we introduced controls for a country's percentile in the 'voice and accountability' and 'control of corruption' sub-components of the World Governance Indicators, and a country's score on the Polity regime authority spectrum (Marshall, Jaggers and Gurr, 2002). We estimate three separate models controlling for each of these indicators separately in each specification (main text Figure 3, Models 2-4), as well as a fourth model including all three additional indicators simultaneously (main text Figure 3, Model 5).

We also estimate a model in which we adjusted for demographic variables which are expected to be associated with higher health expenditures by increasing demand for health-care and services: the fertility rate and the combined share of the population aged under 15 and over 65 as they (main text Figure 3, Model 6) (Nooruddin and Simmons, 2009). In another model we included a control for the level of urbanisation, since large clusters urban dwellers can mobilize demands for additional health-care and services from governments, and cities also offer economies of scale (main text Figure 3, Model 7) (Alesina and Wacziarg, 1998; Baqir, 2002).

It is possible that tariff changes occur in the context of international political integration – including Free Trade Agreement ratification or United Nations membership. These co-inciding

political processes may account for our results as they can influence health policy norms or policy space in ways that may encourage or discourage public health expenditure(Meyer *et al.*, 1997; Brady, Beckfield and Zhao, 2007; Barlow *et al.*, 2018). We therefore re-estimated our models including a control for a country's degree of international political integration using the political globalization sub-component of the KOF Globalization Index (main text Figure 3, Model 8).

Finally, we conducted a test in which we exclude official development assistance from the models, astrade openness can encourage aid and so this variable may mediate the association and attenuate the bias the true tariff coefficient (main text Figure 3, Model 9) (Helble, Mann and Wilson, 2009; Richiardi, Bellocco and Zugna, 2013).

Alternative tariff measures

Our trade-weighted measure of the average tariff rate has certain limitations. For example, a country may have relatively little trade overall because it has prohibitive tariffs (i.e., tariffs set so high as to eliminate imports) in many import categories. In this case a country would have a large share of trade in a few import categories with relatively low tariffs. The trade-weighted average tariff would be relatively low in this case. This would result in a low average tariff being reported for a protectionist country. Alternatively, a country may apply lower tariff rates to WTO members, who are subject to a country's Most-Favoured-Nation (MFN) tariff, compared with non-WTO members. However, these tariffs may only be moderately lower than the tariffs on non-WTO members and so have little effect on who a country trades with. In this case, the weighted average tariff rate might also capture tariffs on non-WTO members, resulting in a slightly higher tariff rate being reported than is appropriate given the lower tariffs on WTO members.

To address the possible issues outlined above we evaluated whether our results were consistent when substituting our original tariff estimate for alternative tariff measures: the simple mean

average tariff (main text Figure 3, Model 10) and the weighted-mean MFN tariff (main text Figure 3, Model 11).

Budget cycles and observer bias

It may take time for changes in tariffs and tax revenues to influence government budgets and health expenditures. For example, many governments set budgets in response to tariff and tax revenue changes in the previous - rather than current – year. We therefore conducted an additional robustness check in which we re-estimate my models with the explanatory variables lagged by 1-year (Figure 3, Model 12).

In addition, those compiling the GE index, or the surveys used to construct it, may hold the view that state capacity is integral to development (Stubbs, King and Stuckler, 2014). This raises the possibility that observers from, or selected by, organisations that hold such a view may code countries that are performing better economically with higher scores on the GE index or its constituent indicators. Although coding bias may not be deliberate, it could still occur as a sub-conscious result of exposure to information about country economic performance. As a result, we would expect GE scores to have upward bias in countries with higher GDP per capita or GDP growth. We therefore conducted an additional robustness test, following Stubbs et al. (2014), in which we control for GDP growth in the previous year when estimating my models (Figure 3, Model 13).

Influential observations

We evaluated the sensitivity of our results to potentially influential observations – including those with very high tariffs and low spending – by calculating the Cook's distance of each unit. Cook's distance (or 'D') is a commonly used measure of the influence exerted by each data point on the predicted outcome (Snijders and Berkhof, 2008; Van der Meer, Te Grotenhuis and Pelzer, 2010). Cook's D for each observation *i* measures the change in fitted values for all observations

with and without the presence of observation *i*. As a rule of thumb, cases are regarded as too influential if the associated value for Cook's D exceeds the cut-off value of 4/n. We re-estimated our models excluding cases with a Cook's D larger than 4/n to test whether their exclusion produced different findings (Figure 3, Model 14).

Robustness check results

Figure 3 in the main text shows that the results from each of the additional tests described above were consistent with our original specification: a 1% reduction in tariffs was associated with an increase in per capita public health spending of between $3.4 (95% CI: 1.19 to 5.62) and $8.1 (95% CI: 5.56 to 10.59) among countries with a government effectiveness score above the 30th percentile, no statistically identifiable change in public health care expenditure among countries with scores between the 10th and 30th percentile, and a reduction in public health expenditure of between $2.3 (95% CI: -4.85 to -0.30) and $5.3 (95% CI: -8.49 to -2.09) among countries with government effectiveness scores below the 10th percentile.

Associations with health expenditure as a share of GDP were slightly less robust but in most specifications tariff reductions were only associated with a rise in public health expenditure as a share of GDP among countries among countries with a government effectiveness score above the 30th percentile, and a reduction among countries with scores below the 10th percentile (Figure S9). Finally, results showing no association between tariff reductions, government effectiveness, and private health expenditure (measured per capita or as a share of GDP) were consistent in these alternative models (See Figure S7 and S8).

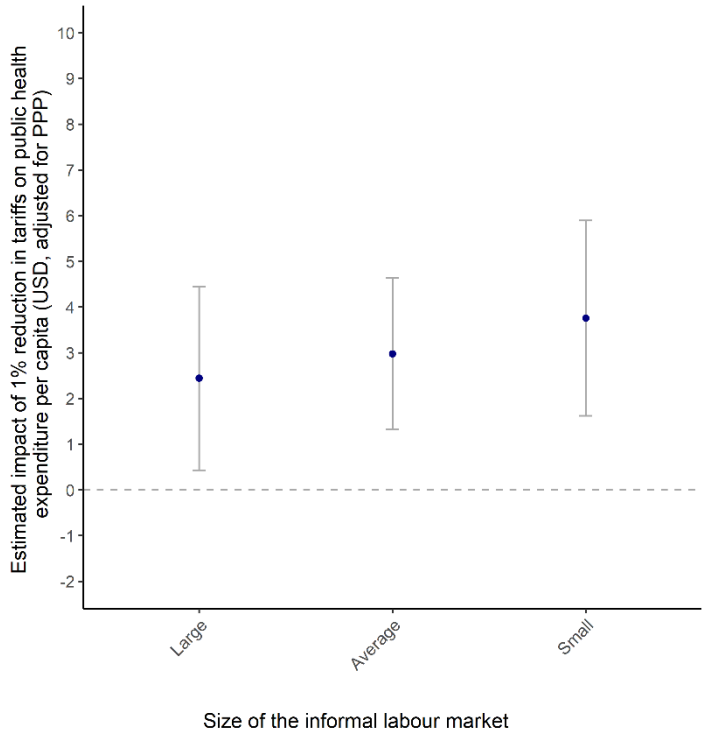
Informal labour market shares

The size of the informal labour market might create challenges in levying taxes and collecting adequate revenue to finance public health spending (Auriol and Warlters, 2005; Lagomarsino *et al.*, 2012). This suggests that tariff reductions may have divergent impacts on public health expenditure depending on the size of the informal labour market. Data pertaining to the size of the informal labour market is fragmentary. We were therefore unable to implement a full interaction test to asses variation in the association between tariff changes and government health spending according to the size of the informal economy in each country/ year. We nevertheless conducted a preliminary assessment by dividing countries according to whether the size of the informal sector is small or large, as follows.

