Bensch, Gunther; Ankel-Peters, Jörg; Vance, Colin

Conference Paper

Spotlight on Researcher Decisions – Infrastructure Evaluation, Instrumental Variables, and Specification Screening

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2023: Growth and the "sociale Frage"

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Bensch, Gunther; Ankel-Peters, Jörg; Vance, Colin (2023) : Spotlight on Researcher Decisions – Infrastructure Evaluation, Instrumental Variables, and Specification Screening, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2023: Growth and the "sociale Frage", ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg

This Version is available at:
https://hdl.handle.net/10419/277703

Spotlight on researcher decisions – Infrastructure evaluation, instrumental variables, and specification screening

Jörg Ankel-Peters[1,2*], Gunther Bensch[1], and Colin Vance[1,3]

[1] RWI – Leibniz Institute for Economic Research, Germany
[2] University of Passau, Germany
[3] Jacobs University, Germany

January 2023

Abstract

This paper revisits the instrumental variable (IV) approach in Lipscomb et al. (2013, 2021, LMB) to study the impacts of electrification. We first make corrections to the construction of the dataset, including the modelled IV. Revised estimates on main outcomes and mechanisms are statistically insignificant, with substantially lower effect sizes. We second develop a framework that accounts for weak IVs and discourages specification screening. Applying it to LMB, we find that most theoretically justified specifications yield insignificant results. The proposed framework is transferable to other IV applications to reduce potential bias stemming from researcher's or replicator's discretion.

JEL: O13, C52, O18

Keywords: replication, instrumental variables, electrification, infrastructure, specification curve analysis, robust inference

Acknowledgements: Funding: This work was supported by the German Research Foundation (DFG) [Grant No. 3473/1-1 within the DFG Priority Program META-REP (SPP 2317)]. We are grateful to Molly Lipscomb, Mushfiq Mobarak, and Dimitri Szerman for fruitful exchange on earlier manuscripts from this replication project and for their support in making all replication data available. We are furthermore grateful to Gunnar Gotz for excellent research assistance to a previous version of this replication exercise. We also thank Boris Blagov, Abel Brodeur, Sylvain Chabé-Ferret, Kenneth Gillingham, Michael Grimm, Magnus Johannesson, Michal Kolesár, Daniel Millimet, Florian Neubauer, Julian Rose, and Jevgenijs Steinbuks for helpful comments and suggestions.
*Corresponding author: joerg.peters@rwi-essen.de

1. Introduction

Causally identifying the effects of infrastructure access on economic development remains challenging, impeding evidence-based policy guidance in this investment-intensive field. In the electrification literature, several papers use geographic variation as instrumental variables (IVs), since randomization of power lines and other networks, as in Lee et al. (2020*a*), is practically impossible in most cases (see Lee et al. 2020*b*).

This paper pursues two aims. First, we provide revised impact estimates from a computational replication of Lipscomb, Mobarak, and Szerman (2021, henceforth LMS 2021), a corrigendum of one of the most influential articles in this literature, Lipscomb, Mobarak, and Barham (2013, henceforth LMB 2013). The LMS (2021) corrigendum responds to a coding error in LMB (2013), diagnosed in our previous computational replication (see next section and Bensch et al. 2021, henceforth BPV 2021). We now show that errors related to the construction of outcome variables and the IV in LMB (2013) remain in LMS (2021); removing these errors leads to statistically insignificant effects in LMS' (2021) main specification that are much smaller in size. We thereby augment the infrastructure and electrification literature by revisiting the findings from this seminal paper.

Second, we argue that theoretically justified alternatives to LMS' main specification exist and hence propose an IV sensitivity testing framework that reduces the room for selective researcher decisions in a structured, simple, and transparent manner. We apply this framework to the LMB/LMS data. The framework integrates recent contributions from the causal inference literature: We draw on specification curve analysis (Simonsohn et al. 2020), inference robust to weak IVs and screening on the first stage (Andrews et al. 2019; Lee et al. 2022; Angrist and Kolesár 2022), reporting and publication bias in the IV-based literature (Andrews et al. 2019; Brodeur et al. 2020; Kranz and Pütz 2022), and more general concerns about sensitivity in IV settings (Young 2022).

The framework comprises two steps, starting with a structured specification inventory that delineates the key researcher-decision domains, including the selection of control variables and the weighting of the regression. We find several choice options for six researcher-decision domains that have a theoretically similar justification as the choices made in the main specification in LMS (2021). The second step involves transparently reporting the range of results that emerges from combining the various choice options in different IV specifications.

For this transparent reporting, we adopt specification curve analyses following Simonsohn et al. (2020), including a simplified one-factor-at-a-time (OFAT) graph. This reveals that our revised main-specification estimates respond strongly to adjustments in the specification. LMS' main specification stands out in terms of statistical and economic significance among a myriad of theoretically justified specifications. We conclude that this framework is transferable to other IV settings to prevent selective reporting and specification screening.

In the case of LMS (2021), selective reporting originates from screening on first-stage strength, whereby researchers pretest theoretically justified specifications and keep specifications with stronger first stages (Andrews et al. 2019). The screening practice typically remains opaque in published articles. LMS (2021), though, make transparent that they pick a different specification than LMB to "retain first-stage power" (LMS 2021:1) in response to our previous replication in BPV (2021). Andrews et al. (2019) argue that this elimination of theoretically justified specifications not only opens scope for reporting and publication bias, but that it is also unnecessary, since inference robust to weak IV is possible.[1] In our analysis, we therefore combine two recently developed approaches that are robust to weak IV and avoid this reporting bias.

The first approach, from Angrist and Kolesár (2022), pretests on the first-stage *sign* and discards specifications with empirically wrong signs (a negative coefficient for the IV in the LMB/LMS case). This, according to Angrist and Kolesár (2022), narrows down bias considerably while not sharing the distorting effect on inference of screening on first-stage *strength*. The second approach, from Lee et al. (2022), implements what they call the *tF*-adjustment as a "standard inference benchmark" that is robust to weak IVs and "not overly cautious" (Lee et al. 2022, p. 3279). This adjustment smoothly increases second-stage standard errors and hence widens confidence intervals as the first-stage *F*-statistic decreases. It thereby makes the common *t*-ratio inference approach robust to weak IVs without imposing further assumptions, in a similar way as Anderson-Rubin confidence intervals. To use the two approaches complementarily, we only flag but do not drop specifications with wrong signs, given that the *tF*-adjustment requires abstaining from screening out specifications based on

[1] Young (2022) furthermore cautions against screening on first-stage *F*-Statistics. As he argues, *F*-Statistics tend to be uninformative guides to the performance of conventional inference procedures in typical IV study setups.

pre-testing.[2] As an additional robustness check we integrate the delete-one sensitivity test, which tests for the maximum change in estimates when deleting one cluster from the sample (see Young 2022).

Our framework therefore provides structure for a comprehensive specification inventory and presentation. While our focus is on the role of first-stage strength and potential screening for strong IVs, the framework likewise helps to reduce room for *p*-hacking, which, according to Brodeur et al. (2020), is particularly pervasive in IV applications. In practice, first-stage screening is likely intertwined with *p*-hacking in that researchers have incentives to jointly check first-stage *F*-statistics and second-stage *p*-values to ultimately pick specifications that deliver pleasant *p*-values at reasonable first-stage strength. In a similar vein, the framework also restricts selective behaviour from replicators (cf. Bryan et al. 2019). More generally, our framework may be useful when pre-registration and pre-analysis plans are difficult, as is the case in many setups where secondary data is used (Christensen and Miguel 2018; Burlig 2018; Ofosu and Posner 2021). In that context, the framework can be used in pre-analysis plans by pre-committing to a transparent post-hoc identification of researcher decision-domains.

2. LMB (2013)'s instrumental variable, BPV (2021)'s replication and the LMS (2021) corrigendum

LMB (2013) study the impact of electrification on two main development indicators, housing values and the Human Development Index (HDI), as well as on 17 additional mechanism indicators.[3] Their identification strategy is to instrument actual electricity grid roll-out in Brazil from 1960 to 2000 by the hypothetical electricity grid coverage that would have emerged if no socio-economic but only geographic features had determined electricity infrastructure investments. Thereby, LMB (2013) aim to account for the likely endogeneity of infrastructure placement.

[2] The two approaches differ in that Angrist and Kolesár (2022) assume that IV endogeneity does not exceed a certain threshold, which Lee et al. (2022) consider as too restrictive. Our approach to combining the two approaches avoids having to decide on such an assumption, which "ultimately does not follow from any econometric result; instead, it rests entirely on how comfortable one is with those additional a priori assumptions." (Lee et al. 2022, p. 3279).

[3] LMB (2013) also show results for another indicator, *life expectancy*. We consider this indicator as redundant, since the set of indicators already includes the HDI sub-component *longevity*, which is a transformation of life expectancy (correlation coefficient of 0.97).

The authors use a sophisticated engineering-based predictive model to construct an IV that is exogenous to economic development and that determines a considerable part of electricity infrastructure investment costs. The model focuses on hydropower, the predominant electricity source in Brazil. It uses a probit regression with geographic covariates to predict a "suitability index"; the index determines for each of around 32,500 evenly spaced grid points the hypothetical sequence of hydropower placement – the more suitable a location technically is, the earlier the hydropower plant is built there. A separate algorithm uses area slope information to determine the cost-minimizing transmission line network linked to these hydropower plants, assuming two substations per plant. It is further assumed that all grid points within 50 kilometres of a predicted plant or substation are electrified. The resulting hypothetical electricity grid coverage rates at the county level are lagged by one decade to form the IV used in the two-stage least squares (2SLS) application, with four data points covering the 30 years between 1970 and 2000. Here, LMB (2013) additionally include an Amazon-decade interaction term to control for time-varying idiosyncrasies in this sizable region with prohibitively high costs of infrastructure development.

In BPV (2021), we detect inconsistencies between the Amazon definitions used in the construction of the IV and the 2SLS estimation in LMB (2013). BPV (2021) furthermore document preliminary analyses using the data originally published on the website of the *American Economic Journal: Applied Economics* (AEJ:AE), suggesting that consistent Amazon definitions lead to statistically insignificant estimates and smaller effect sizes. The analyses were preliminary to the extent that the full data set was not provided on the AEJ:AE website. LMS (2021) provide an online corrigendum that removes the inconsistency, published after extensive exchanges between BPV and LMS and attempts to jointly publish a corrigendum. BPV (2021) summarize this process in a preamble. Lastly, after posting LMS (2021) online, the original authors shared with us, as an addition to the AEJ:AE dataset, the MATLAB® code to run the engineering model described above (Szerman et al. 2022*a*).

In line with BPV (2021), LMS (2021) use a refined vegetation-based definition of the Amazon, for which they find a very low first-stage *F*-statistic of 2.1 in the main specification defined in LMB (2013). In response to this low *F*-statistic, LMS (2021) define a new main specification with a revised set of interaction terms included in the 2SLS estimation, maintaining the same IV. Instead of Amazon-decade interaction terms, decade dummies are interacted with the

quartic of the "suitability index". This IV approach is also included as a robustness check in LMB (2013) and used by the authors as the main identification strategy in Szerman et al. (2022*b*), a follow-on paper examining the deforestation effect of agricultural productivity induced by electrification.

