Data Visualization & Latex

Prof. Susanna Berkouwer

Pre-doc Training, August 2023

Communicating research

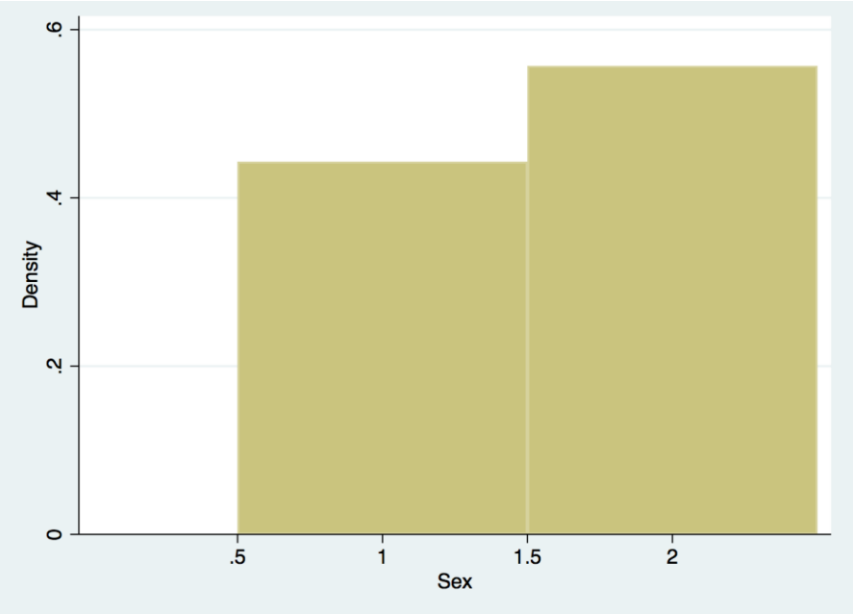
Prof. Susanna Berkouwer

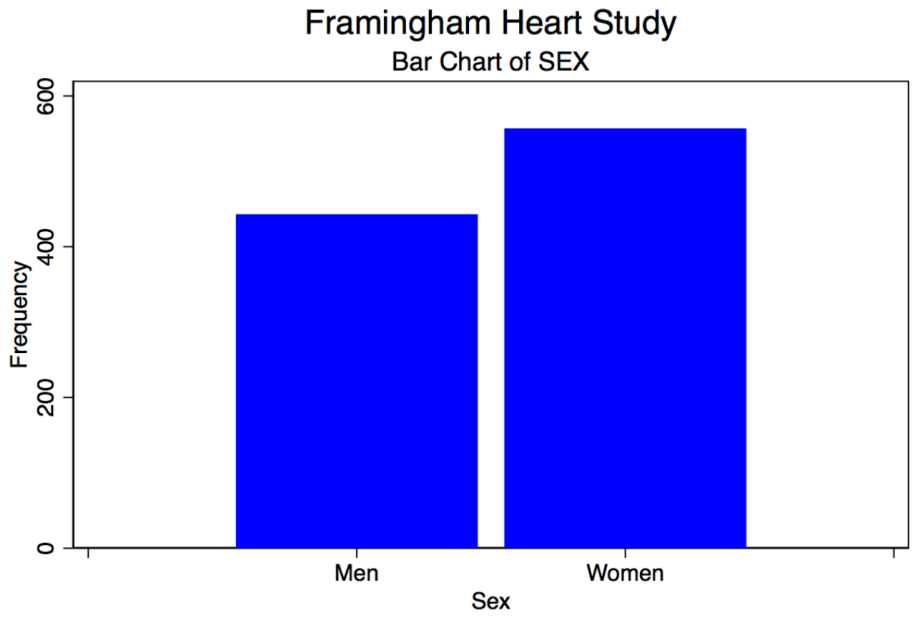
Pre-doc Training, August 2023

Today

1. **Graphing**
   a) Labelled: x-axis variable name, y-axis variable name, title
   b) Scaled: y-axis should (almost*) always start at 0
   c) Footnoted: How was the variable created?
2. Tables
3. Latex
4. Organizing your research output

What should a histogram convey?