First, we obtained data from the International Labour Organization showing the share of informal employment as a percentage of total employment, 1999-2016 (ILOSTAT, 2018). Data availability in each year ranged from just 1 country in 1999 to 33 countries in 2013. Second, we calculated the mean size of the informal labour market across all countries in each year and converted each country-specific value into standardized z-scores. Where data were missing for a particular country and/or year, we used the most recent' year's data to calculate z-scores. Third, we converted these z-scores into 3 categories corresponding to informal labour market shares i) within 1 standard deviation from the mean ('average informal labour market'), ii) at least one standard deviation larger than the mean ('large informal labour market), and iii) at least 1 standard deviation below the mean ('small informal labour market). Finally, we re-estimated our interaction models replacing the 'state capacity' variable with the new 'informal sector size' variable to assess variation in the association between tariff reductions according to whether a country had a large or small informal labour market. We used this model to calculate the Average Marginal Effect of 1% tariff reduction on per capita health expenditure across countries with

small, average, and large informal labour market sizes. Figure 1 plots the results from this additional test.

Figure 1. Average marginal effect of 1% tariff reduction on per capita health expenditure by informal labour market size



Notes: Labour market sizes correspond to z-scores of informal labour market employment as a share of the total population in the most recent year when data were available, as follows. Large: at least 1 standard deviation above the mean. Average: within one standard deviation from the mean. Small: at least 1 standard deviation above the mean.

The results in Figure 1 suggest that tariff reductions do not have divergent associations with public health expenditure depending on the size of the informal labour market as the AMEs are all comparable and the confidence intervals overlap substantially. Although these findings are not in line with the expectations outlined above, the fragmentary nature of the data preclude definitive conclusions.

Baseline tariff rate

There is some evidence that reducing already low tariffs may not deliver large increases in growth (Dhingra *et al.*, 2016). We might therefore expect countries with already-low tariffs to smaller increases in public health spending due to limited economic gains that provide resources for expanding health expenditure. Alternatively, as discussed in the main paper, many developing countries have difficulties in levying domestic taxes and have historically relied on trade taxes in order to finance public services (Cagé and Gadenne, 2018). Hence, the presence of low (or high) tariffs in a period before a tariff reduction may reflect the lack (or presence) of such difficulties. We therefore conducted an additional analysis to test whether the association between tariffs and public health spending depends on the tariff level in the previous year. The results are summarised in Figure 2 below.

Figure 2. Average marginal effect of 1% tariff reduction on per capita health expenditure by tariff rate in the previous year

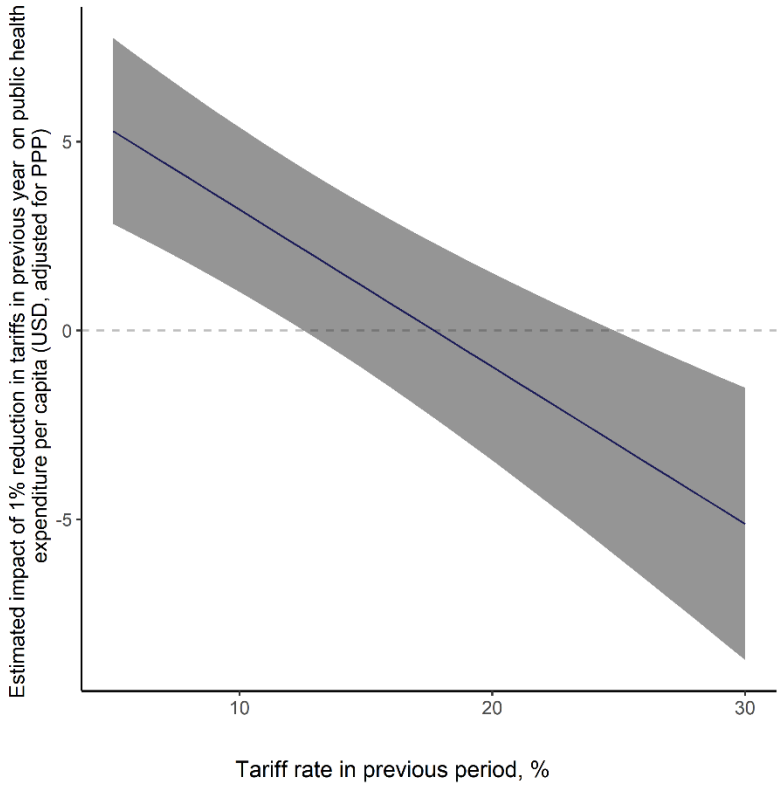


Figure 2 shows that the association between a 1% tariff reduction and public health spending was larger where tariffs were lower in the previous period. This suggests that where tariffs were lower before the reduction, tariff declines corresponded to a larger increase in public health spending.

References for Statistical Appendix

Abadie, A. and Cattaneo, M. D. (2018) 'Econometric Methods for Program Evaluation', *Annual Review of Economics*, 10(1), pp. 465–503. doi: 10.1146/annurev-economics-080217-053402.

Alesina, A. and Wacziarg, R. (1998) 'Openness, country size and government', *Journal of public Economics*. Elsevier, 69(3), pp. 305–321.

Andersen, D., Møller, J. and Skaaning, S.-E. (2014) 'The state-democracy nexus: conceptual distinctions, theoretical perspectives, and comparative approaches', *Democratization*. Taylor & Francis, 21(7), pp. 1203–1220.

Ardant, G. (1971) 'Histoire de l'impôt'. Fayard Paris.

Arndt, C. (2006) *Development Centre Studies Uses and Abuses of Governance Indicators*. OECD Publishing.

Arriola, L. R. (2009) 'Patronage and political stability in Africa', *Comparative Political Studies*. Sage Publications Sage CA: Los Angeles, CA, 42(10), pp. 1339–1362.

Auriol, E. and Warlters, M. (2005) 'Taxation base in developing countries', *Journal of Public Economics*. Elsevier, 89(4), pp. 625–646.

Baqir, R. (2002) 'Districting and government overspending', *Journal of political Economy*. The University of Chicago Press, 110(6), pp. 1318–1354.

Barlow, P. *et al.* (2018) 'Trade challenges at the World Trade Organization to national noncommunicable disease prevention policies: A thematic document analysis of trade and health policy space', *PLOS Medicine*. Public Library of Science, 15(6), p. e1002590.