3. LMB (2013) and LMS (2021) revisited

In this section, we replicate the results for all outcomes and mechanisms in LMB (2013) and revised in LMS (2021). We reconstruct the outcome and control variables based on data we retrieved from the *Instituto de Pesquisa Econômica Aplicada* (IPEA).[4] With the complete dataset at hand, including the code to run the engineering model (Szerman et al. 2022*a*), we detect two seemingly minor technical issues that remained in LMS (2021). Removing these issues and otherwise using the same main specification as LMS (2021), however, leads to statistically insignificant estimates and much smaller effect sizes, as we will show in Sections 3.2 to 3.4.

3.1 Corrections to the data

First, we remedy a few inconsistent variable adjustments and aggregations. For example, LMB (2013) define an adjustment for variables that changed definition in 1990. LMB (2013) and LMS (2021) apply this adjustment to affected variables – except for the outcome variable HDI, even though it changed definition in 1990, too. Furthermore, data was not always correctly aggregated from municipality to the county level (see Appendix A for a comprehensive discussion).

Second, LMS (2021) state that a 'seed' – a starting point of a random number generator – is required in the grid simulation algorithm as part of the IV construction to make exact replication possible. Yet, the results presented in LMS (2021) are not reproducible, since they do not correspond to the output of the replication code in Szerman et al. (2022*a*). Moreover, multiple runs of the grid simulation without setting a seed yield multiple IV outputs and, in turn, qualitatively and quantitatively different outcome estimates. This is due to the algorithm's incapacity to identify the one cost-minimizing electricity grid required to yield

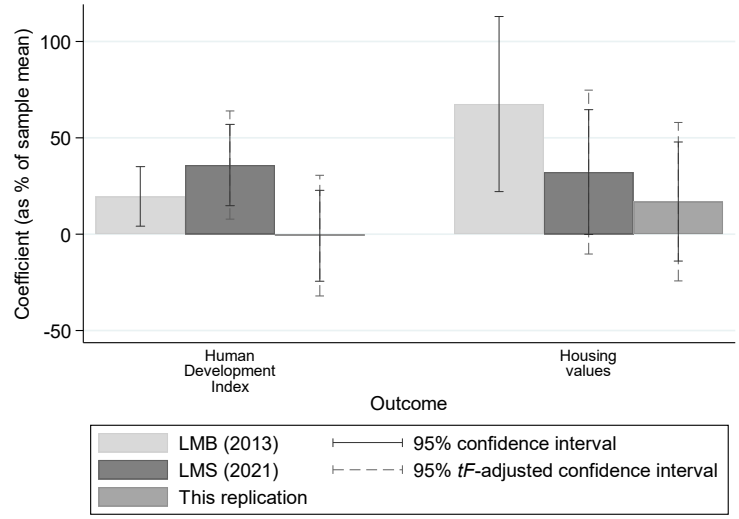
[4] Except for *in-migration*, all outcome variables could be retrieved from IPEA. While LMS (2021) disregard this indicator, we rely on the data originally published by LMB (2013) on the website of AEJ:AE, recognizing that it is unclear to us where this data came from. While the available data on the IV construction would have allowed us to amend the dataset by one decade, IPEA does not provide the necessary data for 2010 for any of the outcome variables.

stable results. The algorithm adopted by LMS (2021) compares 80,000 electricity grid outlines, which is insufficient in the given context where the number of potential electricity networks in the first of the four studied decades alone exceeds 10^{1000}, as we show in Appendix A. This also precludes us from solving the problem by increasing the number of iterations to a degree that is still computationally tractable. Thus, LMS (2021) also involves a convergence problem related to the IV construction.[5] In Appendix A we also show that irrespective of whether a seed is set has no bearing upon our overall findings of vanishing statistical and economic significance for the LMS results. In the present section, we therefore decide to apply the seed included in Szerman et al. (2022*a*) as an imperfect, but for our purposes sufficient remedy that at least heals the reporting error and makes our results reproducible. This is the second reason why our results differ from those printed in LMS (2021).

3.2 Main outcomes

Figure 1 contrasts effect sizes for the two main outcomes of LMB (2013) and LMS (2021) with those from our replication, which corrects the two issues outlined above. The figure expresses effect sizes as percentage of sample means and shows 95% confidence intervals both based on conventional *t*-ratio critical values and based on the *tF*-adjustment procedure proposed by Lee et al. (2022). The latter were also used by LMS (2021) to check robustness.[6]

Figure 1: Comparison of effect sizes for main electrification outcomes



[5] New simulation attempts may overcome the convergence problem, but stricter modelling assumptions seem to be required, which would probably make the IV more dependent on non-geographic factors and thereby counteract the motivation of the IV.

[6] Anderson-Rubin confidence intervals as the widely used weak-IV-robust alternative can be expected to be larger (Lee et al. 2022), which holds true for most of the estimations performed as part of this paper as well.

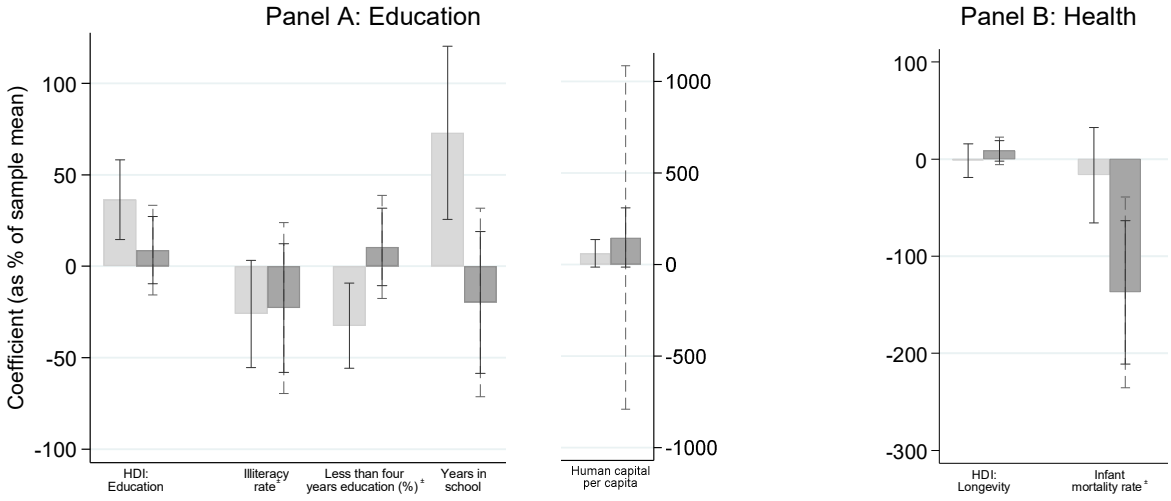
LMS (2021) find the effect size for HDI to almost double relative to LMB (2013)'s results, while the effect size for housing values decreases by about half. The revised HDI effect from the present replication, however, is zero, while the housing values effect size decreases by half compared to LMS (2021) to an estimate equivalent to 16.9% of the sample mean. While the latter would still be economically significant, the related *p*-value is 0.28 and both confidence intervals overlap with zero.

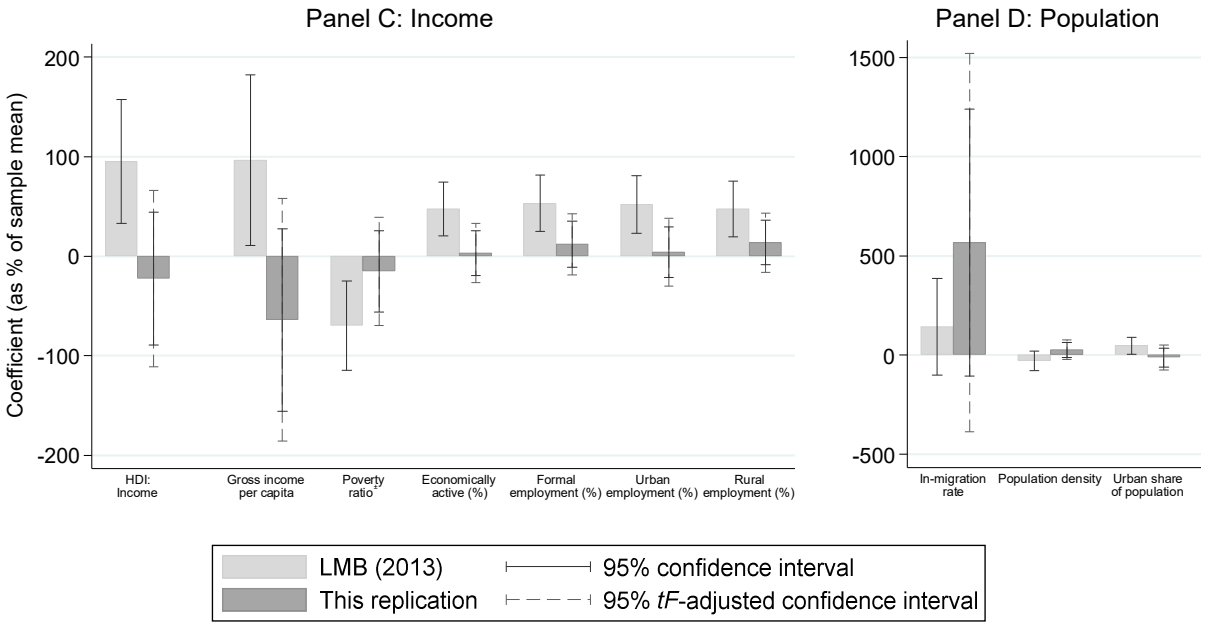
3.3 Mechanisms

LMB (2013) corroborate their findings on main outcomes by additionally demonstrating effects on mechanisms, which underpin the plausibility of their findings. Robust effects on those mechanisms would suggest that parts of the LMB (2013) causal chain hold, even if we do not detect an effect on the final outcome measure. In their effort to unpack underlying mechanisms, LMB (2013) originally found that "[…] development gains are concentrated in the income and education sectors, and not in health" and that "[…] improvement in labor productivity […] rather than general equilibrium re-sorting appears to be the likely mechanism by which these development gains are realized" (LMB 2013, p.224, 200).

Revised results in Figure 2 using the seed proposed by LMS, however, show that little remains that can be confidently said about mechanisms. Except for the health-related infant mortality rate and the in-migration rate, none of the estimates are statistically significant at the 5% level.

