Assessing External Validity in an Instrumental Variable Setting

Evaluation Review
XX(X):1–24
©The Author(s) 2023
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Abstract

We test external validity through instrumental variable estimation, using the case study about the impact of solid fuel usage on women's average cooking time. In the case study, we compare the local average treatment effect (LATE) of the country of interest with the predicted LATE estimated with data from other countries. While the sub-population is an important factor, it does not significantly undermine external validity in our case study. Among six countries examined, three (Ethiopia, Honduras, and Cambodia) exhibit no statistically significant difference between predicted and actual LATE across various specifications. These results give evidence that external validity is not severely harmed in our case study. Conversely, in Nepal and Zambia, the two LATEs are statistically different, indicating distinct sub-populations compared to the sub-populations of other countries. These findings provide evidence that sub-population is a non-trivial factor for external validity.

Keywords

external validity, instrumental variable, generalization, prediction, solid fuel impact, time usage, developing countries

Introduction

Suppose you are a policy-maker or a superintendent that needs to decide what program will be effective compared to its cost when implemented. One piece of evidence to look for might be empirical program evaluation done in other populations or other settings, which include other time periods, sites, and differences in detailed policy implementations. Then, it is questionable if a study conducted with data from a certain population and setting can be generalized to a different population or setting. The illustration above is an example of a broader issue commonly called external validity.

Randomized Controlled Trials (RCT), which are suggested as an ideal methodology for estimating the effect of a program (Leamer, 1983), might not result in high external validity. For instance, if the sites where the experiments are held are not selected randomly, then the RCT study might have low external validity. Allcott (2015) shows that predictions using the first 10 sites over-estimates the treatment effect in the later 101 sites. The program under evaluation by Allcott (2015) is the Opower program which is an RCT. The program tests if providing more information on electricity usage affects consumers to save energy. One explanation of why the first 10 sites over-estimate the later 101 sites is because those who participate in the first 10 sites are expected to be those who are more interested in the program and will actively participate. Thus, the estimated treatment effect might be larger for those first 10 sites than the later 101 sites. This is an example where RCT might fail to have high external validity.

Moreover, RCT might not be available for every program evaluation. It might be unethical or infeasible to implement a randomized experiment. For example, Imbens (2010) states that macroeconomic policies are hard to be subject to randomized experiments. Moreover, for developing countries, it might be hard to use RCT as empirical evidence for program evaluation due to its high costs. For example, a policy-maker might be interested in building a nuclear power plant. However, randomly selecting sites and building nuclear power plants in those sites will be highly costly that might make the experiment infeasible.

Researchers can choose to use quasi-experiments when RCT is not available but observational data is available, and want to have a better research design for causal inference (Angrist and Pischke, 2010). Among methods in quasi-experimental studies, instrumental variables are widely used. If the researcher has valid instruments, then the estimated effect is unbiased and consistent, resulting in high internal validity. However, as similar to the RCT case, high internal validity does not necessarily lead to high external validity.

There is previous research to improve the external validity. Tipton (2013) uses the propensity score method, the subclassification method in particular, as an attempt to generalize the results from experiments. Meanwhile, Tipton and Peck (2017) focus on the design-based approach and suggest a stratification method that does not require random sampling to better select samples to represent the target population. Stuart and Rhodes (2017) give a general review of the external validity issue and do a case study using a reweighting approach. Tipton (2014) and Tipton et al. (2017) present indexes that can be employed to evaluate the degree of similarity between an experimental sample and an inference population.

While some papers specifically focus on generalization, there are papers that place a greater emphasis on prediction. Generalization involves inferring the population's average treatment effect from the average treatment effect observed in experiments. On the other hand, prediction aims to estimate the average treatment effect of the target population based on the average treatment effect observed in other populations or settings. For instance, in addition to Allcott (2015) mentioned above, Hotz et al. (2005) somewhat accurately predicts the average outcomes of a training program of the control group in one location given outcomes in other locations. Hotz et al. (2005) predict the control group to distinguish between two complications that threaten external validity. First, the populations might differ in sites. Second, the detailed implementation of the program might differ between sites. The results among the control group suggest that population difference is mostly adjusted after considering characteristic differences. Meanwhile, Hotz et al. (2005) gives suggestive evidence that difference in detailed program implementation does not harm the prediction much after adjusting for individual characteristics, especially the previous employment status. The data used for evaluation is from four experimental evaluations of WIN demonstration programs, which are randomized control experiments implemented in some parts of the U.S. The prediction method incorporates a matching process.

Generalization and prediction have different perspectives but share similar research methods. Moreover, they share a common objective of estimating the average treatment effect for individuals not included in the sample, using information derived from the individuals who are part of the sample.

This paper studies the external validity issue within the context of instrumental variable (IV) estimation. We present a case study using large observational data from World Bank that covers multiple countries. We check if the estimated local average treatment effect from the IV estimation of one country can be closely predicted using data from other countries. Previous literature point out that IV estimates the treatment effect only locally. This is because IV estimates the effect of treatment for a sub-population that is induced to take the treatment due to the instruments such as policy changes or random assignments. This means that different instruments measure the average treatment effect among different sub-populations (Angrist and Pischke, 2009).[*] However, even if we use the same instrument, if the sub-populations are systematically different, the estimate of local average treatment effect (LATE)

[*] Chapter 4, section 4.4, p. 150-173

will be biased for out-of-sample prediction while the estimate is unbiased in the sample (Allcott, 2015). We suspect that the sub-population of different countries will be systematically different from each other. If this is the case, predicting a treatment effect with IV estimation using other countries' data will be biased.