Baron, R. M. and Kenny, D. A. (1986) 'The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations', *Journal of Personality and Social Psychology*, 51(6), pp. 1173–1182.

Besley, T. (1995) 'Property Rights and Investment Incentives: Theory and Evidence from Ghana', *Journal of Political Economy*, 103(5), pp. 903–937.

Besley, T. and Persson, T. (2013) 'Taxation and development', in *Handbook of public economics*. Elsevier, pp. 51–110.

Brady, D., Beckfield, J. and Zhao, W. (2007) 'The Consequences of Economic Globalization for Affluent Democracies', *Annual Review of Sociology*, 33(1), pp. 313–334.

Cagé, J. and Gadenne, L. (2018) 'Tax revenues and the fiscal cost of trade liberalization, 1792–2006', *Explorations in Economic History*, 70, pp. 1–24.

Carbone, G. and Memoli, V. (2015) 'Does democratization foster state consolidation? Democratic rule, political order, and administrative capacity', *Governance*. Wiley Online Library, 28(1), pp. 5–24.

Delavallade, C. (2006) 'Corruption and distribution of public spending in developing countries', *Journal of economics and finance*. Springer, 30(2), pp. 222–239.

Dhingra, S. *et al.* (2016) *The consequences of Brexit for UK trade and living standards*. LSE.

Fong, C., Hazlett, C. and Imai, K. (2018) 'Covariate balancing propensity score for a continuous treatment: application to the efficacy of political advertisements', *The Annals of Applied Statistics*. Institute of Mathematical Statistics, 12(1), pp. 156–177.

Franco, Á., Álvarez-Dardet, C. and Ruiz, M. T. (2004) 'Effect of democracy on health: ecological study', *BMJ*, 329(7480), pp. 1421–1423.

Grindle, M. S. (2012) *Jobs for the Boys*. Boston, MA: Harvard University Press.

Hanson, J. K. and Sigman, R. (2013) 'Leviathan's latent dimensions: measuring state capacity for comparative political research', in *APSA 2011 Annual meeting paper*.

Helble, M., Mann, C. and Wilson, J. S. (2009) *Aid for trade facilitation*. The World Bank.

Hendrix, C. S. (2010) 'Measuring state capacity: Theoretical and empirical implications for the study of civil conflict', *Journal of peace research*. Sage Publications Sage UK: London, England, 47(3), pp. 273–285.

ILOSTAT (2018) *Employment by economic activity*. Geneva, Switzerland: International Labour Organization. Retrieved May 22, 2018 (https://www.ilo.org/global/statistics-and-databases/lang--en/index.htm): International Labour Organization.

Kaufmann, D., Kraay, A. and Mastruzzi, M. (2011) 'The worldwide governance indicators: methodology and analytical issues', *Hague Journal on the Rule of Law*. Cambridge University Press, 3(2), pp. 220–246.

Lagomarsino, G. *et al.* (2012) 'Moving towards universal health coverage: health insurance reforms in nine developing countries in Africa and Asia', *The Lancet*. Elsevier, 380(9845), pp. 933–943.

Leeper, T. J. (2017) 'Interpreting regression results using average marginal effects with R's margins', *Available at the comprehensive R Archive Network (CRAN)*.

Lindberg, S. I. *et al.* (2014) 'V-Dem: A new way to measure democracy', *Journal of Democracy*. Johns Hopkins University Press, 25(3), pp. 159–169.

Marshall, M. G., Jaggers, K. and Gurr, T. R. (2002) 'Polity IV project: Dataset users' manual', *College Park: University of Maryland*.

Van der Meer, T., Te Grotenhuis, M. and Pelzer, B. (2010) 'Influential cases in multilevel modeling: a methodological comment', *American Sociological Review*. Sage Publications Sage CA: Los Angeles, CA, 75(1), pp. 173–178.

Meyer, O. W. *et al.* (1997) 'World Society and the Nation-State', *American Journal of Sociology*, 103(1), pp. 144–181. doi: 10.1086/231174.

Morgan, S. L. and Winship, C. C. N.-H. . M. 2007 300. 7. 22 (2007) *Counterfactuals and causal inference: methods and principles for social research*. Cambridge: Cambridge University Press.

Mustillo, S. A., Lizardo, O. A. and McVeigh, R. M. (2018) 'Editors' comment: A few guidelines for quantitative submissions'. SAGE Publications Sage CA: Los Angeles, CA.

Nooruddin, I. and Simmons, J. W. (2009) 'Openness, uncertainty, and social spending: Implications for the globalization—welfare state debate', *International Studies Quarterly*. Blackwell Publishing Ltd Oxford, UK, 53(3), pp. 841–866.

Pomeranz, D. and Vila-Belda, J. (2019) 'Taking State-Capacity Research to the Field: Insights from Collaborations with Tax Authorities', *Annual Review of Economics*. Annual Reviews.

Rauch, J. E. and Evans, P. B. (2000) 'Bureaucratic structure and bureaucratic performance in less developed countries', *Journal of public economics*. Elsevier, 75(1), pp. 49–71.

Richiardi, L., Bellocco, R. and Zugna, D. (2013) 'Mediation analysis in epidemiology: methods, interpretation and bias', *International Journal of Epidemiology*, 42(5), pp. 1511–1519. doi: 10.1093/ije/dyt127.

Rothstein, B. O. and Teorell, J. A. N. (2008) 'What is quality of government? A theory of impartial government institutions', *Governance*. Wiley Online Library, 21(2), pp. 165–190.

Snijders, T. A. B. and Berkhof, J. (2008) 'Diagnostic checks for multilevel models', in *Handbook of multilevel analysis*. Springer, pp. 141–175.

Stubbs, T., King, L. and Stuckler, D. (2014) 'Economic growth, financial crisis, and property rights: observer bias in perception-based measures', *International Review of Applied Economics*. Taylor & Francis, 28(3), pp. 401–418.

Teorell, J. *et al.* (2016) 'The quality of government standard dataset, version Jan16', *University of Gothenburg: The Quality of Government Institute*.

Xu, G. (2018) 'The costs of patronage: Evidence from the british empire', *American Economic Review*, 108(11), pp. 3170–3198.