Figure 2: Comparison of effect sizes for outcomes to test mechanisms





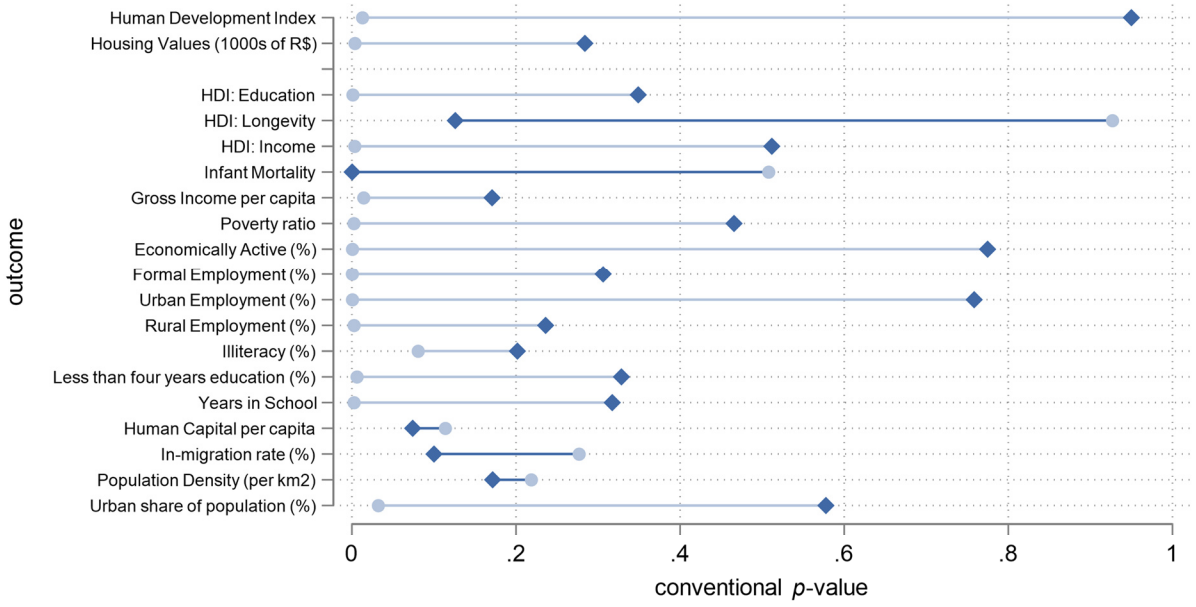
Note: ± represents indicators such as the poverty ratio, for which a decrease is considered positive, and vice versa.

3.4 Summary on effect sizes and p-values

The considerable loss in statistical significance that became apparent in Figure 1 and Figure 2 is further visualized in the dumbbell plot in Figure 3, which summarizes *p*-value differences for the outcomes in those figures between LMB (2013) and the present paper. While LMS (2021) found fewer statistically significant results than LMB (2013), they generally conclude that "the second stage results […] are similar to those originally reported" (LMS 2021: 1). Figure 3 emphasizes that this conclusion is not supported by the results emerging from our corrections to the code. Notably, the numbers below the figure indicate that none of the 14 outcomes that LMB (2013) identified as statistically significant remain significant in this replication when applying conventional *p*-values. The picture is very similar when applying the stricter definition of statistical significance at the 5% level using *tF*-adjusted standard errors.

Turning to the size of the effects paints a similar picture in that the economic significance of the effects decrease considerably. When we express differences in effect sizes between LMB (2013) and this replication in percentage points of the respective sample mean, the mean and median of these differences across the two main and 17 mechanism outcomes are 79 and 51 percentage points, respectively (not shown in the figure).

Figure 3: Changes in statistical significance for main outcomes and outcomes to test mechanisms



	# significant outcomes			Change in significance (10% level)			Abs. difference in *p*-values	
	** / ***	*	insignificant	sig→insig	insig→sig	insig→insig	mean	median
LMB (2013)	13	1	5	14	2	3	0.40	0.32
This replication	1	1	17					

*Note: *** / ** indicates statistical significance at the 1, and 5 percent level, respectively.*

4. Sensitivity testing framework

Section 3 presented revised main-specification results, where we adopted the main specification selected by LMS (2021). We now examine the sensitivity of results when other theoretically justified specifications are used. This sensitivity testing framework begins with a structured specification inventory to identify researcher-decision domains.

4.1 Step one: Structured specification inventory

We first go through each of the components of the following generic 2SLS estimation command to identify researcher-decision domains for the LMB/LMS case:

$$outcome = (endogenous\_treatment\_variable = IV)\ control\_variables\ if\_condition\ [weight].$$

The main objective is to identify researcher-decision domains that provide alternative theoretically justified specifications for specification curve analyses. Since the decision on these domains is inherently subjective and researcher-dependent, the very idea of this

framework is to make the underlying considerations as explicit as possible. For the case of LMB/LMS, we identify the following six domains:

- the choice of the *endogenous_treatment_variable*
- *control_variables* for the Amazon and other topographic factors
- *control_variables* for other infrastructure
- the choice regarding the inclusion of the Amazon region (*if_condition*)
- the definition of the Amazon region (affecting the *IV*, *control_variables*, and the *if_condition*)
- the *weight* used in the regressions.

Table 1 summarizes the rationale and challenge related to each of these six researcher-decision domains with alternative options to the choices underlying the specification adopted in LMS (2021). We extend discussions and sensitivity analyses that were partly already conducted in either LMB (2013), LMS (2021), or the follow-on paper by the original authors, Szerman et al. (2022*a*). In the same way as for the domains, the decision on *alternative choice options* within each domain is inherently subjective, and the framework is supposed to facilitate the transparent reporting on the theoretical relevance and substantive plausibility of these alternative choice options.

Table 1: Researcher-decision domains in LMB (2013) and LMS (2021)

Decision domain and rationale	LMB and LMS choices as basis of revised main specification	Alternative choice options and OFAT selection	
		#‡	Description
Endogenous treatment variable *Reflect electricity grid connections*	*Lagged electricity infrastructure* LMB/LMS define this variable as the share of a county within 50km of a transmission substation, lagged by one decade in the same way as the IV to reflect the time lag between grid availability and grid connection.	1	*Share electrified* The *share electrified* among houses in a county is an alternative that is "[…] likely measured with less error, and […] a direct measure of household-level connectivity" (LMB 2013, p.220).
Controls for topography *Account for potentially remaining time-varying differences*	*Quartic of Suitability Index x decade* After LMB used *Amazon x decade* interaction terms, LMS replace these by interactions of the *quartic of Suitability Index x decade*.	16	*Multiple options* LMS discuss 14 alternative options. We further checked interactions of decade dummies with the square and cubic of the suitability index. OFAT selections are the one originally used in LMB (2013) – *Amazon x decade* – and another randomly selected one, *water flow* x *budget*, where the latter refers to the lagged 10-year national budget available for hydroelectric dams.

Decision domain and rationale	LMB and LMS choices as basis of revised main specification	Alternative choice options and OFAT selection	
		#‡	Description
Controls for other infrastructure *Account for the possibility "that electricity proxies for a broader package of infrastructure investments" (LMB 2013, 219)*	*No controls for other infrastructure* LMB/LMS do not control for other infrastructure in the main specification.	2	*Lagged other infrastructure* & *other infrastructure* LMB control in a robustness check for *Lagged other Infrastructure*, namely water and sanitation access, as well as land slope and water trend as proxies for roads, all lagged by one decade. As an alternative, we propose a control set for *other infrastructure* that lags only the proxies for road, not the water and sanitation control variables, since the latter refer to effective access rates (as for houses electrified). Using this adjusted control variable set also increases the usable data points from the LMB dataset.
Inclusion or (partial) exclusion of the Amazon *Account for the prohibitive costs of electrification in the Amazon due to high material transport and low population density*	*Inclusion of entire Amazon* LMB/LMS include the *entire Amazon* in their analysis, which makes up more than half the area of Brazil. At the same time, there is not much variation in predicted treatment status in the Amazon. The IV changes in Amazon areas by only 0.1 percentage points over the entire observation period, compared to 12.4 p.p. in non-Amazon areas.	1+	*Exclusion of entire Amazon & Exclusion of (combinations of) individual Amazon states* The most obvious alternative is to drop *All Amazon States*. Szerman et al.'s (2022*b*) follow-on paper to LMB (2013) applies this sensitivity test, plus they drop (combinations of) individual Amazon states. This makes the sample surface area decrease by up to 59%, but the number of county observations by at most 6%, which dispels concerns about loss in statistical power. Our two OFAT selections are the *exclusion of all Amazon states* as the obvious alternative and the *exclusion of Pará*, the second largest state in the Amazon, as a randomly picked subregion.
Amazon definition *Account for a region that is "fundamentally different […] compared to the rest of Brazil, and […] plays an important role in the forecasting model" (LMB 2013, p. 214)*	*Amazon biome* LMS use the vegetation-based *Amazon biome* definition following the Brazilian Institute for the Environment and Natural Resources (IBAMA).	2	*Legal Amazon & extended Amazon* LMS discuss two alternative options, one administrative definition of the Amazon, known as *legal Amazon*, and an *extended Amazon* definition originally used by LMB based on Brazil's geo-political macro-regions.
Weighting of regressions *Adjust for factors that motivate weighting (or not) reviewed in Solon et al. (2015)*	*County area weights* LMB/LMS apply *county area weights.* They explain this with their IV being based on data at the level of the evenly spaced grid points and as "the number of grid points is not the same in each county" (LMB 2013, p.212).	2	*No weights & population weights* *No weights* and *population weights* are two alternatives that have also been used by other IV-based studies that assess infrastructure outcomes at a regional level of aggregation (Dinkelman 2011; Mettetal 2019). Opting for population weights is also justified in the present case since all outcome variables are population averages. *No weights* may be reasonable since the IV is already defined in relative terms – as the percentage of grid points within a county allocated with electricity – so that area weighting seems dispensable.

Note: ‡ # = number of alternative choice options

In many applications, not all researcher decisions can be included in a quantitative specification curve analysis, but some can be pondered qualitatively for a comprehensive appreciation of the quantitative results – what we call qualitative deliberation. We identify

two such decisions for the present setting, the *IV construction* and the definition of one of the *outcomes* adopted by LMB and LMS and discuss them further in Appendix B. While we encourage researchers to generally include these qualitative deliberations in the main body of their analyses, this is of subordinate importance to the main point of our paper, the spotlight on researcher decisions and specification screening.

4.2 Step two: Results exposition using specification curve analyses

4.2.1 Analytical approach

As can be inferred from Table 1, a myriad of specifications can be derived by combining the different alternatives for the respective researcher domain. Different handlings of the Amazon alone, which has proven to entail severe robustness issues in BPV (2021), in combination with the choice options for other domains easily leads to thousands of specifications. To present these alternatives in a structured, concise and still manageable way, we use a procedure that can serve as a blueprint for similar cases. It combines a simplified graphical analysis with a full-fledged specification curve analysis of theoretically justified specifications, with the former being a sniff test for the latter. For the simplified graphical analysis, we generate a one-factor-at-a-time (OFAT) sensitivity graph, where we take the main specification of LMS (2021) with our revised data as the point of departure to consecutively change one choice option from one researcher-decision domain.

For each of the six researcher-decision domains we select two alternative choice options to the one applied in LMS (2021). For two domains, *controls for topography* and the *inclusion of the Amazon*, we identified multiple alternatives and hence pick two alternative options, an obvious one and an additional random pick; for one domain, *endogenous treatment variable*, we only avail of one alternative (see OFAT selections in Table 1). We thus yield a tractable and more comparable set of $5 \times 2 + 1 \times 1 = 11$ alternative specifications. We refer to this as *OFAT graph* in the following, which is depicted in Figure 4, one for each of the two main outcomes. Yet since this is only a selection out of the multiverse of specifications, we conduct a full-fledged specification curve analysis in Appendix D. For this as well, we take the restricted set of choice options from the OFAT selection (note again that this does not exhaustively cover all possible combinations). We thus run $3^5 \times 2^1 = 486$ specifications, referring to our five researcher-decision domains with three choice options and the sixth researcher-decision domain with two options, each including the choice option applied in LMS (2021). We abstain from

distinguishing different combinations by their substantive plausibility but note that, in some contexts, certain combinations may be clearly more justifiable than others.