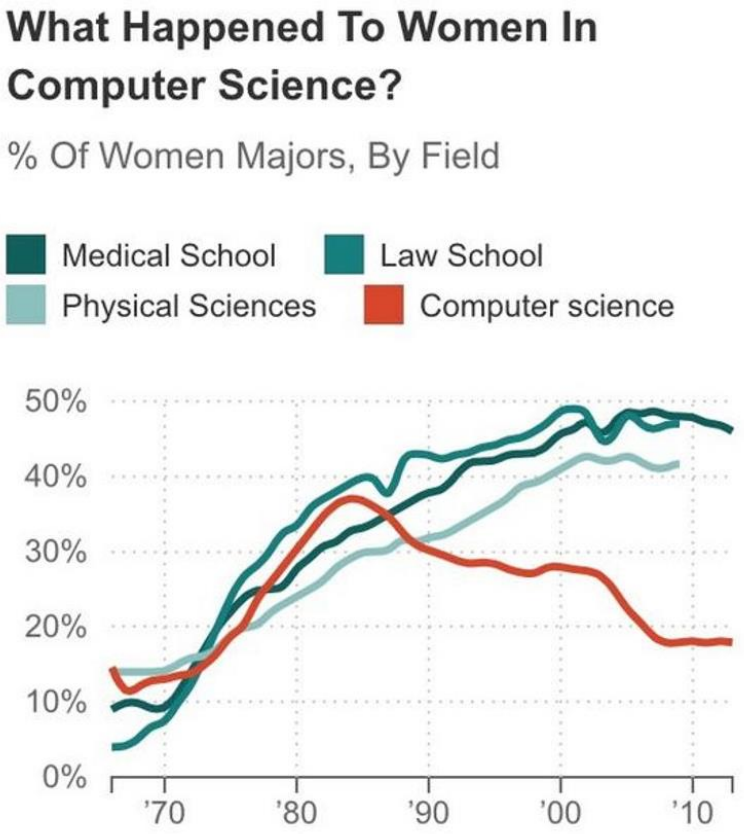


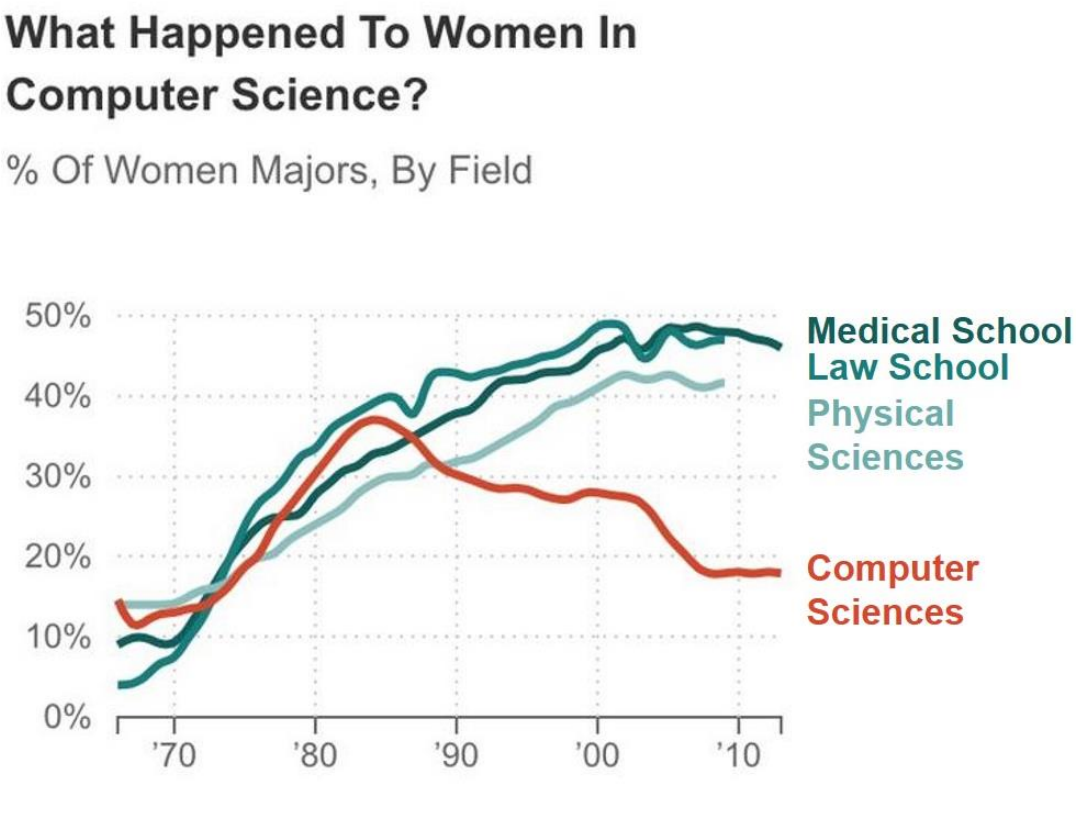
Data Visualization w Stata (UMASS guide)

- What variable is on the x-axis? Is this continuous or discrete?
- What is the sample size?
- Where does the data come from?
- What do the values mean?

Wharton
University of Pennsylvania

Label lines with quick legibility in mind





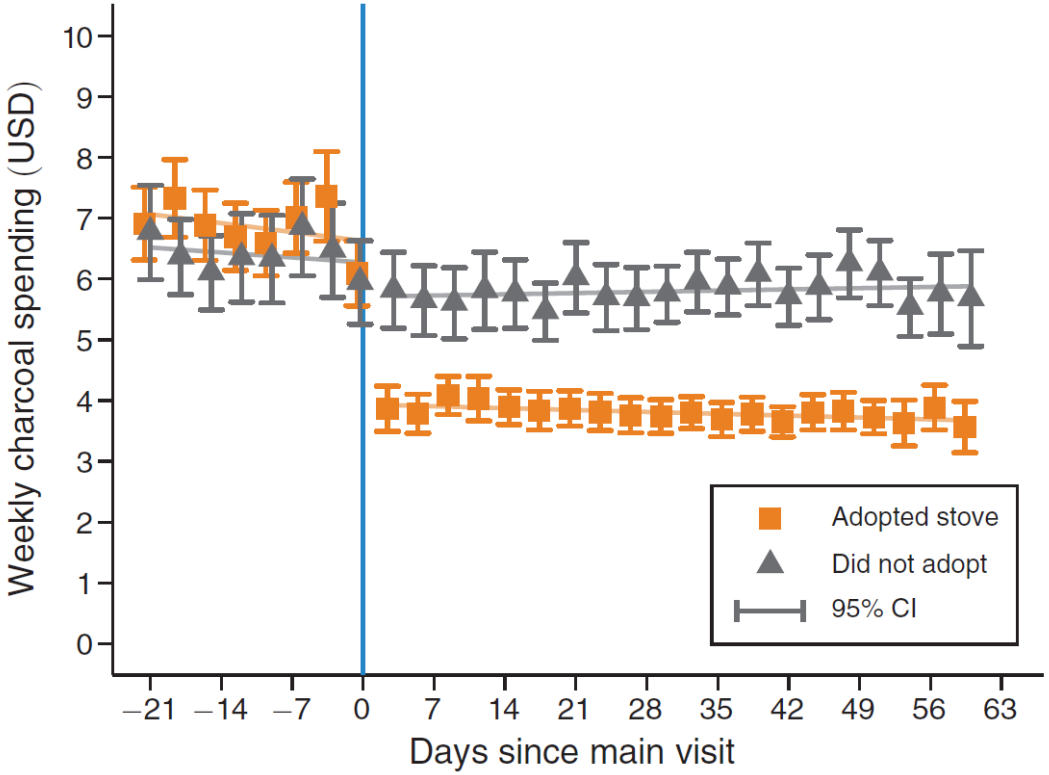
Ann K. Emery, depict data studio
(https://depictdatastudio.com/directly-labeling-line-graphs/)

Wharton
University of Pennsylvania

Other important notes

- Graphs should be legible for color-blind folks and with a grayscale printer
  - https://colorbrewer2.org/#type=qualitative&scheme=Dark2&n=3

Coding in practice *(Berkouwer & Dean; AER, 2022)*



- 1 key take-away: **“after main visit, spending dropped for adopters”**
- 1 key point in time:
  - ‘Main visit’
- 2 key groups:
  - Adopters
  - Non-adopters
- 3 key components:
  - Means (by day + adoption)
  - Confidence Intervals
  - Lines of best fit