1. Appendix Tables

Table S1. Data sources and measurement of variables in main analysis

Variable	Description	Source
Domestic general government ('public') health expenditure	Public spending on health-care and services from domestic sources, measured in i) US dollars per capita, adjusted for inflation and differences in purchasing power, and, ii) as a share of GDP.	World Development Indicators
Private health expenditure	Spending on health-care and services from private sources, measured in i) US dollars per capita, adjusted for inflation and differences in purchasing power, and ii) as a share of GDP. Private sources include households, corporations and non-profit organizations.	World Development Indicators
Weighted mean tariff rate	Weighted mean applied tariff as a percentage of the import value. This captures the average of effectively applied rates weighted by the product import shares corresponding to each partner country.	World Development Indicators
Gross Domestic Product (GDP) per capita	Gross Domestic Product, measured in US dollars per capita, adjusted for inflation and differences in purchasing power.	World Development Indicators
Government effectiveness	Composite indicator of the perceptions of the quality of public services, civil service, policy formulation, implementation, and the credibility of the government's commitment to such policies. Measured as a percentile rank indicating the country's rank among all countries covered by the aggregate indicator, with 0 corresponding to lowest rank, and 100 to highest rank.	Worldwide Governance Indicators
ODA per capita	Net overseas development assistance per capita in US dollars	World Development Indicators
War dummy	Dichotomous indicator coded as 1 in a country if there was an armed conflict resulting in 1000 or more deaths in that year, 0 otherwise	UCDP/PRIO Armed Conflict Dataset

Table S2. Descriptive statistics

Variable	**Mean or number (Standard deviation or %)**
Tariff rate, %	8·59 (4·78)
Government health expenditure per capita, US dollars	227·53 (213·62)
Private health expenditure per capita, US dollars	162.71 (170.52)
WGI government effectiveness score, percentile rank	38·84 (20·49)
GDP per capita, US dollars	6,188.40 (4,538.57)
Official development assistance per capita, US dollars	69.65 (100.74)
Country at war (1=at war, 0 otherwise)[a]	20 (3·2)
Countries	65
Years	17
Country-years[b]	632

Notes: a: figure shows number of cases at war, all other figures are means. See Table S1 for list of data sources and variable measurement. See Table S4 for list of countries included in the analysis and number of years of data for each country. All US dollar figures adjust for inflation and differences in purchasing power.

Table S3. Data source and measurement for variables included in robustness checks

Variable	Description	Source
Voice and accountability score percentile	Composite indicator of the perceptions of the extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association, and a free media. Percentile rank indicates the country's rank among all countries covered by the aggregate indicator, with 0 corresponding to lowest rank, and 100 to highest rank.	World Governance Indicators
Control of corruption score	Composite indicator of the perceptions of the extent to which public power is exercised for private gain, including both petty and grand forms of corruption, as well as "capture" of the state by elites and private interests. Percentile rank indicates the country's rank among all countries covered by the aggregate indicator, with 0 corresponding to lowest rank, and 100 to highest rank.	World Governance Indicators
Polity score	Index of regime authority measured on a 21-pont scale ranging from -10 (hereditary monarchy) to +10 (consolidated democracy). The Polity score consists of six component measures that record key qualities of executive recruitment, constraints on executive authority and political competition.	Center for Systemic Peace
Population age structure	Population between the ages 0 to 14 or over 65 as a percentage of the total population	World Bank World Development Indicators
Fertility rate	The number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year.	United Nations Population Division
Urbanisation	Urban population as a share of the total population	World Bank World Development Indicators
Political globalization	Index capturing engagement in international political integration. It is measured using the number of multilateral treaties signed since 1945, the number of memberships in international organizations and a measure for the treaty partner diversity. The raw data are aggregated and countries are assigned a score of 1-100 based on the percentiles of the distribution in each year.	KOF Swiss Economic Institute

Unweighted mean tariff rate	Simple mean applied tariff is the unweighted average of effectively applied rates for all products subject to tariffs calculated for all traded goods, measured as a percentage of the import value	World Bank World Development Indicators
Weighted Most Favoured Nation (MFN) tariff rate	Average of Most Favoured Nation rates (applied to WTO members), weighted by the product import shares corresponding to each partner country	World Bank World Development Indicators

Table S4. Countries included in analytical sample

Country	No. years of data (% of all country-years)
Albania	14 (2.2)
Algeria	10 (1.6)
Angola	12 (1.9)
Bangladesh	13 (2.1)
Belize	13 (2.1)
Benin	10 (1.6)
Bhutan	5 (0.8)
Botswana	13 (2.1)
Brazil	14 (2.2)
Bulgaria	3 (0.5)
Burkina Faso	8 (1.3)
Burundi	13 (2.1)
Cabo Verde	8 (1.3)
Cameroon	12 (1.9)
Central African Republic	12 (1.9)
Chad	11 (1.7)
China	9 (1.4)
Colombia	11 (1.7)
Comoros	7 (1.1)
Costa Rica	16 (2.5)
Cote d'Ivoire	14 (2.2)
Ecuador	16 (2.5)
Egypt, Arab Rep.	11 (1.7)
Ethiopia	7 (1.1)
Gambia, The	8 (1.3)
Ghana	7 (1.1)
Grenada	10 (1.6)
Guatemala	13 (2.1)
Guinea-Bissau	8 (1.3)
Guyana	13 (2.1)
Indonesia	8 (1.3)
Jordan	13 (2.1)
Kenya	12 (1.9)
Kiribati	0 (0.0)
Lao PDR	8 (1.3)
Lesotho	2 (0.3)
Madagascar	16 (2.5)
Malawi	10 (1.6)
Malaysia	8 (1.3)
Mauritania	6 (0.9)

Mauritius	11 (1.7)
Mexico	3 (0.5)
Mongolia	13 (2.1)
Morocco	7 (1.1)
Namibia	8 (1.3)
Niger	8 (1.3)
Nigeria	14 (2.2)
Pakistan	11 (1.7)
Papua New Guinea	10 (1.6)
Peru	13 (2.1)
Philippines	15 (2.4)
Rwanda	13 (2.1)
Senegal	14 (2.2)
Sierra Leone	5 (0.8)
Sri Lanka	4 (0.6)
St. Lucia	12 (1.9)
St. Vincent and the Grenadines	10 (1.6)
Sudan	8 (1.3)
Suriname	7 (1.1)
Thailand	6 (0.9)
Togo	1 (0.2)
Tonga	6 (0.9)
Tunisia	10 (1.6)
Vanuatu	3 (0.5)
Zambia	13 (2.1)
Zimbabwe	3 (0.5)

Notes: Total number of countries: 65; total number of: 17.

2. Appendix Figures

Figure S1. Association between country mean tariff rate and public health expenditure per capita in each year, 1996-2015