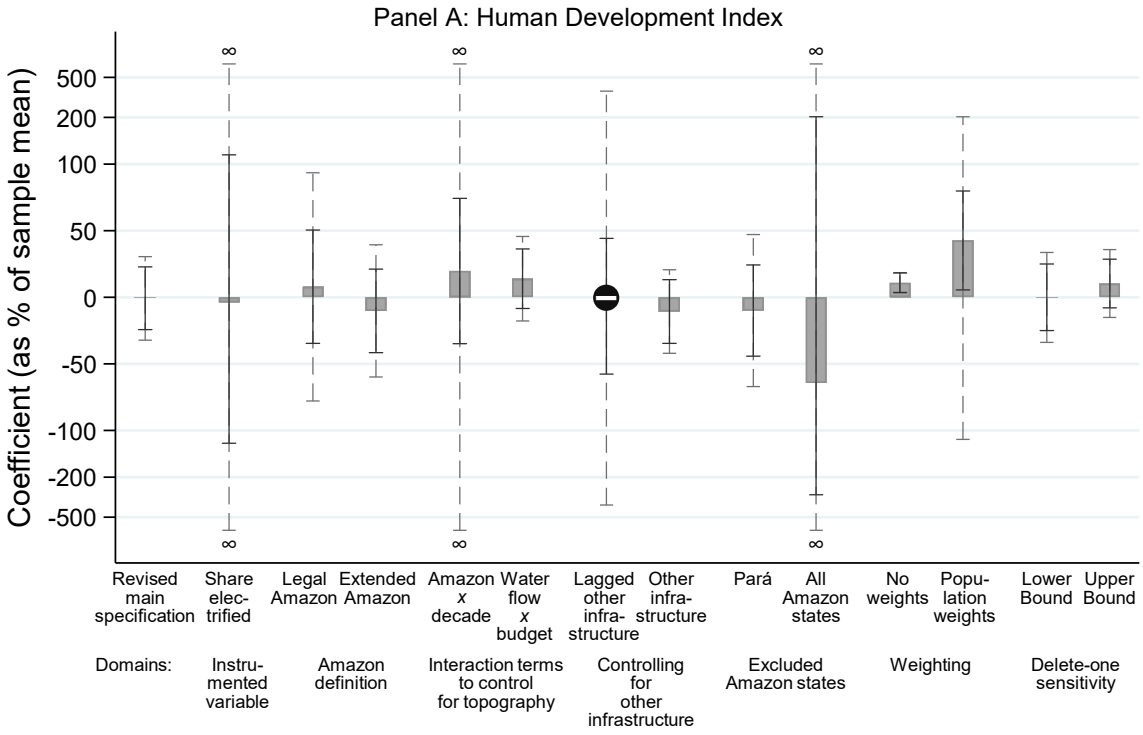
In the OFAT graphs, we show conventional and *tF*-adjusted confidence intervals and flag first-stages with a wrong sign, in our case a negative sign. We do not discard them entirely as suggested by Angrist and Kolesár (2022), to make explicit that theoretically justified specifications yield the wrong sign. Moreover, this allows us to use the simultaneous application of the *tF*-adjustments, which require abstaining from any pre-testing. As an additional sensitivity tests beyond those related to researcher-decision domains we propose the delete-X sensitivity test applied by Young (2022) in his analysis of IV applications in AEA publications. X refers to the number of clusters deleted. Here, we adopt the delete-one sensitivity to assess the change in estimates and *p*-values from deleting one cluster, that is one county, from the analysis sample.

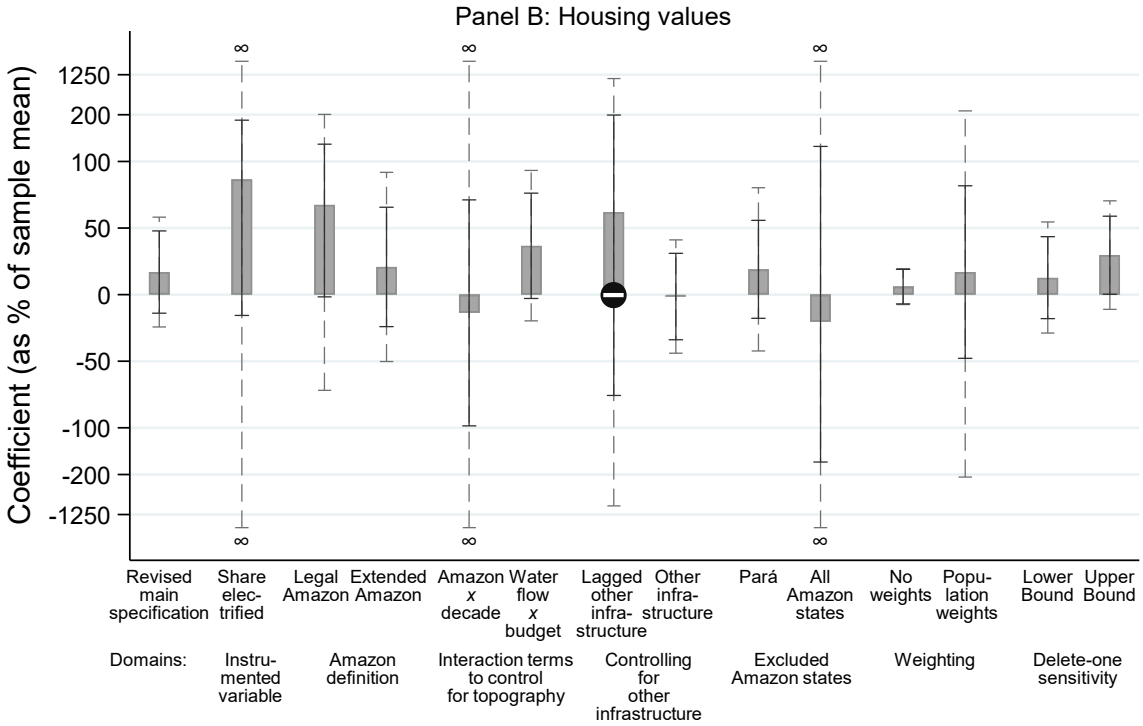
4.2.2 OFAT graphs and full-fledged specification curves

The two panels of Figure 4 show the OFAT graphs for the two main outcomes, HDI and housing values. The choices taken in the main specification for the individual researcher-decision domains presented in Table 1 are reproduced below the figure. Looking at the first-stage *F*-statistics in the OFAT graph, we find for all domains that theoretically valid specifications yield IVs that are considered weak according to the $F<10$ rule of thumb (Staiger and Stock 1997). All specifications in our OFAT graph have *F*-statistics below 104.7, the threshold identified by Lee et al. (2022) to guarantee a true 5 percent test.

Among the 11 alternative specifications in our OFAT graphs, one specification has a negative first stage (the one controlling for lagged other infrastructure, see Figure 4). In our full-fledged specification curve analysis, we find that 18% among the 486 specifications have a negative first stage, almost all (93%) of which include the lagged other infrastructure controls. We thus see a good share of specifications that would be discarded using the Angrist and Kolesár procedure, yet virtually all are induced by only one decision domain, other infrastructure, and here again by one specific way of controlling for it, namely lagged other infrastructure. This seems intuitively reasonable given that this specification is indeed most prone to endogeneity, as already acknowledged by the original authors who noted "[…] that the provision of other infrastructure may also be endogenous and we do not have exogenous instruments for their availability" (LMB 2013, p.219).

Figure 4: OFAT graphs with effect sizes across changes in different decision domain choices





***F*-Stat**	20.0	3.8	8.1	11.3	2.0	16.1	4.7	21.9	11.3	0.4	76.5	4.6	18.5	-.8 17.2
Decision domain choices in revised main specification		Lagged electricity infrastructure		Amazon biome		Quartic of suitability in-dex x decade		none		none		County area weights		–

95% confidence interval — 95% *tF*-adjusted confidence interval

Note: ∞ = estimates with F-Statistics below 3.84 have infinite confidence intervals; ⛔ = wrong first-stage sign. The specification used in LMS (2021) corresponds to the revised main specification, LMB (2013) used Amazon x decade to control for topography. Delete-one sensitivity as adopted by Young (2022).

As for second-stage results, all the estimates presented in Figure 4 are statistically insignificant and exhibit large confidence intervals. The only exception is the unweighted regression on the HDI in Panel A, which is robust to whether traditional confidence intervals are used or weak-IV robust ones. Young's delete-one sensitivity test shown at the right of Figure 4 furthermore exhibits a high degree of sensitivity of second-stage results: merely deleting one of the 2,182 Brazilian counties from the main-specification analysis causes the two outcomes to vary considerably; expressed in terms of variations in conventional *p*-values, the significance level of HDI estimates ranges between 0.27 and 0.99 and that of housing values between 0.05 and 0.42. Such sensitivity occurs because individual counties of large size – and thus with large data weights – experience small over-time changes in the continuous IV combined with large over-time changes in the outcome, creating a large lever in the estimations.

The findings from the OFAT graphs are confirmed by our full-fledged specification curve analysis in Appendix D, now including estimates for all potential, non-redundant combinations of choice options considered. HDI delivers relatively more statistically significant estimates where the original and revised estimate have the same sign (20% at the *tF*-adjusted 5% level), which almost all come from specifications where no weighting is applied in the regressions, a researcher decision that has not been considered in LMB (2013) and LMS (2021).

Accordingly, only one option of one of our researcher-decision domains, the omission of weights, yields results that provide indications for a quantitatively discernible impact with the given identification strategy. While the plausibility of this decision is debatable, it is worth noting that these weight options have not been considered in LMB (2013) and LMS (2021). In short, we are hard-pressed to find evidence of a consistently positive effect of electrification on either the HDI or housing values based on the specification curve analyses.

5. Conclusions

This article has done two things. First, we have reanalysed LMS (2021) and identified errors in the construction of the dataset. Upon correcting these errors, we obtain statistically insignificant estimates with much smaller effect sizes for main-specification results on the impact of electrification. Second, we have applied a sensitivity testing framework to the LMB/LMS identification approach, finding little robustness in estimates and little evidence for electrification effects. The framework's first step is a structured specification inventory that we

consider essential for any IV sensitivity testing. The second step includes specification curves and an OFAT graph that can be combined to summarize and present the sensitivity. This framework can be used in primary IV-based research to reduce leeway of researcher decisions (Breznau et al. 2022; Huntington-Klein et al. 2021), but also in similar replications like ours to reduce what Bryan et al. (2019) call "replicator's degrees of freedom".

Our framework does not cover concerns related to the IV's exclusion restriction, as they are discussed for a comparable case in Bensch et al. (2020). For LMB (2013), the exclusion restriction could be violated by water availability and land gradient spurring population growth and economic growth independently of electricity provision, for example via the suitability for coffee production. Coffee production is a key driver of economic development trajectories in Brazil (Peláez, 1972; Kruger, 2007). More generally, Lee et al. (2020*b*) note that "it is hard to rule out the possibility that the correlation between the instrument and the dependent variable runs through additional channels beyond electrification", which "raises questions about the validity of any geographic cost-based instrument". Such concerns have been increasingly voiced in the literature in recent years, also beyond electrification (see, for example, Lal et al. 2021, Mellon 2022, Gallen and Raymond 2023).