Wharton
University of Pennsylvania

Footnotes convey crucial information about variables

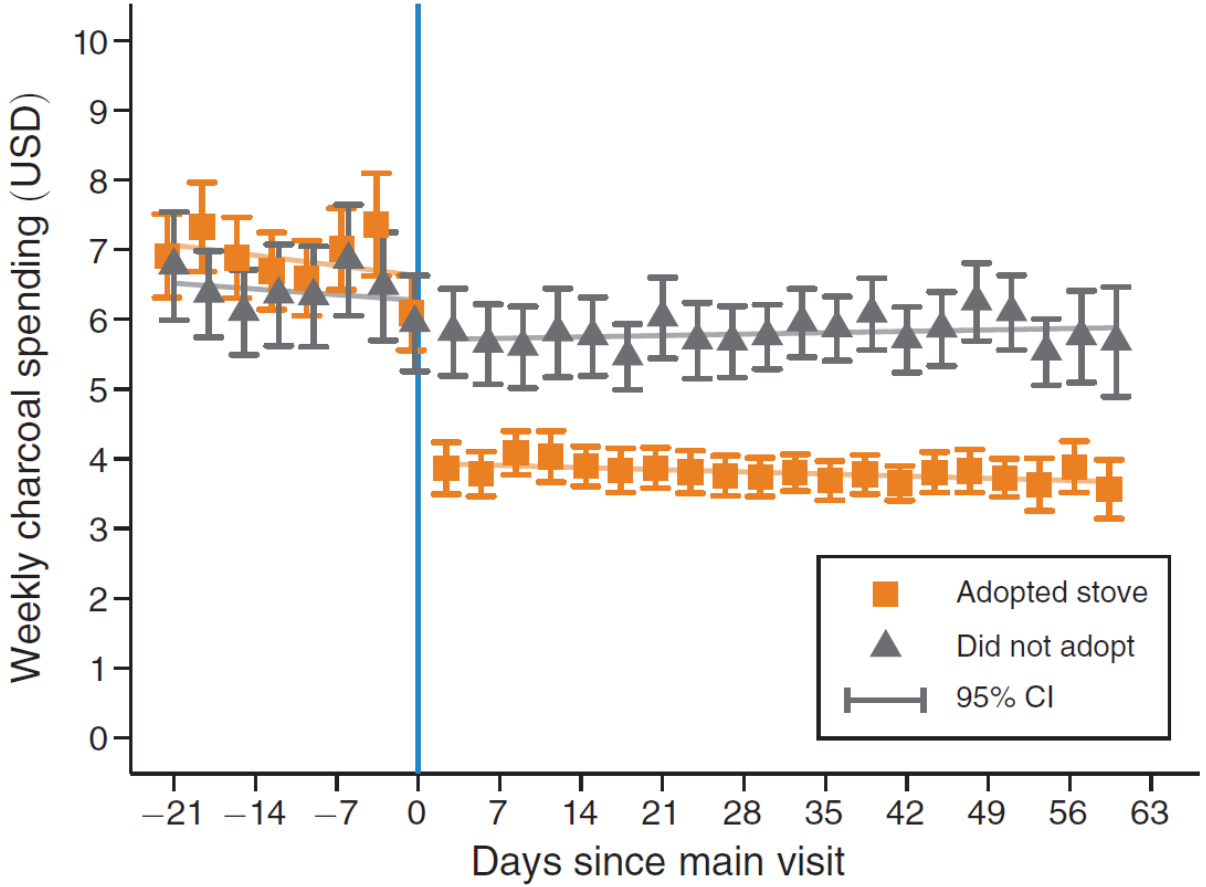
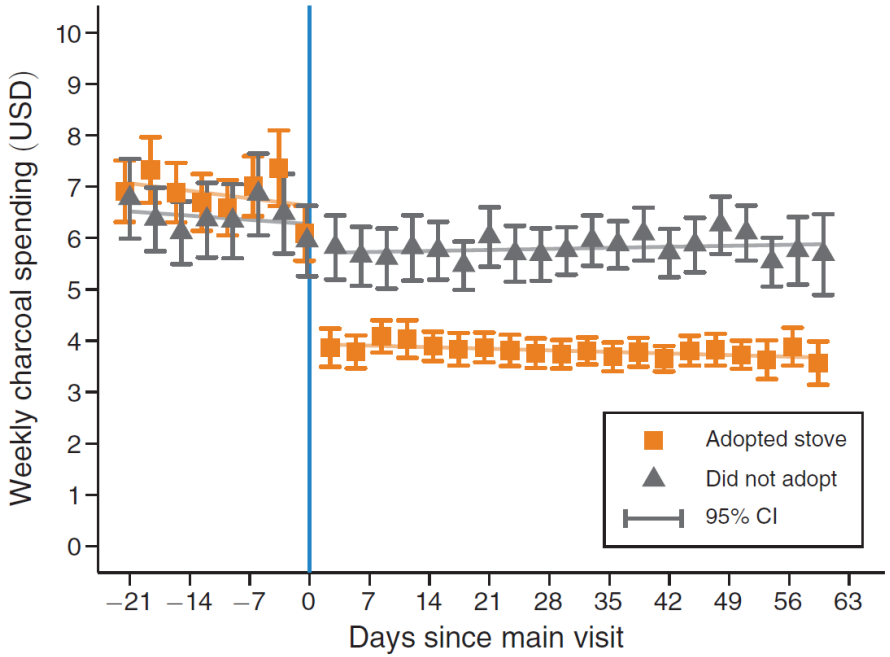


FIGURE 4. ENERGY EFFICIENT STOVES REDUCE ENERGY SPENDING

Notes: Weekly charcoal spending by adopters and nonadopters of the energy efficient stove before and after the main visit (visit 2). Charcoal spending is elicited through a recurring three-day SMS survey. Adoption of the stove causes charcoal expenditures to drop by US$2.28 per week (39 percent relative to the control group). The causal estimates presented in online Appendix Figure C3 are similar.

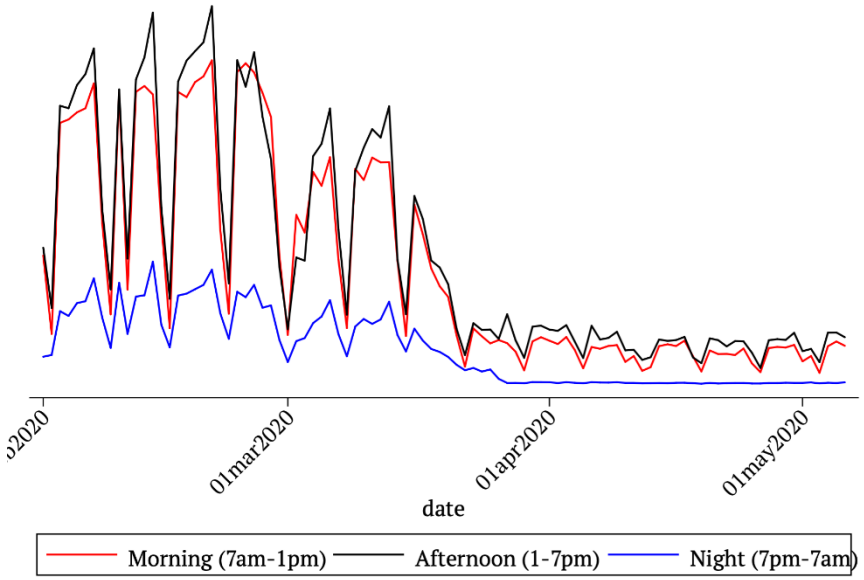
(Ask your PI / other pre-docs for code)