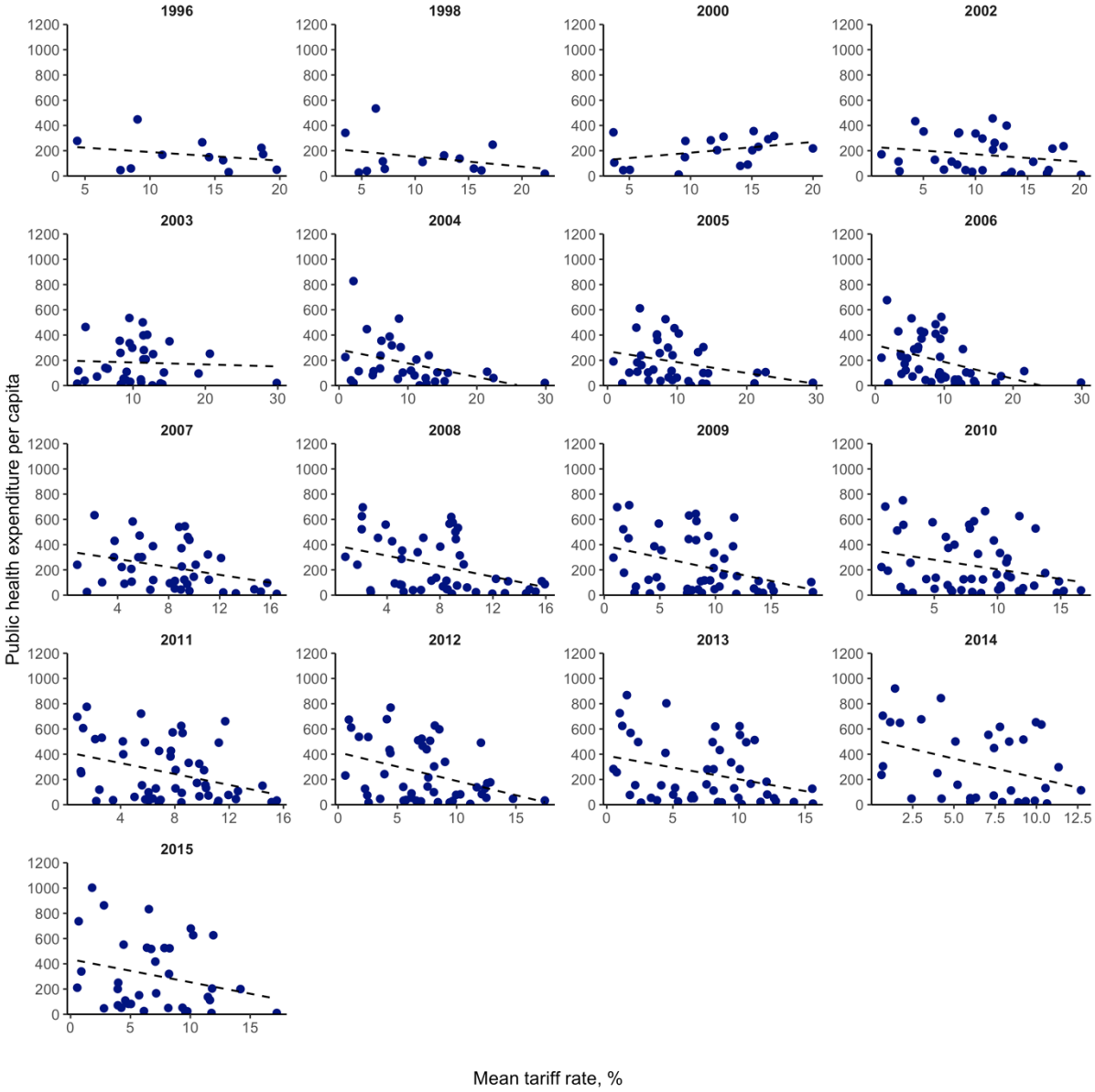


Figure S2. Association between country mean tariff rate and public health expenditure as a percentage of GDP, 1996-2015

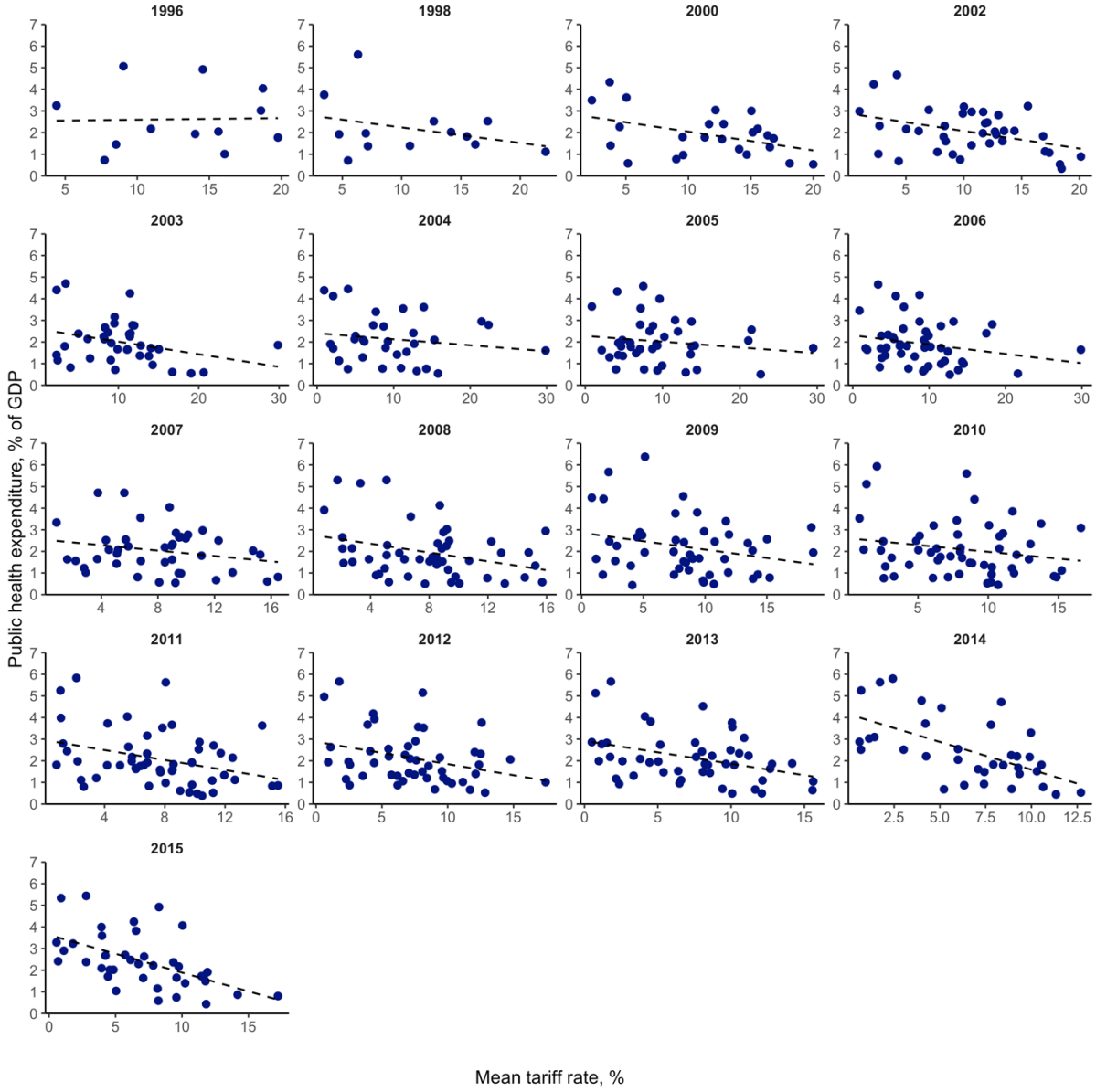


Figure S3. Association between country mean tariff rate and private health expenditure per capita and as percentage of GDP, 1996-2015