Nevertheless, IVs are often proclaimed as a second-best solution for a policy-relevant research area in which randomization is hardly possible. We argue that especially under this pragmatic perspective, all available due diligence tests should be applied. Our study cautions against an excessive focus on the results of one single main specification, and instead suggests that sensitivity testing should be at the centre of IV-based papers. Recent replication debates, like that between Spamann (2022) and Hayes and Saberian (2022), confirm that views across teams of researchers may differ strongly about the preferred main specification, as is also emphasized in Simonsohn et al. (2020). Our framework can serve to guide such debates and enable a fruitful exchange between replicators and original authors – or help to carve out the specific points of disagreement.

In terms of the overarching research question – what are the impacts of electrification on economic development – our results do not necessarily question the positive effect of electricity. Yet, they raise concerns about what one can learn about this relationship based on the adopted identification approach. They also raise the meta-scientific question of how the overall picture in the literature would change if our framework was applied to all high-quality studies in the field.

References

Andrews, I., Stock, J., & Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, *11*, 727-753.

Angrist, J., & Kolesár, M. (2022). One instrument to rule them all: The bias and coverage of just-ID IV. *arXiv:2110.10556v5*.

Bensch, G., Peters, J., & Vance, C. (2021). Development effects of electrification in Brazil – A comment on Lipscomb et al. (2013). *USAEE Working Paper* 21-529.

Bensch, G., Gotz, G., & Peters, J. (2020). Effects of rural electrification on employment: A comment on Dinkelman (2011). *Ruhr Economic Papers*, 840. Essen: RWI.

Bi, N., Kang, H., & Taylor, J. (2020). Inferring treatment effects after testing instrument strength in linear models. *arXiv:2003.06723v1*.

Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H., Adem, M., Adriaans, J., ... & Van Assche, J. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*, *119*(44), e2203150119.

Brodeur, A., Cook, N., & Heyes, A. (2020). Methods matter: *p*-hacking and publication bias in causal analysis in economics. *American Economic Review*, *110*(11), 3634-60.

Bryan, C. J., Yeager, D. S., & O'Brien, J. M. (2019). Replicator degrees of freedom allow publication of misleading failures to replicate. *Proceedings of the National Academy of Sciences*, *116*(51), 25535-25545.

Burlig, F. (2018). Improving transparency in observational social science research: A pre-analysis plan approach. *Economics Letters*, *168*, 56-60.

Christensen, G., & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, *56*(3), 920-80.

De Chaisemartin, C. & X. D'Haultfœuille (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: a survey. *The Econometrics Journal*, forthcoming.

Dinkelman, T. (2011). The effects of rural electrification on employment: New evidence from South Africa. *American Economic Review*, *101*(7), 3078-3108.

Gallen, T. & Raymond, B. (2023). Broken instruments. Available at: http://tgallen.com/Papers/Gallen_Raymond_BrokenInstruments.pdf.

Hall, A. R., Rudebusch, G. D., & Wilcox, D. W. (1996). Judging instrument relevance in instrumental variables estimation. *International Economic Review*, *37*(2), 283-298.

Heyes, A., & Saberian, S. (2022). Correction to "Temperature and Decisions: Evidence from 207,000 Court Cases" and Reply to Comment by Spamann. *American Economic Journal: Applied Economics, 14*(4), 529-533.

Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J. R., Burli, P., ... & Stopnitzky, Y. (2021). The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, *59*(3), 944-960.

Kranz, S., & Pütz, P. (2022). Methods matter: *p*-hacking and publication bias in causal analysis in economics: Comment. *American Economic Review*, *112*(9), 3124-3136.

Kruger, D. I. (2007). Coffee production effects on child labor and schooling in rural Brazil. *Journal of Development Economics*, *82*(2), 448-463.

Lal, A., Lockhart, M. W., Xu, Y., & Zu, Z. (2021). How much should we trust instrumental variable estimates in political science? Practical advice based on over 60 replicated studies. Available at SSRN: https://ssrn.com/abstract=3905329.

Lee, D. S., McCrary, J., Moreira, M. J., & Porter, J. R. (2022). Valid *t*-ratio Inference for IV. *American Economic Review*, *112*(10): 3260-3290.

Lee, D. S., McCrary, J., Moreira, M. J., & Porter, J. R. (2020*c*). Valid *t*-ratio Inference for IV. *arXiv:2010.05058v1*.

Lee, K., Miguel, E., & Wolfram, C. (2020*a*). Experimental evidence on the economics of rural electrification. *Journal of Political Economy*, *128*(4): 1523-1565.

Lee, K., Miguel, E., & Wolfram, C. (2020*b*). Does household electrification supercharge economic development? *Journal of Economic Perspectives*, *34*(1), 122-144.

Lipscomb, M., Mobarak, A. M., & Barham, T. (2013). Development effects of electrification: Evidence from the topographic placement of hydropower plants in Brazil. *American Economic Journal: Applied Economics*, *5*(2), 200-231.

Lipscomb, M., Mobarak, A. M., & Szerman, D. (2021). Another look at the precision of IV estimates of the development effects of access to electricity in Brazil. Available at: https://www.mollylipscomb.com/_files/ugd/a200af_629245100d8f46f592f171ada715c492.pdf.

Mellon, J. (2022). Rain, Rain, Go Away: 192 potential exclusion-restriction violations for studies using weather as an instrumental variable. Available at SSRN: https://ssrn.com/abstract=3715610.

Mettetal, E. (2019). Irrigation dams, water and infant mortality: Evidence from South Africa. *Journal of Development Economics*, *138*, 17-40.

Ofosu, G. K., & Posner, D. N. (2021). Pre-analysis plans: An early stocktaking. *Perspectives on Politics*, 1-17.

Peláez, C. M. (1972). História da industrialização no Brasil. São Paulo: ANPEC.

Reiff, L. O. & Reis, E. J. (2016). Estoque de capital em residências no Brasil (1970-1999). *Texto para Discussão* 2265. Rio de Janeiro: IPEA.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208-1214.

Solon, G., Haider, S. J., & Wooldridge, J. M. (2015). What are we weighting for? *Journal of Human Resources*, *50*(2), 301-316.

Spamann, H. (2022). Comment on "Temperature and Decisions: Evidence from 207,000 Court Cases". *American Economic Journal: Applied Economics, 14*(4), 519-528.

Staiger, D. & Stock, J. (1997). Instrumental variables regression with weak instruments. *Econometrica*, *65*, 557-586.

Szerman, D., Lipscomb, M., Mobarak, A. M., & Barham, T. (2022*a*). Replication data for IV modelling in LMB (2013) using MATLAB®.

Szerman, D., Assunção, J., Lipscomb, M., & Mobarak, A. M. (2022*b*). Agricultural productivity and deforestation: Evidence from Brazil. *Economic Growth Center Discussion Paper*, Yale University.

Young, A. (2022). Consistency without inference: Instrumental variables in practical application. *European Economic Review*, *147*, 104112.

Appendix A. Adjustments to LMB (2013) and LMS (2021) datasets

A1. Variable adjustments and aggregations

Some outcomes changed definition in the 1990's. These are the *HDI* and its three subcomponents *education*, *longevity*, and *income* as well as the *poverty ratio*, *illiteracy rate*, the *share of people with education less than four years*, and the *years of schooling*. For these outcomes y, the dataset includes for the decade 1990, $d = 9,$ variables with both the old and new definition, i.e. $y_{d=9}^{old}$ and $y_{d=9}^{new}$.

As noted for *HDI* in online appendix 3 of LMB (2013), the data overlap in that decade can be used to adjust data for the year 2000 ($d = 10$) by multiplying it by the following ratio as an adjustment factor:

$$y\_ratio_m = \frac{y_{m,d=9}^{old}}{y_{m,d=9}^{new}}.$$

Here, m refers to the unit of observation of the raw data, which is the municipality. This data is later aggregated at the level of a county, c.

LMB (2013) and LMS (2021) made this adjustment for decade 10 to all outcomes listed above except for *HDI*. Furthermore, they used the new-definition variable of the *HDI* in decade 9 as well. In the revision as part of this replication, we therefore adjust the *HDI* for decade 9 and 10 in the same fashion as all other variables with changing definition in the 1990's. This is summarized in Table A1.

Table A1: HDI variable adjustment

		Variable definition applied		**mean (sd) of $hdi_{c,d}$, the HDI at county level**	
Variable		LMB & LMS	Revision	LMB & LMS	Revision
HDI in decade 9 and 10	$hdi_{m,d=9}$	$hdi_{m,d=9}^{new}$	$hdi_{m,d=9}^{old}$	0.626 (0.098)	0.567 (0.145)
	$hdi_{m,d=10}$	$hdi_{m,d=10}^{new}$	$hdi_{m,d=10}^{new} \times hdi\_ratio_m$	0.709 (0.081)	0.639 (0.137)

We further considered a minor correction necessary that was not made by LMB (2013) and LMS (2021). It again affects the variables listed above for which definition changed in the 1990's. Since the number of municipalities increased from about 4500 to about 5500 in that decade, no data from before the creation of the municipality was available for about 1000 municipalities, mostly small municipalities with less than 20,000 inhabitants (Brandt 2010). Hence, no adjustment factor could be derived and the decade-10 variables for these municipalities were set to missing by LMB (2013) and LMS (2021). However, in aggregating the municipality data to the county level, LMB (2013) and LMS (2021) used the population of

all municipalities in deriving population-weighted averages. This obviously led to understating some indicators. We therefore only considered those municipalities in deriving population-weighted averages for which the respective indicator was non-missing. This is again summarized in tabular form in Table A2.

The table also shows an alternative approach: instead of discarding the decade-10 data of so many municipalities because y_ratio cannot be determined, one may approximate a municipality's adjustment factor from other municipalities in the same county. A plausible approximation is the population-weighted average of municipality-level adjustment factors that are available in the same county. We tried this and found the resulting county data to differ only marginally. Correlation between variables for decade 10 defined according to the two approaches always exceeds 0.998. Accordingly, we abstain from showing the sensitivity to these different definitions and apply the first approach above, which is closer to how LMB (2013) and LMS (2021) handled this issue.

Table A2: Aggregation of municipality data

		Variable definition applied	
Variable		LMB & LMS	Revision
Aggregated municipality data	$y_{c,d}$	$\sum_m y_{m=o,d} \times \frac{pop_{m=o,d}}{\sum_m pop_{m=n}}$	$\sum_m y_{m=o,d} \times \frac{pop_{m=o,d}}{\sum_m pop_{m=o}}$
Variables with changing definition, for which y_ratio can not be determined	$y_{m=u,d=10}$	$y^{new}_{m=u,d=10}$	$y^{new}_{m=u,d=10} \times \overline{y\_ratio}_m$, with $\overline{y\_ratio}_m = \sum_{-m} y\_ratio_{-m} \times \frac{pop_{-m=o,d}}{\sum_m pop_{-m=o}}$

Note: $m = u$ refers to municipalities with unknown adjustment factor; $m = o$ to municipalities, for which data is available for the respective outcome y; $m = n$ represents municipalities with available population data and $-m$ other municipalities in the same county as municipality m.

For two variables – *income* as HDI sub-component and the *share of electrified households* – the adjustments caused the values from 2000 to exceed the maximum possible value of either 1.00 or 100%, both in LMB (2013) and the revision. This happened in less than five percent of counties. We decided to top-code these cases to 1.00 and 100%, respectively.