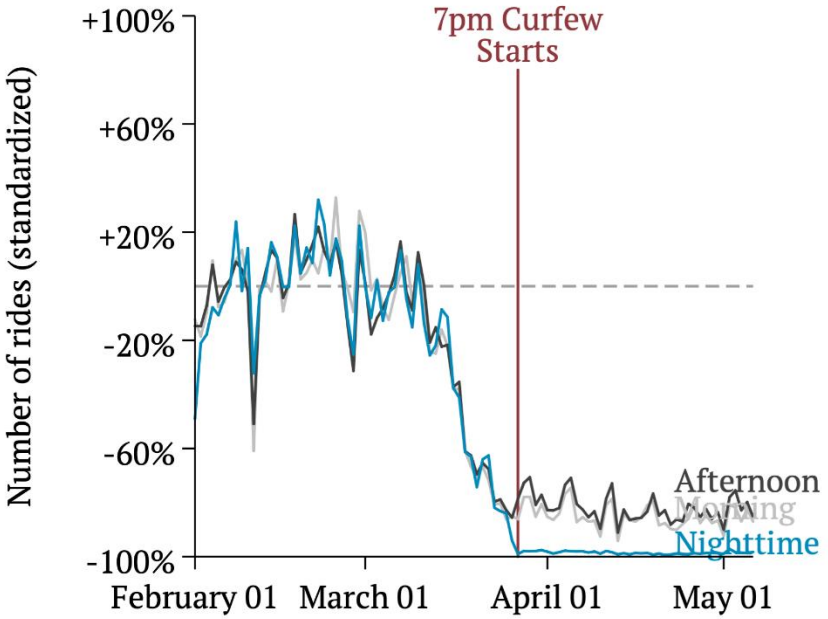



```
twoway (lfit mean CsinceV2 if jikokoa==0 & CsinceV2<=0, lcolor(gs10) connect(l)) ///
 (lfit mean CsinceV2 if jikokoa==0 & CsinceV2> 1 , lcolor(gs10) connect(l)) ///
 (lfit mean CsinceV2 if jikokoa==1 & CsinceV2<=0, lcolor(dkorange*0.5) connect(l)) ///
 (lfit mean CsinceV2 if jikokoa==1 & CsinceV2> 1 , lcolor(dkorange*0.5) connect(l)) ///
 (rcap ub lb Csince if jikokoa==1, mcolor(dkorange dkorange) lcolor(dkorange dkorange)) ///
 (rcap ub lb Csince if jikokoa==0, mcolor(gs6 gs6) lcolor(gs6 gs6)) ///
 (scatter mean CsinceV2 if jikokoa==1, mcolor(dkorange) msymbol(S) lcolor(dkorange) lpattern(dash) connect(none)) ///
 (scatter mean CsinceV2 if jikokoa==0, mcolor(gs6) msymbol(T) lcolor(gs6) lpattern(shortdash) connect(none)), ///
 `white' scale(1.5) ///
 xline(0, lcolor(midblue)) ///
 xsize(2.5) ysize(1.2) ///
 yscale(range(0 `l')) ///
 ylabel(#10, nogrid angle(0)) xlabel(-21(7)64, angle(0)) ///
 ytitle("Weekly charcoal spending (`c')") ///
 xtitle("Days since main visit") ///
 legend(cols(1) pos(3) label(8 "Did not adopt") label(7 "Adopted stove") label(6 "95% CI") order(7 8 6) region(lcolor(white)) )
```

Wharton
University of Pennsylvania

Label lines with quick legibility in mind

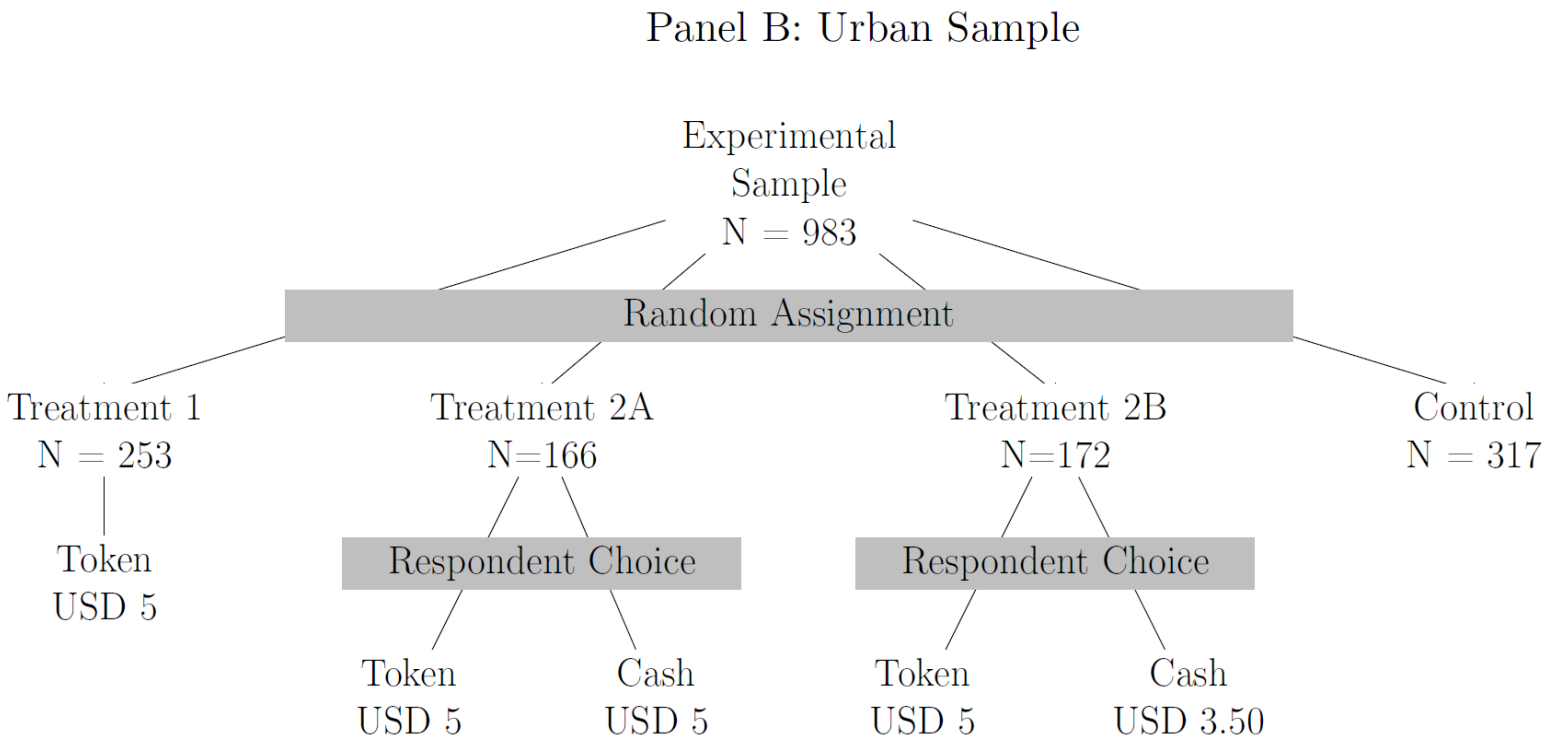




Berkouwer & Eggers (2020)

Wharton
University of Pennsylvania

Tikz for graphs (Latex)



Panel A displays the non-experimental sample as well as treatment assignment within the experimental sample for rural households. Panel B displays treatment assignment among urban households. Attrition causes modest discrepancies between groups and subgroups.