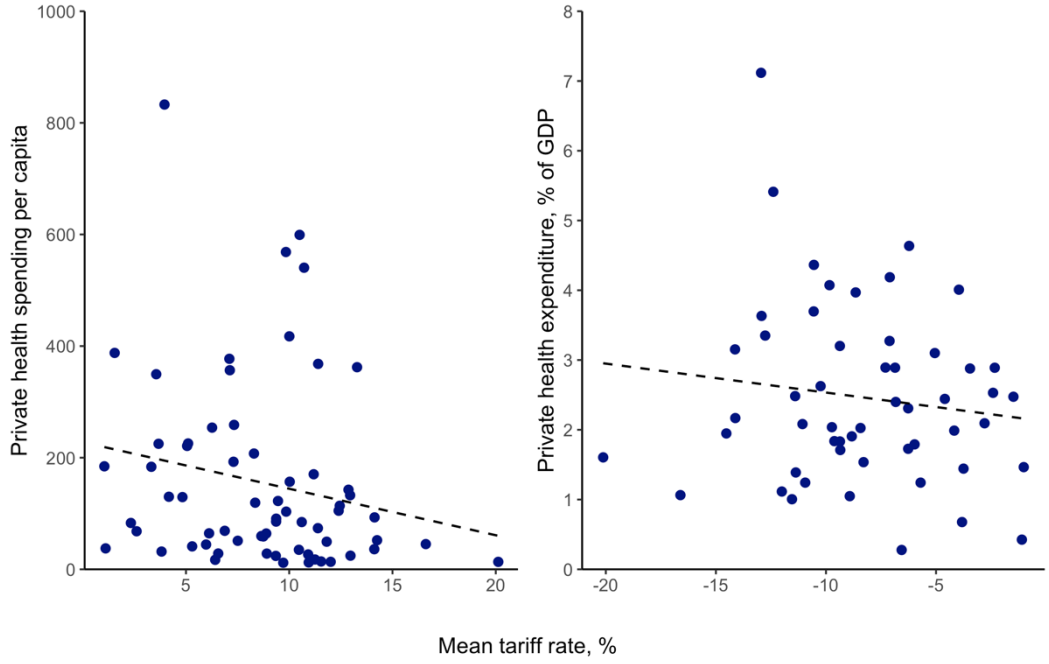


Figure S4. Average marginal effect of 1% tariff reduction on government and private health expenditure as a percentage of GDP