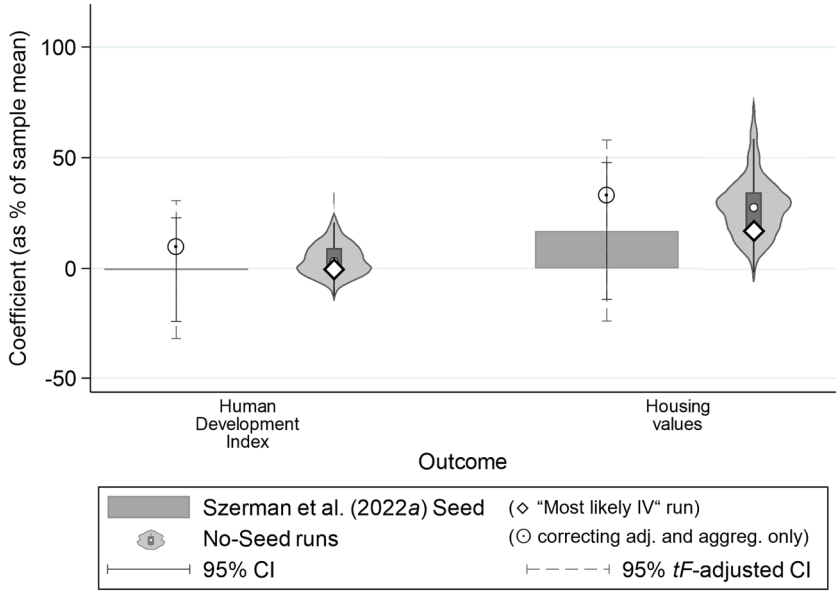
A2. IV construction

Both LMB (2013) and LMS (2021) use simulations to construct the IV, that is the lagged hypothetical electricity grid coverage rates in individual counties, which may change from decade to decade. LMS (2021) note that setting a randomly selected seed as starting point in the grid simulation algorithm at the centre of the IV construction is required to make exact replication of their results possible. The replication code in Szerman et al. (2022*a*) includes such

a seed but using this seed does not replicate the results in LMS (2021). The IV values used in the analysis presented in LMS (2021) and those resulting from the setting of the seed included in the code of Szerman et al. (2022*a*) differ notably: depending on the decade, they differ in 17 to 18 percent of counties, and in these cases, IV values measuring the extent of counties covered by the hypothetical grid differ by 63 to 64 percentage points, on average.

Figure A1 underpins the consequentiality of setting a seed and the fact that running the IV construction multiple times without setting a seed yields different IVs. The figure reproduces the point estimates of the two main outcomes when using the seed included in Szerman et al. (2022*a*), which are also shown in the main text of this replication. In addition, violin plots show the distribution of estimates for 100 runs when not setting a seed. This distribution of estimates goes along with conventional *p*-values ranging from 0.05 to 0.99 and from 0.00 to 0.92 for HDI and housing values, respectively. The reason for this variability in IV outputs and estimates seems to lie where the algorithm searches for the least-cost electricity network. For the first decade, for example, 480 sub-stations have to be placed. On a grid with around 32,500 points, this makes $\binom{32500}{480}$ or 7×10^{1083} potential combinations. LMS' algorithm runs with 80,000 iterations, which in light of the immensely high number of potential combinations seems insufficient to spot the single least-cost network configuration and thus to yield sufficiently consolidated results.

Figure A1: Sensitivity due to non-convergence in IV modelling outputs



Note: The figure presents 95% confidence levels both based on conventional t-ratio critical values and those based on the tF adjustment procedure proposed by Lee et al. (2022).

For different reasons, which we outline below under *alternative IV modelling*, we do not see that the convergence problem can be solved in the given simulation framework. There, we also discuss alternative approaches including one, where we condense the information from the 100 simulation runs undertaken as part of our analysis into what we refer to as the "most likely

IV". This approach cannot overcome the convergence problem either, and does therefore not represent a viable alternative to the IV adopted by LMB/LMS.

Meanwhile, the point estimates when applying this "most likely IV" shown in Figure A1 are quite close to the estimates using the seed proposed by Szerman et al. (2022*a*), thus lending support to the use of the LMS-seed simulation for the main-specification results. Similarly, at least the HDI estimate using the seed proposed by Szerman et al. (2022*a*) is very close to the median estimate of the estimates from the no-seed runs.

Lastly, Figure A1 shows point estimates when only the adjustment and aggregation issues are corrected. Comparing these estimates to those from LMS (2021), it becomes clear that the setting of a seed affects both point estimates, whereas the adjustment and aggregation issues primarily play a role for HDI.

A3. Miscellaneous minor corrections made by LMS (2021) to the LMB (2013) dataset

The LMS (2021) dataset includes a few minor adjustments to the LMB (2013) dataset that we also incorporated in the dataset used for our analysis. First, the LMS (2021) dataset includes one county less, because data is missing for this county to construct the suitability index used in the new main specification. Second, LMS (2021) removed grid points that could not be linked to municipality data before running the probit regressions underlying the IV. In LMB (2013), this was only done at a later stage. Accordingly, the coefficients of the probit regressions are slightly different, as can be seen in Appendix C. Third, according to notes in the replication code of Szerman et al. (2022*a*), LMS (2021) rectified a few errors in the MATLAB® code used by LMB (2013) to model the IV, such as wrong budgets defining the number of plants to be built by decade.

There is only one minor change by LMS (2021) to the LMB (2013) dataset, where we sticked to LMB's (2013) choice: LMS (2021) deviate from using the IPEA data for the variable *county area*, which is used for the weighting of regressions and the calculation of the outcome *population size*, and instead use the county area determined based on GIS data from 2000 for all decades.

Alternative IV modelling

Our replication exposed that the setting of a seed (random starting point) in the IV construction is consequential. Abstaining from setting a seed, it follows that each run of the grid simulations at the centre of the IV construction yields different IV output and, in turn, different impact estimates for the outcome indicators. There are different reasons why we do

not see that the convergence problem can be solved in the given simulation framework. For example, it seems infeasible to substantively reduce the number of potential combinations (e.g. by considering only buffers of grid points around the sub-stations and thereby reducing the upper index in the binomial coefficient, the set of potential electricity grid locations linked to a sub-station). Alternative modelling assumptions related to non-geographic factors (such as population agglomerations) may be introduced as well, but these would work against the ambition to let only geographic and engineering parameters determine the electricity network.

Recent work on electrification impacts applies simple algorithms borrowed from the transportation infrastructure (Faber 2014) to solve the least-cost grid optimization problem. However, these solutions are not transferable to the LMB/LMS case, because those works can take the terminal points of the grids as given, whereas LMB/LMS rightly consider them as endogenous choice parameters. For example, Kassem (2021) studies the effects along the electricity grids that interconnect existing and thus far disconnected colonial power plants in Indonesia, and Budjan (2022) studies a programme that densified the grid between existing sub-stations in Nigeria.

Against this background, we consider another approach that seeks to condense the information from the multiple runs of the grid simulations into *one* IV instead of trying to fix the simulation in the first place. We thereby intend to arrive at the most likely IV, that is the most likely development of the hypothetical electricity grid. A straightforward representation of this is the IV in a simulation run that shows the smallest absolute difference to the IV values from all other simulation runs. Concretely, we first determine the pairwise absolute IV difference across simulation runs, $\Delta pIV^{r,-r}$, aggregate them across all pairs to the absolute IV difference for a specific simulation run, $\Delta IV^{r}(-r)$, and drop the IV from the simulation run with the highest absolute IV difference (see also Table A3 below). We then follow the same procedure by recalculating $\Delta IV^{r}(-\check{r})$ for the narrowed down set of other simulation runs, $-\check{r}$, and consecutively drop the IVs with highest $\Delta IV^{r}(-r)$ until only three simulation runs remain, from which we select the one with the lowest $\Delta IV^{r}(-r)$.

Table A3: Components of the algorithm to determine the "most likely IV"

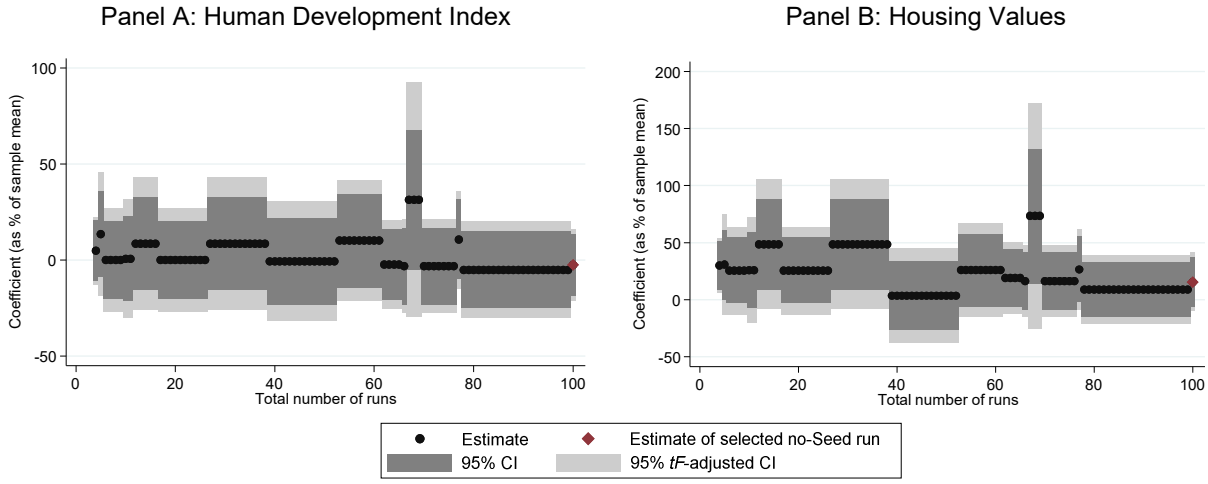
Variable		Formula
Pairwise absolute IV difference across simulation runs *r* and *-r*	$\Delta pIV^{r,-r}$	$\sum_c \sum_d \lvert IV^{r}_{c,d} - IV^{-r}_{c,d} \rvert$
Absolute IV difference for simulation run *r*	$\Delta IV^{r}(-r)$	$\sum_{-r} \Delta pIV^{r,-r}$

Note: c refers to counties as the unit of observation and d to the decade.

We tried this approach using 100 runs of the IV modelling. Figure A2 shows estimates for the two main development outcomes according to the IV selected after a certain number of runs of the IV construction. We see that the estimates continue to change also after dozens of runs

and that they do not clearly converge. This approach is, hence, not clearly preferable, but estimates of the selected run at least fall well into the middle range of all estimates. Given the extent of the underlying convergence problem outlined under A.2 above, we generally do not see that this approach can be useful, even if thousands of runs would be performed.

Figure A2: Estimates with "most likely IV" depending on the total number of IV modelling runs



Additional references

Brandt, C. T. (2010). A criação de municípios após a Constituição de 1988. *Revista de Informação Legislativa*, *47*(187), 59-75.

Budjan, A. (2022). Move on up – Electrification and internal migration. Paper presented at the annual conference of the Verein für Socialpolitik 2022: Big Data in Economics, ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg.

Faber, B. (2014). Trade integration, market size, and industrialization: evidence from China's National Trunk Highway System. *Review of Economic Studies*, *81*(3), 1046-1070.