Berkouwer, Biscaye, Hsu, Kim, Lee, Miguel, Wolfram (2021)

Wharton
University of Pennsylvania

Tikz for graphs (Latex)

```
\begin{tikzpicture}[label distance=10mm]
\tikzstyle{every node}=[font=\Large]
    \begin{pgfonlayer}{nodelayer}
        \node [style=none, fill=black] (0) at (-6.25, -0.25) {};
        \node [text width=3cm, align=center, fill=white] (1) at (-6.25, -0.75) {Treatment 2A \\ N=166};
        \node [text width=3cm, align=center, fill=white] (2) at (-12, -0.75) {Treatment 1 \\ N = 253};
        \node [text width=3cm, align=center, fill=white] (3) at (0.5, -0.75) {Treatment 2B \\ N=172};
        \node [text width=3cm, align=center, fill=white] (4) at (-8, -4.25) {Token \\ USD 5}; %  \\ N = 6 \\ (4.5\%)
        \node [text width=3cm, align=center, fill=white] (5) at (-4.75, -4.25) {Cash \\ USD 5}; %  \\ N = 127 \\ (95.5\%)
        \node [text width=3cm, align=center, fill=white] (6) at (-1.25, -4.25) {Token \\ USD 5 }; %  \\ N = 26 \\ (17.8\%)
        \node [text width=3cm, align=center, fill=white] (7) at (2.25, -4.25) {Cash \\ USD 3.50}; %  \\ N = 120 \\ (82.2\%)
        \node [text width=3cm, align=center, fill=white] (21) at (-12, -2.75) {Token \\ USD 5};
        \node [text width=3cm, align=center, fill=white] (12) at (6, -0.75) {Control \\ N = 317 };
        \node [text width=3cm, align=center, fill=white] (40) at (8, 3.5) {~};
        \node [style=none] (8) at (-12, -0.25) {};
        \node [style=none] (9) at (0.5, -0.25) {};
        \node [style=none] (11) at (6, -0.25) {};
        \node [text width=3cm, align=center, fill=white] (10) at (-3, 2.5) {Experimental Sample \\ N = 983 };
        \node [text width=13cm, align=center, fill=lightgray] (91) at (-3, 0.75) {                                                    Random Assignment
                                };
        \node [text width=5cm, align=center, fill=lightgray] (92) at (-6.25, -2.5) {       Respondent Choice     };
        \node [text width=5cm, align=center, fill=lightgray] (93) at (0.5, -2.5) {       Respondent Choice     };
    \end{pgfonlayer}
    \begin{pgfonlayer}{edgelayer}
        \draw (0.center) to (1);
        \draw (1) to (4);
        \draw (1) to (5);
        \draw (3) to (6);
        \draw (3) to (7);
        \draw (9.center) to (3);
        \draw (8.center) to (2);
        \draw (0.center) to (10);
        \draw (10) to (11.center)
        \draw (11.center) to (12)
        \draw (8.center) to (10);
        \draw (10) to (9.center);
        \draw (21) to (2);
    \end{pgfonlayer}
\end{tikzpicture}
```

Define labelled nodes

Some invisible nodes too:

Add decision rectangles

Connect nodes with lines

Wharton
University of Pennsylvania

Today

1. Graphing
2. **Tables**
3. Latex
4. Organizing your research output

Wharton
University of Pennsylvania

What should a table convey?

TABLE 2—CAUSAL IMPACT OF STOVE ADOPTION ON WEEKLY CHARCOAL SPENDING

	OLS	First Stage	IV Estimate (1-month endline)			IV Estimate (1-year endline)	
	(1) USD	(2) Bought Stove	(3) USD	(4) IHS(USD)	(5) IHS(KG)	(6) USD	(7) IHS(USD)
BDM Price (USD)	0.01	-0.03					
	(0.01)	(0.00)					
WTP (USD)	-0.01	0.02	-0.00	-0.00	0.00	0.01	0.00
	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.02)	(0.00)
Bought Cookstove (=1)	-1.89		-2.28	-0.50	-0.48	-2.50	-0.56
	(0.28)		(0.29)	(0.07)	(0.08)	(0.42)	(0.09)
Observations	7853	913	7853	7853	796	6979	6979
Control Mean	5.72		4.97	2.16	1.55	5.30	2.21
Socioeconomic controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Data Source	SMSes	Midline	SMSes	SMSes	Buckets	SMSes	SMSes

Note: Results from an instrumental variables regression that uses the (randomly assigned) BDM price as an instrument for stove adoption to estimate the causal impact of adoption on weekly charcoal expenditures. Columns (1) and (2) present the OLS and first-stage estimates, respectively. Column (3) uses weekly charcoal expenditures in USD as the outcome variable. Column (4) uses the inverse hyperbolic sine (IHS) conversion of the USD amount. A 0.50 IHS reduction corresponds to a 39 percent reduction relative to the control group. Column (5) uses the IHS of the weight of the charcoal bucket one month after stove adoption as the outcome variable. Columns (6) and (7) conduct the same analyses as columns (3) and (4) respectively, but using data from the SMS survey conducted one year after the main visit. Socioeconomic controls include baseline savings, income, risk aversion, credit constrainedness, number of adults and children. In regressions using SMS data, errors are clustered by respondent. SE in parentheses.

Berkouwer & Dean (AER, 2022)

Wharton
University of Pennsylvania

What should a table convey?

- Data source
- How variables were constructed
- Sample size
- Regression specification
  - e.g. OLS, IV
- Outcome variable
- Independent variables
- Other controls (fixed effects, controls)
- Significance
  - Stars (some journals don't allow this anymore!)
  - Standard errors

Today

1. Graphing
2. Tables
3. **Latex**
4. Organizing your research output

Installing LaTeX

1. Install a TeX distribution (TeX Live, MiKTeX, or MacTeX)
2. Install an editor (Lyx, TeXmaker, or other)
3. Export to PDF

Installing LaTeX

~~1. Install a TeX distribution (TeX Live, MiKTeX, or MacTeX)~~

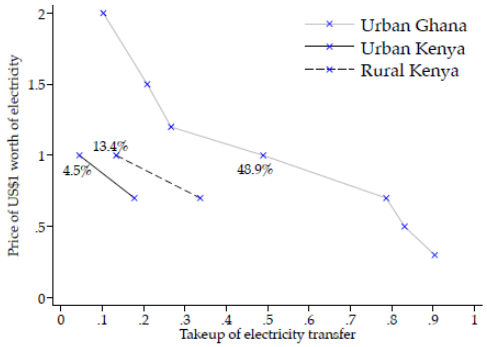
~~2. Install an editor (Lyx, TeXmaker, or other)~~

~~3. Export to PDF~~

1. Just use Overleaf.com

Calling the graphs & tables you made

Figure 2: Demand for Electricity Transfers by Context



Demand for USD 1 of electricity expressed in USD of mobile money. The numbers shown indicate the share of respondents who prefer USD 1 of electricity to USD 1 in mobile money in an equal trade-off, by context. The choice is incentivized in Kenya, as respondents receive the transfer they choose, while in Ghana the choice is presented as hypothetical.

mobile banking service to a mobile money account corresponding to a phone number provided by the respondent. Electricity transfer tokens were first purchased at a local Kenya Power office,[7] and the token ID was then sent to respondents via SMS, who could then enter it into their meter to activate the credit.

4 Results

In Kenya, households overwhelmingly prefer mobile money, and are willing to forego on average 40% of value to receive mobile money instead of an electricity transfer. By contrast, in Ghana, respondents are on average willing to forego 4% of value to receive an electricity transfer instead of mobile money. Electricity transfers modestly increase electricity usage in Kenya and decrease electricity spending and transactions in Ghana, but have limited impacts on broader socioeconomic outcomes.