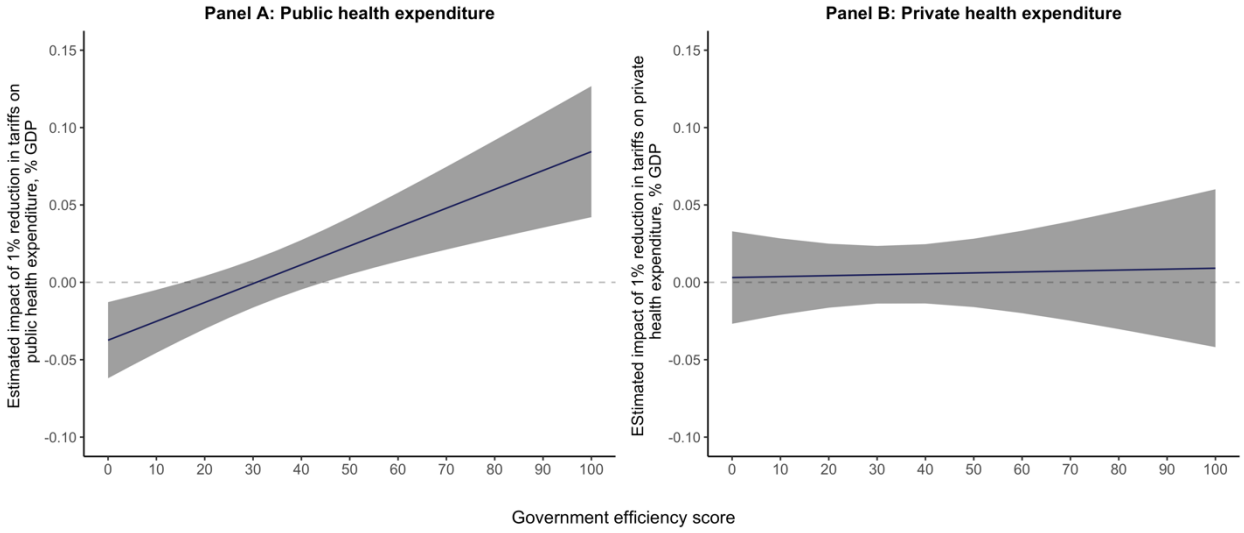


Figure S5. Mediation analyses Step 3 and Step 4

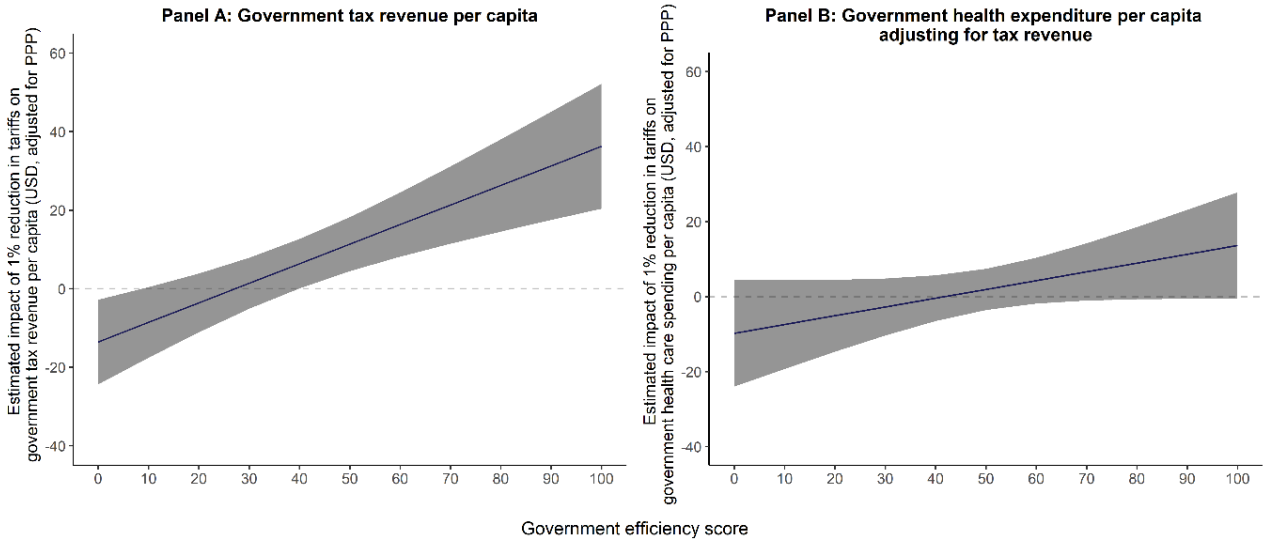
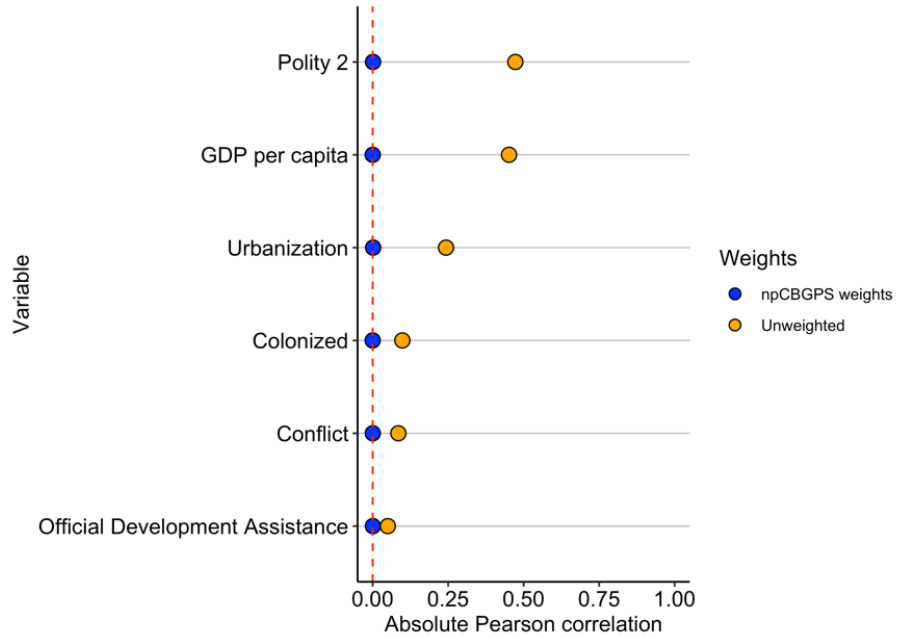
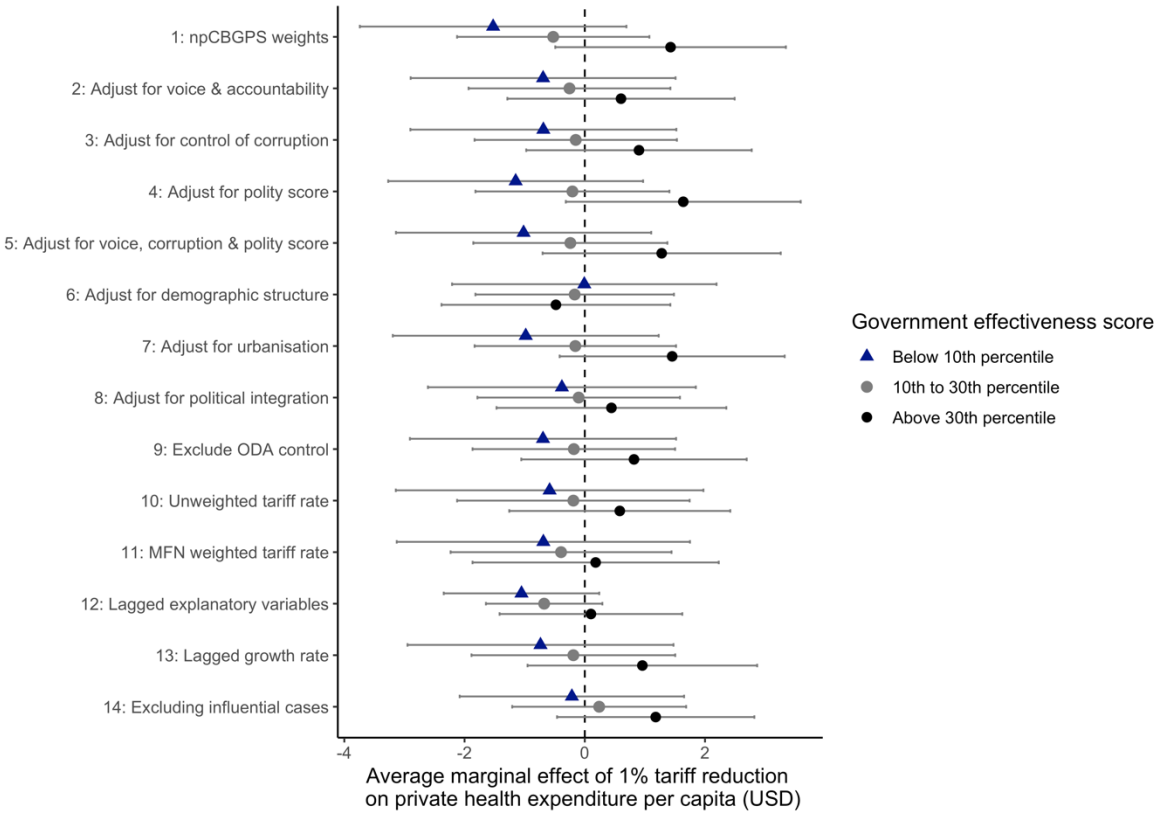


Figure S6. Covariate balance in npCBGPS specifications



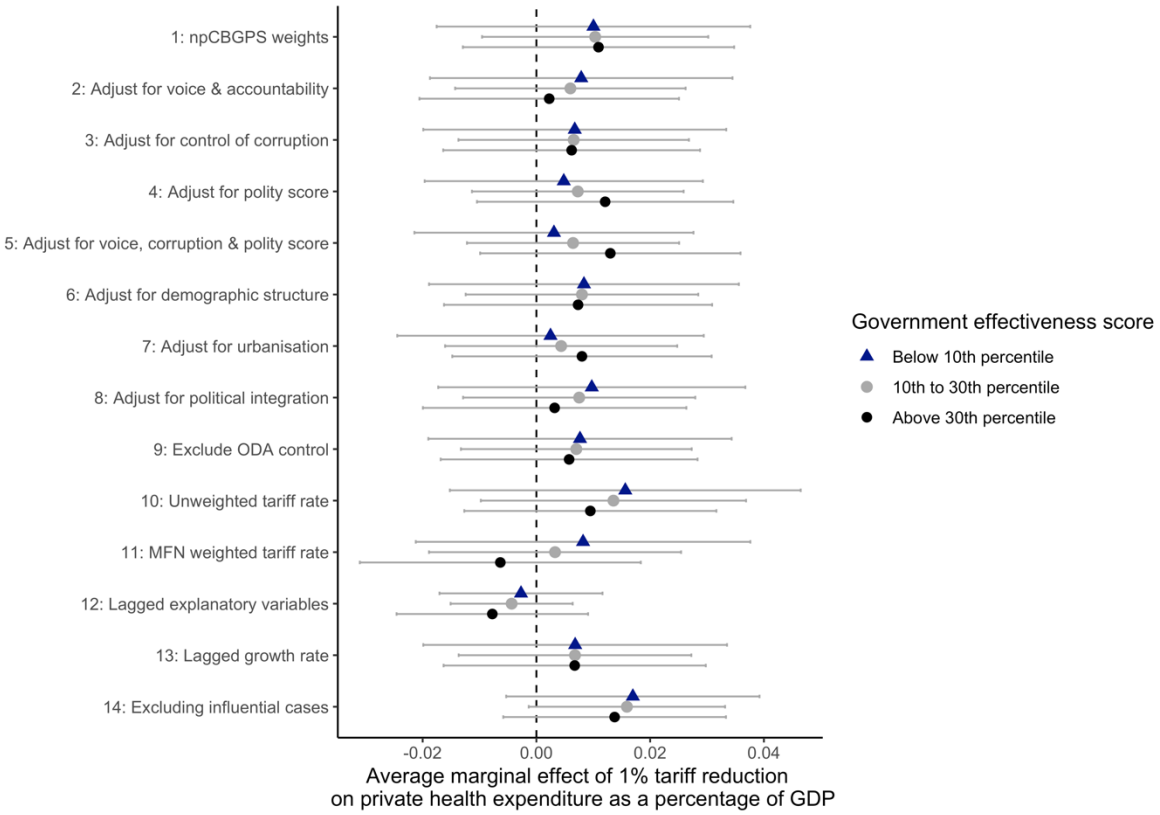
Notes: npCBGPS weights are 'non-parametric Covariate Balancing Propensity Score' weights.(Fong, Hazlett and Imai, 2018) Mean absolute Pearson correlation of tariff covariates and predictions reduces from 0·20 to 0·0002 when applying npCBGPS weights. These are estimated using an algorithm that minimises the Pearson correlation between covariates and treatment assignment in the sample whilst simultaneously maximising the prediction of treatment assignment.