Kassem, D. (2021). Does electrification cause industrial development? Grid expansion and firm turnover in Indonesia. *CRC TR 224 Discussion Paper 52*.

Appendix B. Qualitative assessment of additional researcher decisions

In the present case, much of the sensitivity analysis related to researcher-decision domains could be integrated in the quantitative sensitivity analyses, already yielding a clear and consistent picture of the study results. In other cases, qualitative deliberations may play a larger role in gauging the strength of the presented evidence. Here, we briefly touch on a researcher-decision domain that can be covered in the present case through qualitative deliberations only:

- *Outcome*, to represent development results of electrification,
- *IV construction*, to reflect the hypothetical electricity network determined by geographic factors alone to avoid endogeneity.

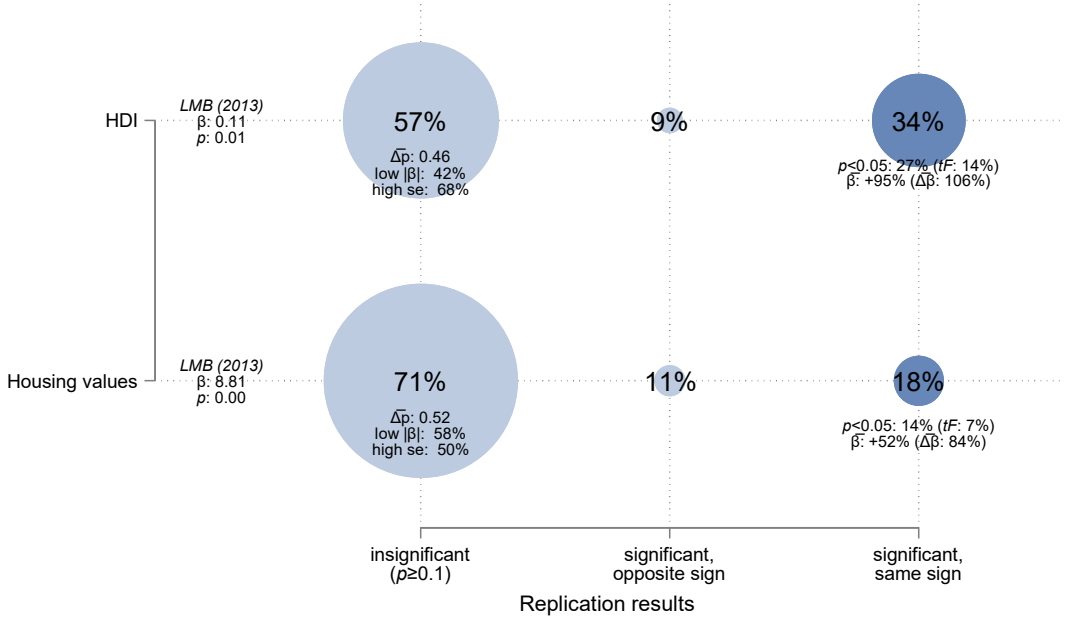
As one of their two main development *outcomes*, LMB/LMS use a housing value variable provided by the public research institution IPEA. The sensitivity question on this outcome is less about alternative options, but about its added value in generating evidence on the underlying research question in the first place. This added value is debatable since the analysis in LMB (2013) merely extrapolates results from a cross-sectional analysis of sample data from the end of the study's observation period over space and time. IPEA constructed the variable based on a hedonic model using data from rented houses in the 1999 National Household Sample Survey (PNAD). The hedonic model finds a statistically significant estimate of 0.24 for the electricity access attribute, which is then used to simulate the housing value variable over time and space (see Reiff and Reis 2016).

Regarding the *IV construction*, LMB/LMS use a cost-minimizing algorithm to predict the hypothetical electricity network that, however, does not converge and yield a distribution of IV's (see Section 2 and Appendix A). This issue does not present alternative choice options for an assessment of the researcher-decision domain *IV construction* in specification curve analyses in the main text. Nevertheless, we may go beyond the seed included in the code of Szerman et al. (2022*a*) underlying the specification curve analyses to capture the distribution of IV's in a complementary sensitivity analysis. For that purpose, we further increase the considered multiverse of specifications by the results of the 100 no-seed runs of the grid simulation described under *Alternative IV modelling* in Appendix A. With these runs always using the biome definition of the Amazon, this provides $100 \times 3^4 \times 2^1 \times 1^1 = 16{,}200$ specifications in addition to the 486 specifications used for the full-fledged specification analysis in the main text.

Figure B1 summarizes the information in what we refer to as a sensitivity dashboard, with the two main outcomes on the y-axis. The results can be interpreted as a more representative picture of the replicability of original results in that they are not conditioned on the seed from the code of Szerman et al. (2022*a*). For the housing values indicator, we find that 18% of

specifications yield statistically significant estimates at the 10% level, and 7% at the *tF*-adjusted 5% level. Excluding specifications with negative first stages, which would have been dropped according to sign screening proposed by Angrist and Kolesár (2022), these shares would be 12% and 4%, respectively, and thus close to what one would expect to occur by chance (not shown in the figure).

Figure B1: Sensitivity dashboard for main outcomes with extended multiverse of specifications



Note: Based on 16,686 specifications for each of the two outcomes. Δ = mean absolute deviation; β = estimate; se = standard error; low |β| (high se) refers to the share of specifications where the revised |β| (se) is sufficiently low (high) to turn the overall estimate insignificant. Macrons (upper bars) indicate mean values.

For the second main outcome, HDI, we find somewhat higher shares of statistically significant estimates – 34% at the 10% level and 14% at the *tF*-adjusted 5% level – while none of these estimates is based on specifications with negative first stages. These significant estimates differ substantively from the original estimate in that their mean is 95% higher and their mean deviation from the original estimate amounts to 106% of the original estimate. Insignificant estimates for this outcome are far from statistical significance, with a mean deviation in the *p*-values of 0.46, where the loss of significance is rather due to higher standard errors than to lower estimates.

To conclude, additional ambiguities exist that affect the strength of the evidence presented in LMB (2013) and LMS (2021). This is in line with the key findings from the quantitative sensitivity analyses, where a broad array of specifications indicates that there is no result that is robust across equally plausible specifications in the present setup.

Appendix C. Comparison of probit regressions underlying the instrumental variable

LMB (2013) use a probit regression with hydropower geographic cost parameters to construct the "suitability index" as basis of the subsequent grid simulations to construct the IV. Column (1) of Table C1 reproduces the estimation results from table 1 of LMB (2013). Columns (2) and (3) show results for the same specification with the LMS (2021) dataset, which corresponds to the data used for this paper. Results in Column (2) using also the same sample area of entire Brazil differ slightly from those in Column (1) because LMS (2021) already removed grid points that could not be linked to municipality data. In LMB (2013), this was only done at a later stage.

Importantly, we find that estimates are very similar when excluding the Amazon as done in Column (3). Notwithstanding the importance of the Amazon at different parts of LMB's analytical framework, this suggests that grid points located in the Amazon do not drive the probit estimates as basis of the modelled IV. Hence, the inclusion of the Amazon does not affect the predicted hydropower suitability of non-Amazon grid points, which represent the area where most of Brazil's population is living.

Table C1: Probit regression for hydropower geographic cost parameters

Dataset	LMB (2013)	LMS (2021)	
Coverage	entire Brazil	entire Brazil	non-Amazon Brazil
	(1)	(2)	(3)
Log of maximum flow accumulation	0.029**	0.018	0.021
	(0.014)	(0.014)	(0.017)
Average slope in the river	0.044	0.036	0.037
	(0.030)	(0.030)	(0.031)
Maximum slope in the river	0.062***	0.061***	0.057***
	(0.012)	(0.012)	(0.013)
Amazon indicator	-0.753***	-0.632***	–
	(0.066)	(0.050)	
Indicator for location has a river	-0.030	0.010	0.002
	(0.063)	(0.064)	(0.075)
Number of Observations	33,342	32,608	13,541
Pseudo R2	0.117	0.115	0.052

*Note: *** / ** indicates statistical significance at the 1, and 5 percent level, respectively.*

Appendix D. Specification curve analysis

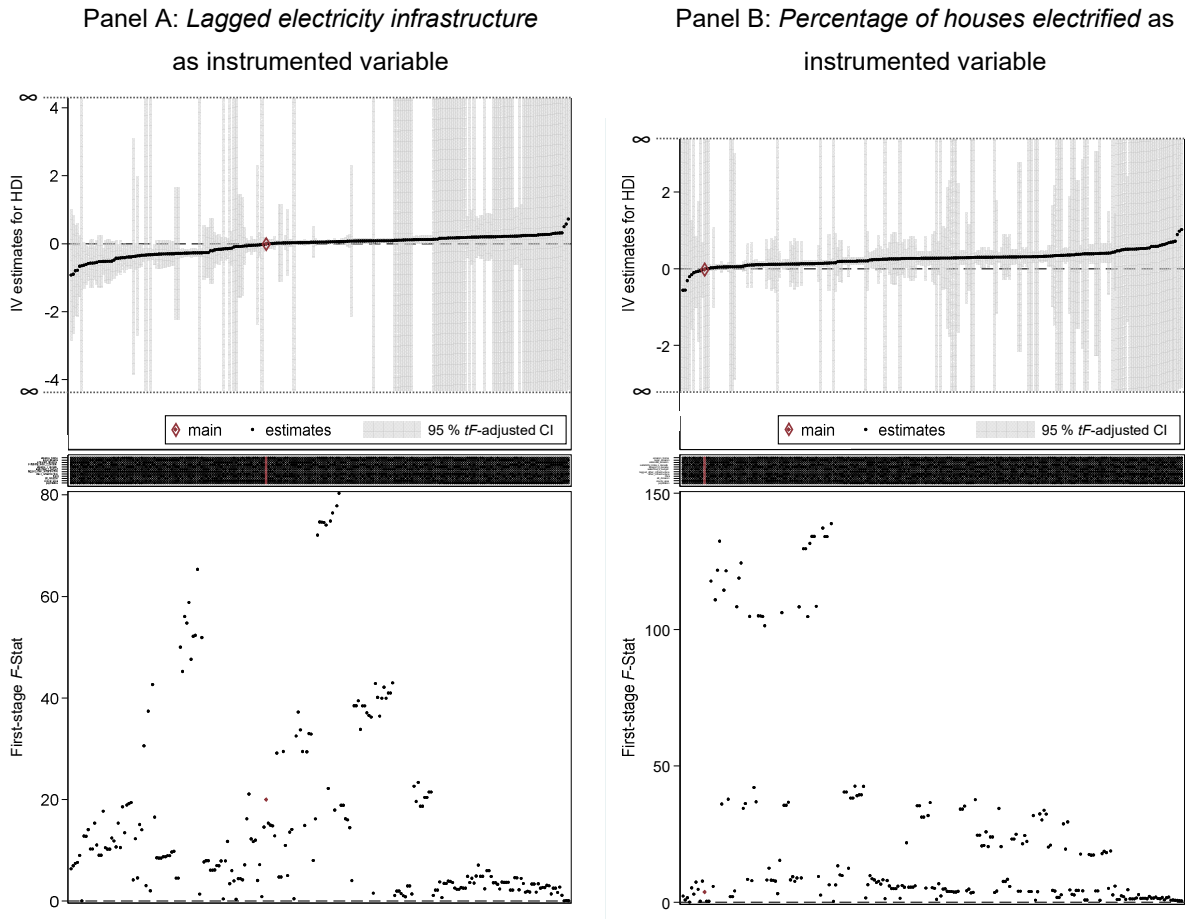
Figure D1 and Figure D2 show specification curves for the two main outcomes HDI and housing values, each split into two panels according to the instrumented variable used. Each panel includes 243 estimates according to the following set of combinations:

3 (Amazon definitions) x 3 (Interactions with controls for topography) x 3 (Controls for other infrastructure) x 3 (Exclusion of parts of the Amazon) x 3 (Weighting of regressions) = 3^5 = 243.