4.1 Preferences for electricity and mobile money

Figure 2 displays take-up of electricity expressed as the implicit choice price indicating the value of the mobile money transfer offer relative to the electricity credit transfer offer. In Kenya, when the implicit price of USD 1 of electricity is USD 1 of mobile money, 5% of urban participants take up the electricity transfer (choosing USD 5 in electricity over USD 5 in mobile money). Even

[7] To prevent fraud, each token is tied to a respondent's Kenya Power account number.

Table 3: Government relief and energy and other consumption, Ghana

	N	Control Mean (SD)	Relief in last 30 days (SE)	Relief before last 30 days (SE)
Electricity usage since previous survey (kWh)	246	168.64	-24.46	-104.91**
		(137.38)	(50.45)	(53.48)
Electricity usage since previous survey (USD)	246	18.43	-2.67	-11.46**
		(15.01)	(5.51)	(5.84)
Electricity spending in past month (USD)	2312	15.20	-1.49*	-1.32
		(13.49)	(0.83)	(0.98)
Current balance on prepaid meter (USD)	759	5.05	0.23	0.01
		(7.01)	(0.88)	(1.15)
Pre-paid topups in last 30 days	2000	1.90	-0.19*	-0.10
		(1.40)	(0.11)	(0.13)
Average topup amount in last 30 days (USD)	2004	10.21	-0.13	-0.41
		(9.56)	(0.57)	(0.71)
Total consumption in the past 7 days (USD)	2349	113.27	0.18	-0.63
		(110.53)	(6.92)	(8.41)
Energy spending (excl. electricity) in the past 7 days (USD)	2325	4.39	0.09	0.13
		(6.37)	(0.52)	(0.58)
Food spending in the past 7 days (USD)	2329	32.63	0.44	1.82
		(25.97)	(1.69)	(1.84)
Worried about having enough food in past 7 days (=1)	2350	0.22	-0.04	-0.04
		(0.41)	(0.03)	(0.03)
Number of days in past 7 that adults skipped meals	2348	0.86	-0.13	0.02
		(1.88)	(0.12)	(0.13)

Estimates of Equation 2 for energy and consumption outcomes in Ghana. From left to right, the columns show the number of observations, the control mean, and the effects of receiving government electricity relief in the last 30 days and of previously receiving relief but not in the last 30 days (relative to never receiving relief). Timing of transfer receipt is not randomly assigned, so these estimates should be interpreted as suggestive correlations. Regressions include household and week fixed effects, and we cluster SEs at the household level. FDR q-values for all coefficients are statistically insignificant at the $\alpha = 0.10$ level.

for the rural sample, pooling survey rounds 2 and 3 to increase statistical power. Table 3 estimates equation 2 for Ghana.

In Kenya, electricity transfers increased electricity usage by 30 kWh (worth USD 3) for the rural sample and 43 kWh for the urban sample. They also led to higher electricity meter balances at endline for both groups, and to decreases in recent electricity spending. These effects are large: relative to the control group, treated households increased electricity usage by 66% in the rural sample and 105% in the urban sample.

On the other hand, the T2A treatment arm had no effect on electricity usage among urban respondents—95.5% of whom chose mobile money. Given that electricity is storable and transaction costs for using mobile money to purchase electricity tokens are low, there is no reason *ex ante* for the electricity transfers in T1 to generate larger increases in electricity usage. Instead, mental accounting, or increased attention to electricity usage due to the treatment, might account for increased electricity consumption among T1.

In Ghana, electricity expenditure falls by USD 1.50 in the last 30 days for respondents that received electricity relief in that time period. This reduction is less than the median amount received in the last 30 days among recipients (7 USD), suggesting recipients are increasing electricity consumption relative to non-recipients. Pre-paid meter balance at the time of the survey increases, but unlike in Kenya this is not significant—households appear to be largely consuming rather than storing their transfers. This indicates that monthly relief transfers were generally inframarginal to

Calling graphs you made (ask your PI / other pre-docs for code)

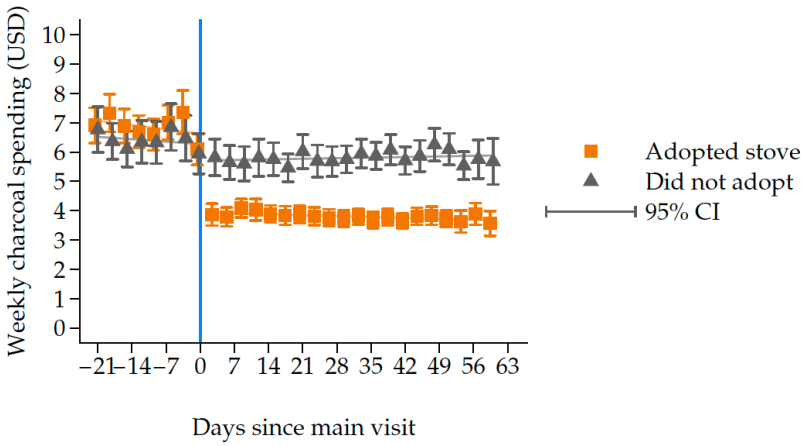


FIGURE 4. ENERGY EFFICIENT STOVES REDUCE ENERGY SPENDING

Note: Weekly charcoal spending by adopters and non-adopters of the energy efficient stove before and after the main visit (visit 2). Charcoal spending is elicited through a recurring 3-day SMS survey. Adoption of the stove causes charcoal expenditures to drop by USD 2.28 per week (39 percent relative to the control group). The causal estimates presented in Figure C4 are similar.

```
\begin{figure}[h]

\caption{Energy efficient stoves reduce
energy spending}

\begin{flushright}
\begin{minipage}{0.9\textwidth}
\label{fig:diffindiff}

\includegraphics[width=1\textwidth]{"Images/
DiffInDiffCharcoal_USD_PAPER".pdf}
\end{minipage}
\end{flushright}

\justify \footnotesize Weekly charcoal
spending by adopters and non-adopters of the
energy efficient stove before and after the
main visit (visit 2). Charcoal spending is
elicited through a recurring 3-day SMS
survey. Adoption of the stove causes
charcoal expenditures to drop by USD 2.27
per week (40 percent relative to the control
group). The causal estimates presented in
\autoref{fig:diffindiffIV} are similar.

\end{figure}
```

Wharton
UNIVERSITY of PENNSYLVANIA

Calling tables you made (ask your PI / other pre-docs for code)

Added in Latex

TABLE 2—CAUSAL IMPACT OF STOVE ADOPTION ON WEEKLY CHARCOAL SPENDING

	OLS	First Stage	IV Estimate (1-month endline)			IV Estimate (1-year endline)	
	(1) USD	(2) Bought Stove	(3) USD	(4) IHS(USD)	(5) IHS(KG)	(6) USD	(7) IHS(USD)
BDM Price (USD)	0.01	-0.03					
	(0.01)	(0.00)					
WTP (USD)	-0.01	0.02	-0.00	-0.00	0.00	0.01	0.00
	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.02)	(0.00)
Bought Cookstove (=1)	-1.89		-2.28	-0.50	-0.48	-2.50	-0.56
	(0.28)		(0.29)	(0.07)	(0.08)	(0.42)	(0.09)
Observations	7853	913	7853	7853	796	6979	6979
Control Mean	5.72		4.97	2.16	1.55	5.30	2.21
Socioeconomic controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Data Source	SMSes	Midline	SMSes	SMSes	Buckets	SMSes	SMSes