Figure S7. Alternative specifications: Average Marginal Effect (AME) of 1% tariff reduction on private health expenditure per capita, by government effectiveness score percentile



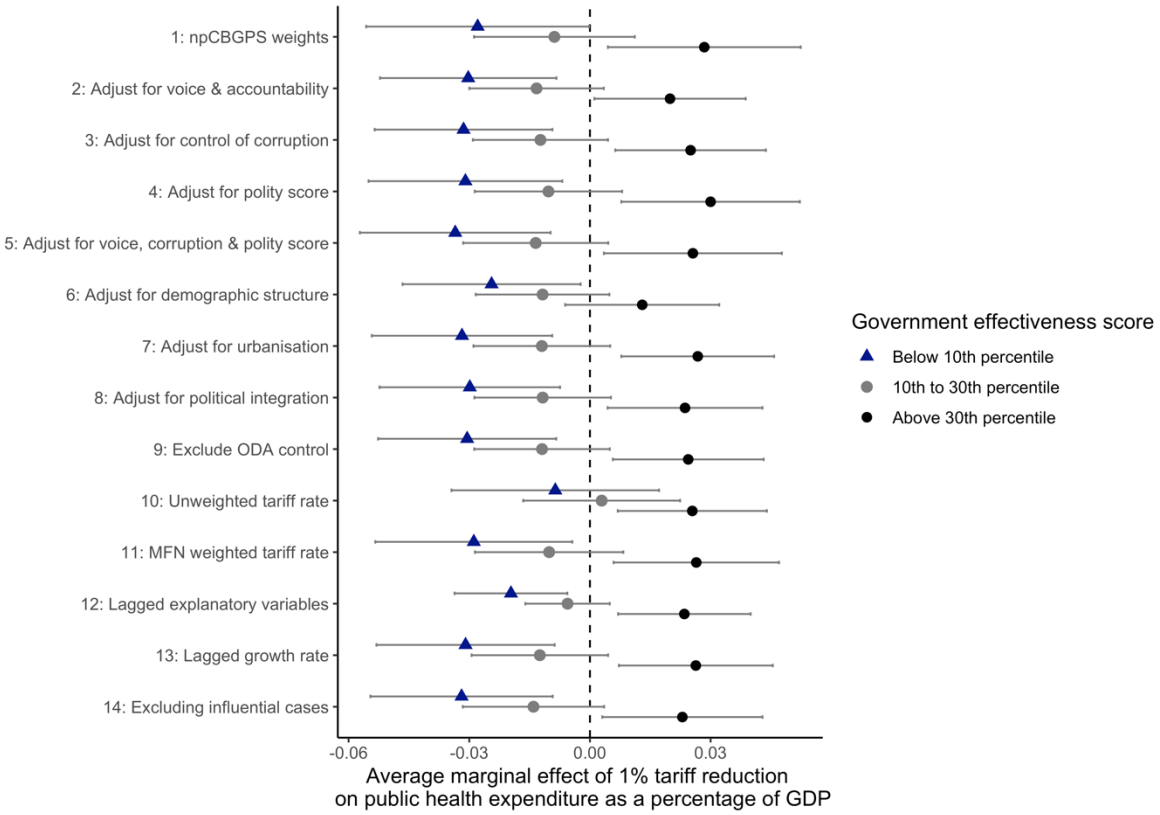
Notes: Table S3 lists the data sources and measurement of variables included in robustness checks. Model 1 is a weighted fixed-effects regression model which includes the same controls as the main model and also re-weights country observations using non-parametric Covariate Balancing Generalised Propensity Score weight. Mean absolute Pearson correlation of tariff covariates and predictions reduces from 0·20 to 0·0002 when using npCBGPS weights; see Figure S6 for visualisation of covariate balance in npCBGPS specifications. Models 2-9 include/exclude the listed variables as controls in the model. Model 10 uses the unweighted mean tariff as predictor rate rather than the import-weighted average tariff rate and Model 11 uses the MFN, trade-weighted tariff rate. Model 12 lags all explanatory variables by one year to account for budget cycles. Model 13 adjusts for the growth rate in the previous year to reduce 'observer bias'. Model 14 excludes influential cases with Cook's D larger than 4/n. Appendix S3 describes each test in detail.

Figure S8. Alternative specifications: Average Marginal Effect (AME) of 1% tariff reduction on private health expenditure as a percentage of GDP, by government effectiveness score percentile



Notes: Model 1 is a weighted fixed-effects regression model which includes the same controls as the main model and also re-weights country observations using non-parametric Covariate Balancing Generalised Propensity Score weight. Mean absolute Pearson correlation of tariff covariates and predictions reduces from 0.09 to 0.0008 when using npCBGPS weights; see Figure S6 for visualisation of covariate balance in npCBGPS specifications. Models 2-9 include/exclude the listed variables as controls in the model. Model 10 uses the unweighted mean tariff as predictor rate rather than the import-weighted average tariff rate and Model 11 uses the MFN, trade-weighted tariff rate. Model 12 lags all explanatory variables by one year to account for budget cycles. Model 13 adjusts for the growth rate in the previous year to reduce 'observer bias'. Model 14 excludes influential cases with Cook's D larger than 4/n. Appendix S3 describes each test in detail.

Figure S9. Alternative specifications: average marginal effect of 1% tariff reduction on public health expenditure as a percentage of GDP by government effectiveness score percentile



Notes: Model 1 is a weighted fixed-effects regression model which includes the same controls as the main model and also re-weights country observations using non-parametric Covariate Balancing Generalised Propensity Score weight. Mean absolute Pearson correlation of tariff covariates and predictions reduces from 0.09 to 0.0008 when using npCBGPS weights; see Figure S6 for visualisation of covariate balance in npCBGPS specifications. Models 2-9 include/exclude the listed variables as controls in the model. Model 10 uses the unweighted mean tariff as predictor rate rather than the import-weighted average tariff rate and Model 11 uses the MFN, trade-weighted tariff rate. Model 12 lags all explanatory variables by one year to account for budget cycles. Model 13 adjusts for the growth rate in the previous year to reduce 'observer bias'. Model 14 excludes influential cases with Cook's D larger than 4/n. Appendix S3 describes each test in detail.