We find the following:

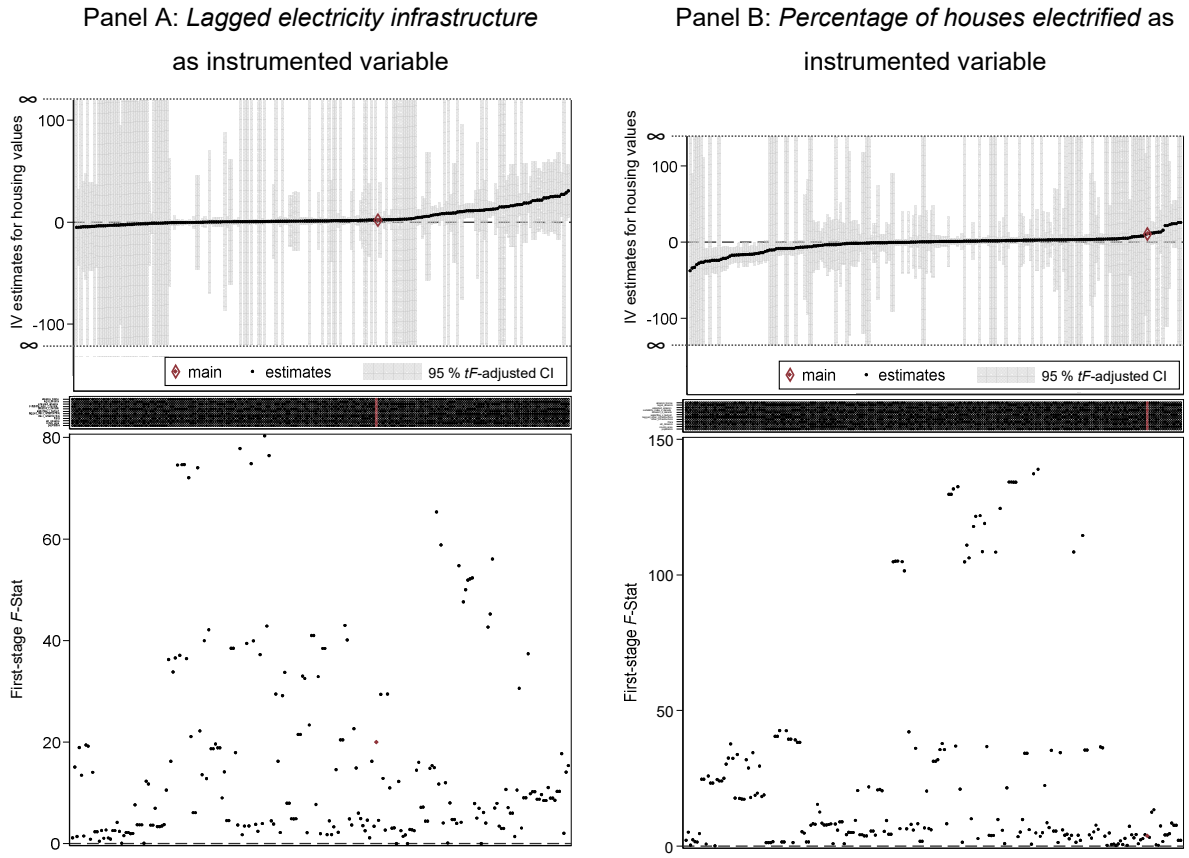
- **Instrument strength and effect sizes**: the stronger the first stage, the smaller the effect size; conversely, the larger the estimates are, the more variance they have, often having infinite confidence intervals for the more extreme estimates
- **Effect sizes with main vs. other valid specifications**: estimates for the main specifications are located fairly in the middle of the specification curves when the instrumented

Figure D1: Specification curve for the outcome HDI



Note: Graphs generated using the Stata command speccurve prepared by Martin E. Andresen. To improve readability, the lowest and highest five estimates were removed from each panel figure, all of which have both high point estimates and confidence intervals (CIs). These graphs show raw coefficients of IV estimates instead of coefficients expressed as percentages of sample mean, as presented in the main text.

Figure D2: Specification curve for the outcome Housing Values



Note: To improve readability, the lowest and highest ten estimates were removed from each panel figure, all of which have both high point estimates and confidence intervals (CIs). These graphs show raw coefficients of IV estimates instead of coefficients expressed as percentages of sample mean, as presented in the main text.

variable is the *lagged electricity infrastructure*, implying that other valid specifications equally yield lower and higher estimates; conversely, when the instrumented variable is the *percentage of houses electrified*, estimate for the main specification is at the lower end of the curve for HDI as outcome and at the higher end for housing values as outcome.

To get a better sense of potential patterns of factors or combination of factors in these specification curves, we zoom into those estimates that are found significant. Starting with the stricter definition of statistical significance at the *tF*-adjusted 5% level, we find that all statistically significant coefficients where the original and revised estimate have the same sign come from either

- specifications where no weighting is applied in the regressions; these make up 20% [0%] of all specifications for the outcome HDI [housing values]; here, estimates turn out to be mostly positive
- specifications that control for other infrastructure and where – like in a robustness check by LMB (2013) – all these other infrastructures are lagged by one decade; these make up

2% [3%] of all specifications for the outcome HDI [housing values]; first stages are virtually always negative when using these controls.

Figure D3 and Figure D4 show specification curves with conventionally calculated confidence intervals. Another 28% [13%] of estimations are significant at the conventional 10% level, most of them either with *F*-Statistics below 10 (18% [3%]), negative first stages (7% [6%]), or again without weighting (2% [1%]).

Figure D3: Specification curve for the outcome HDI, significant estimates only

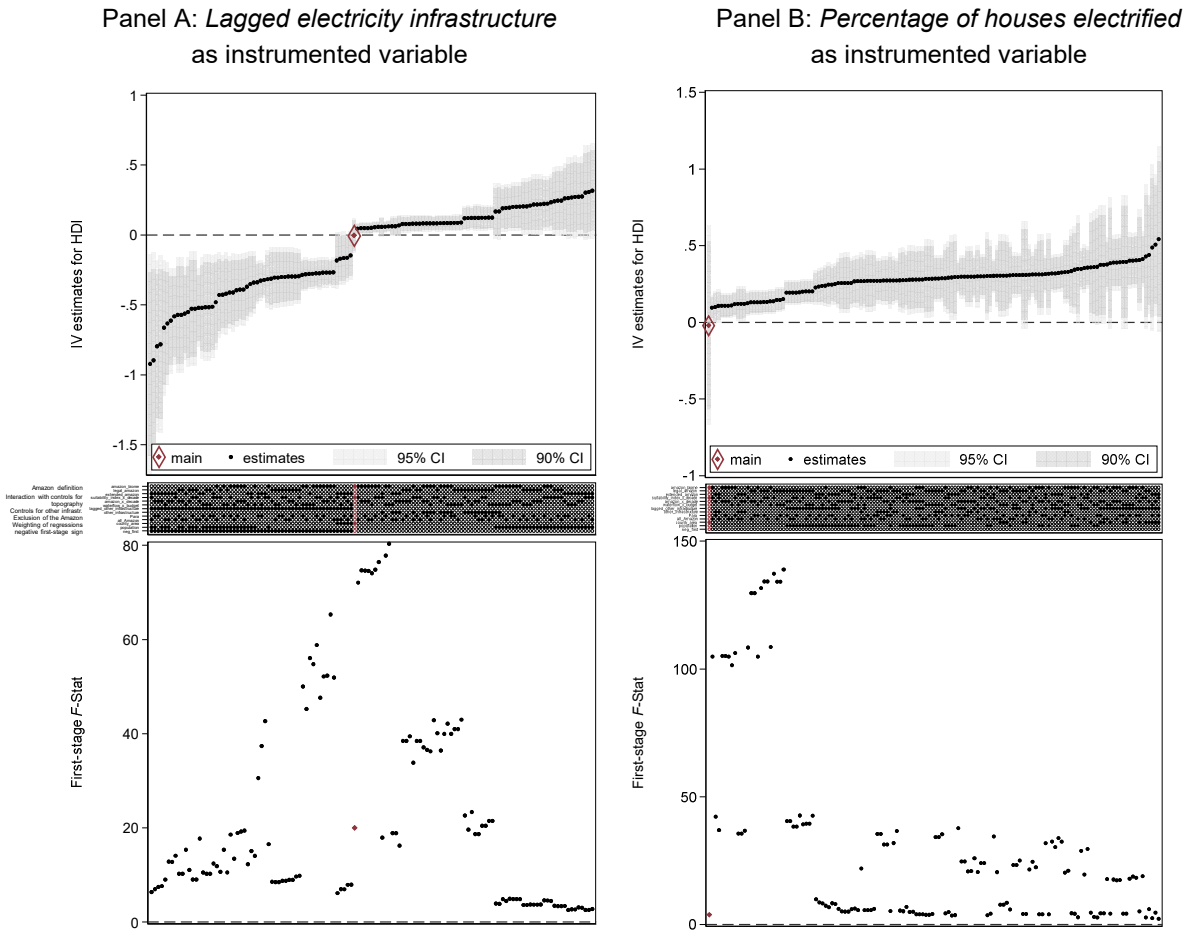
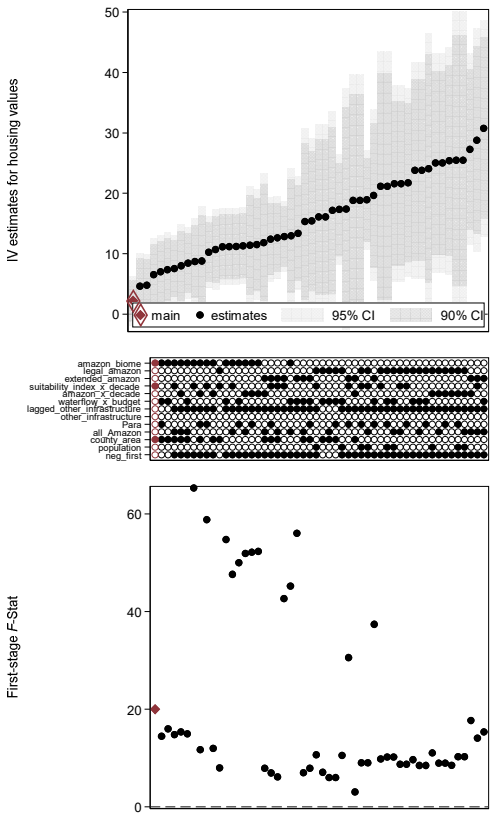


Figure D4: Specification curve for the outcome Housing Values, significant estimates only

Panel A: *Lagged electricity infrastructure* as instrumented variable



Panel B: *Percentage of houses electrified* as instrumented variable