What you made in Stata/R

Note: Results from an instrumental variables regression that uses the (randomly assigned) BDM price as an instrument for stove adoption to estimate the causal impact of adoption on weekly charcoal expenditures. Columns (1) and (2) present the OLS and first-stage estimates, respectively. Column (3) uses weekly charcoal expenditures in USD as the outcome variable. Column (4) uses the inverse hyperbolic sine (IHS) conversion of the USD amount. A 0.50 IHS reduction corresponds to a 39 percent reduction relative to the control group. Column (5) uses the IHS of the weight of the charcoal bucket one month after stove adoption as the outcome variable. Columns (6) and (7) conduct the same analyses as columns (3) and (4) respectively, but using data from the SMS survey conducted one year after the main visit. Socioeconomic controls include baseline savings, income, risk aversion, credit constrainedness, number of adults and children. In regressions using SMS data, errors are clustered by respondent. SE in parentheses.

Added in Latex

Berkouwer & Dean (AER, 2022)

Wharton
University of Pennsylvania

Calling tables you made (ask your PI / other pre-docs for code)

Table 2: Causal impact of stove adoption on weekly charcoal spending

	OLS	First Stage	IV Estimate (1-month endline)			IV Estimate (1-year endline)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	USD	Bought Stove	USD	IHS(USD)	IHS(KG)	USD	IHS(USD)
BDM Price (USD)	0.003	-0.029***					
	(0.013)	(0.001)					
WTP (USD)	-0.003	0.024***	-0.003	-0.001	0.003	0.008	0.004
	(0.010)	(0.001)	(0.011)	(0.003)	(0.003)	(0.016)	(0.003)
Bought Cookstove (=1)	-1.939***		-2.273***	-0.493***	-0.479***	-2.423***	-0.539***
	(0.294)		(0.295)	(0.071)	(0.083)	(0.427)	(0.091)
Observations	7870	913	7870	7870	796	6982	6982
Control Mean	5.717		4.964	2.154	1.545	5.303	2.212
Socioeconomic controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Data Source	SMSes	Midline	SMSes	SMSes	Buckets	SMSes	SMSes

Results from an instrumental variables regression that uses the (randomly assigned) BDM price as an instrument for stove adoption to estimate the causal impact of adoption on weekly charcoal expenditures. Columns (1) and (2) present the OLS and first-stage estimates, respectively. Column (3) uses weekly charcoal expenditures in USD as the outcome variable. Column (4) uses the inverse hyperbolic sine (IHS) conversion of the USD amount. A 0.498 IHS reduction corresponds to a 40 percent reduction relative to the control group. Column (5) uses the IHS of the weight of the charcoal bucket one month after stove adoption as the outcome variable. Columns (6) and (7) conduct the same analyses as columns (3) and (4) respectively, but using data from the SMS survey conducted one year after the main visit. Socioeconomic controls include baseline savings, income, risk aversion, credit constrainedness, number of adults and children. In regressions using SMS data, errors are clustered by respondent. SE in parentheses. * $\leq$ 0.10,** $\leq$ .05,*** $\leq$ .01.

```
\begin{table}[h]

\begin{center}
\caption{Causal impact of stove adoption on weekly charcoal
spending} \label{tab:ivresults}
\addtolength{\tabcolsep}{-5pt}

\begin{tabular}{@{\extracolsep{6pt}}lccccccccc}
\toprule
\input{Tables/IVcharcoal_PAPER.tex}
\bottomrule
\end{tabular} \\
\end{center}

\footnotesize Results from an instrumental variables regression
that uses the (randomly assigned) BDM price as an instrument for
stove adoption to estimate the causal impact of adoption on
weekly charcoal expenditures. Columns (1) and (2) present the
OLS and first-stage estimates, respectively. Column (3) uses
weekly charcoal expenditures in USD as the outcome variable.
Column (4) uses the inverse hyperbolic sine (IHS) conversion of
the USD amount. A 0.498 IHS reduction corresponds to a 40
percent reduction relative to the control group. Column (5) uses
the IHS of the weight of the charcoal bucket one month after
stove adoption as the outcome variable. Columns (6) and (7)
conduct the same analyses as columns (3) and (4) respectively,
but using data from the SMS survey conducted one year after the
main visit. Socioeconomic controls include baseline savings,
income, risk aversion, credit constrainedness, number of adults
and children. In regressions using SMS data, errors are
clustered by respondent. SE in parentheses. ${^*}\leq0.10,
^{**}\leq.05, ^{***}\leq.01$.

\end{table}
```

Wharton
University of Pennsylvania

Today

1. Graphing
2. Tables
3. Latex
4. **Organizing your research output**

Organizing your research output

Challenge:

1. Many datasets with many variables and many ways of cleaning them
2. Many draft versions that may not make it into the paper
3. Many regressions
4. Many analyses
5. Many revisions of old analyses

Solution:

1. Label and describe everything
2. Date everything

e.g. a continuously updated PDF with versions of results

The research process is revise, revise, revise

1. Is each regression OLS/IV/…?
2. What is the outcome variable?
3. Which sample did you use?
4. How did you construct variables?
5. What control variables did you include?
6. What fixed effects and clustering are included?

Table 15: Stages Progression

	Transformers		LMCP Sites		Completed Sites (P1)		Completed Sites (P2)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Ward voted pro-gov in 2013	38.24***	19.91	19.48***	15.67***	1.87	3.35	6.94***	6.80**
	(9.30)	(13.23)	(2.81)	(3.87)	(1.59)	(3.31)	(1.90)	(3.29)
Aligned with MP in 2013	0.12	-14.27	0.29	-7.52*	1.05	1.18	0.36	0.12
	(7.73)	(13.60)	(2.46)	(4.24)	(1.18)	(2.59)	(1.49)	(2.70)
Observations	776	243	776	243	608	174	598	165
Opposition Mean	148.76	149.35	40.93	39.62	7.83	8.29	15.35	15.16
Treatment Effect (%)	25.71	13.33	47.60	39.55	23.87	40.36	45.22	44.84
Adjacent wards only?	No	Yes	No	Yes	No	Yes	No	Yes

Note: All regressions are at the ward level. Regressands are given in per 100,000 ward population terms. Pro-gov't voting is a dummy for it Kenyatta's voteshare in the relevant presidential election was over 50%. Phase 1 completed sites come from the phase 1 progress dataset (a subset of the total number of phase 1 transformers). Phase 2 completed sites come from the AFDB phase 2 dataset (which comprises the entirety of AFDB2 transformers as of late 2022). Adjacent wards are defined as wards which border at least 1 other ward which voted for a different party in the 2013 election. Control variables include land area of the ward, proportion of households with corrugated iron roofs, proportion of household heads with primary/secondary education, ward-specific Gini coefficient, number of laborers who work for pay, total dependency ratio, average household size, and average land gradient.

The research process is revise, revise, revise

1. Is each regression OLS/IV/…?
2. What is the outcome variable?
3. Which sample did you use?
4. How did you construct variables?
5. What control variables did you include?
6. What fixed effects and clustering are included?

Table 16: Stages Decomposition

	Transformers		LMCP Sites		Completed Sites (P1)		Completed Sites (P2)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Ward voted pro-gov in 2013	38.24***	19.91	0.08	0.16***	-0.09**	0.02	-0.01	0.04
	(9.30)	(13.23)	(0.05)	(0.04)	(0.04)	(0.06)	(0.03)	(0.05)
Aligned with MP in 2013	0.12	-14.27	0.04	0.02	-0.02	-0.04	0.05*	0.07
	(7.73)	(13.60)	(0.03)	(0.03)	(0.03)	(0.06)	(0.03)	(0.05)
Observations	776	243	776	243	608	174	598	165
Opposition Mean	148.76	149.35	0.33	0.32	0.42	0.38	0.77	0.79
Treatment Effect (%)	25.71	13.33	22.91	48.98	-20.98	5.72	-1.00	4.83
Adjacent wards only?	No	Yes	No	Yes	No	Yes	No	Yes

Note: All regressions are at the ward level. Regressands are given divided by the previous stage - from left to right, these are: total transformers per 100,000 population, LMCP-selected transformers per transformer, completed LMCP transformer per selected LMCP transformer (in phase 1), andcompleted LMCP transformer per selected LMCP transformer (in phase 2). Pro-gov't voting is a dummy for it Kenyatta's voteshare in the relevant presidential election was over 50%. Phase 1 completed sites come from the phase 1 progress dataset (a subset of the total number of phase 1 transformers). Phase 2 completed sites come from the AFDB phase 2 dataset (which comprises the entirety of AFDB2 transformers as of late 2022). Adjacent wards are defined as wards which border at least 1 other ward which voted for a different party in the 2013 election. Control variables include land area of the ward, proportion of households with corrugated iron roofs, proportion of household heads with primary/secondary education, ward-specific Gini coefficient, number of laborers who work for pay, total dependency ratio, average household size, and average land gradient.

The research process is revise, revise, revise

1. Is each regression OLS/IV/…?
2. What is the outcome variable?
3. Which sample did you use?
4. How did you construct variables?
5. What control variables did you include?
6. What fixed effects and clustering are included?

Table 17: Stages Progression, Not Trimmed

	Transformers		LMCP Sites		Completed Sites (P1)		Completed Sites (P2)		LMCP Meters	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Ward voted pro-gov in 2013	38.24***	19.91	19.48***	15.67***	1.87	3.35			526.49	980.72**
	(9.30)	(13.23)	(2.81)	(3.87)	(1.59)	(3.31)			(385.89)	(417.84)
Ward voted pro-gov in 2017							6.99***	6.78**		
							(1.86)	(3.30)		
Aligned with MP in 2013	0.12	-14.27	0.29	-7.52*	1.05	1.18	0.48	0.09	220.25	-435.35
	(7.73)	(13.60)	(2.46)	(4.24)	(1.18)	(2.59)	(1.50)	(2.68)	(300.94)	(460.24)
Observations	776	243	776	243	608	174	599	165	768	238
Opposition Mean	148.76	149.35	40.93	39.62	7.83	8.29	15.20	14.83	148.76	149.35
Treatment Effect (%)	25.71	13.33	47.60	39.55	23.87	40.36	46.01	45.71	353.92	656.66
Adjacent wards only?	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

Note: All regressions are at the ward level. Regressands are given in per 100,000 ward population terms. Pro-gov't voting is a dummy for it Kenyatta's voteshare in the relevant presidential election was over 50%. Phase 1 completed sites come from the phase 1 progress dataset (a subset of the total number of phase 1 transformers). Phase 2 completed sites come from the AFDB phase 2 dataset (which comprises the entirety of AFDB2 transformers as of late 2022). Adjacent wards are defined as wards which border at least 1 other ward which voted for a different party in the 2013 election. Control variables include land area of the ward, proportion of households with corrugated iron roofs, proportion of household heads with primary/secondary education, ward-specific Gini coefficient, number of laborers who work for pay, total dependency ratio, average household size, and average land gradient.

The research process is revise, revise, revise

1. Is each regression OLS/IV/…?
2. What is the outcome variable?
3. Which sample did you use?
4. How did you construct variables?
5. What control variables did you include?
6. What fixed effects and clustering are included?

Table 18: Stages Decomposition, Not Trimmed

	Transformers		LMCP Sites		Completed Sites (P1)		Completed Sites (P2)		LMCP Meters	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Ward voted pro-gov in 2013	38.24***	19.91	0.08	0.16***	-0.09**	0.02			5.52	28.97***
	(9.30)	(13.23)	(0.05)	(0.04)	(0.04)	(0.06)			(10.70)	(8.98)
Ward voted pro-gov in 2017							-0.01	0.02		
							(0.03)	(0.06)		
Aligned with MP in 2013	0.12	-14.27	0.04	0.02	-0.02	-0.04	0.05*	0.07	6.90	-0.39
	(7.73)	(13.60)	(0.03)	(0.03)	(0.03)	(0.06)	(0.03)	(0.05)	(7.99)	(7.96)
Observations	776	243	776	243	608	174	599	165	768	238
Opposition Mean	148.76	149.35	0.33	0.32	0.42	0.38	0.78	0.80	59.11	44.73
Treatment Effect (%)	25.71	13.33	22.91	48.98	-20.98	5.72	-1.23	2.73	9.33	64.75
Adjacent wards only?	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

Note: All regressions are at the ward level. Regressands are given divided by the previous stage - from left to right, these are: total transformers per 100,000 population, LMCP-selected transformers per transformer, completed LMCP transformer per selected LMCP transformer (in phase 1), and completed LMCP transformer per selected LMCP transformer (in phase 2).

Wharton
University of Pennsylvania

A replication dataset has 5 key folders

DATA
- **RAW**
  - .xlsx, .csv, .txt, .doc, .pdf, etc. (direct download from website/research partner)
- **MEDIUM**
  - .dta, .r (imported, merged, cleaned of typos)
- **CLEAN**
  - .dta, .r, (variables used in analysis)

DO
- Master file with paths and settings (.do, .R, etc.)
- Cleaning file(s) (.do, .R, etc.)
- Analysis file(s) (.do, .R, etc.)

SURVEYING INSTRUMENTS
- Baseline survey (word/excel)
- Endline survey (word/excel)

PACKAGES
- .ado

RESULTS
- **GRAPHS**
  - .eps, .pdf, .png
- **TABLES**
  - .tex
- **MAPS**
  - .eps, .pdf, .png

- ReadMe.docx (sources of raw data files, explanation of files)
- RunAll.do, RunAll.R (replicates entire project)

Wharton
University of Pennsylvania


Wharton
University of Pennsylvania