The contribution of this paper is that it empirically investigates the external validity issue with IV estimation. While there are many papers that discuss external validity using examples from RCTs (Allcott, 2015; Hotz et al., 2005; Stuart and Rhodes, 2017), to the best of our knowledge, it is challenging to find literature on external validity that specifically employs IV estimation. Since RCT is not always available and IV estimation is widely used with observational data, checking the external validity within the IV estimation setting is valuable.

The main focus of this paper is prediction but shares close similarities with the work of Tipton et al. (2017) that focus on generalization. Following the terminology in Tipton et al. (2017), the target population is the observations from the country that is not used as a sample while the sample is the rest of the observations. For example, if we are interested in the LATE of Ethiopia, Ethiopia's population is the target population while data from other countries are the sample. Meanwhile, our methodology, mostly from Allcott (2015), extrapolates the estimated average treatment effect to the target population by estimating heterogeneous treatment effects. Therefore, this paper employs the term "prediction" as our primary objective in this case study following Allcott (2015). One note is that, in the main text, we do not use household weights since the results are not significantly different from using weights. The results using household weights are in the appendix.

The treatment effect that we use as a case study is the effect of using clean fuel on the average cooking hours of women inside the household. We use Multi-Tier Framework Survey (MTF) data which contains multiple variables related to energy access, cooking solutions, and health. The advantage of MTF data is that it covers 13 countries. 10 countries are used in this paper. We categorize the main fuel used in cooking into a binary variable called "solid fuel", which includes charcoal, animal waste, and wood.[†] Following previous literature[‡], if the household uses solid fuel, it is expected to increase the average cooking hours of women in that household on average. Solid fuels, such as wood, typically require more time to ignite compared to non-solid fuels like LPG, kerosene, and electricity. However, the impact of fuel type on cooking time can vary due to many differences across countries or regions, including geographical and cultural differences. Among the 10 countries, 6 countries show a result of positive and significant effect of solid fuel on the average cooking hours of women after IV estimations. We predict each of these 6 countries using data from 9 other countries.

[†] The specific list of fuels categorized as solid and non-solid can be found in the data section.
[‡] Afridi et al. (2023); Williams et al. (2020)

There are several reasons why the prediction may fail in our case study. Among many reasons, we point out three problems. First, each country may have a different distribution of household characteristics. For example, each country may have a different distribution of age, gender, and cooking fuel usage. The second reason is that there may be some unobservable factors that cause heterogeneity. For example, each country might have a different food culture that affects the usage of cooking fuel and affect cooking time. Lastly, if the sub-populations are systematically different between different countries, then the prediction with IV estimation might fail. We use different specifications in an attempt to control the first and second problems. The population difference is accounted for with the re-weighting approach and estimation of heterogeneous treatment effects. The unobservable factors are controlled with control variables and IV estimation. Then, we check if the third problem, which is our main interest, remains.

We find that while the sub-population is an important factor for external validity, it does not significantly undermine external validity in our case study. Overall, 3 out of 6 countries (Ethiopia, Honduras, and Cambodia) fail to reject the test that has a null hypothesis of no difference between the predicted LATE and the actual LATE of the country of interest. For these 3 countries, the predicted LATE and actual LATE are consistently similar under different specifications. This gives suggestive evidence that the sub-population of each of these 3 countries is not systematically different from the other 9 countries. 2 countries (Nepal and Zambia) consistently reject the null hypothesis of no difference. This result gives suggestive evidence that the sub-populations are different for these 2 countries from other countries. For Kenya, the result changes depending on the specification. The actual LATE and predicted LATE are mostly similar for Kenya but if we predict LATE using selected variables after lasso estimation, the test rejects the null hypothesis. In the appendix, we include aggregate variables to account for country differences in estimating heterogeneous treatment effects and use household weights to better represent the population of each country. If we include aggregate variables, the results are similar for other countries but the results of the test mostly change for Kenya and Zambia. If we use household weights, the predicted LATE and actual LATE are similar to each other for 4 countries, which are Ethiopia, Honduras, Kenya, and Zambia.

Research Method

In this section, we focus on the research method related to external validity. The IV estimation for internal validity is explained in more detail when we introduce our case study.

External Validity

To predict LATE in the target population with data from other countries, we incorporate assumptions from Allcott (2015). Let T_i represent a treatment status in which $T_i = 1$ if the individual receives treatment and $T_i = 0$ otherwise. Let Y_i represent the outcome variable and X_i the covariates. $Y_i(1)$ represent the outcome when treatment is received and $Y_i(0)$ the outcome when treatment is not received. $D_i = 0$ means that the data is in the target population while $D_i = 1$ otherwise.

First, for convenience, let's assume that we are using RCT for causal inference. Since RCT randomly assigns the treatment, the potential outcome given the treatment status $(Y_i(T_i))$ can be thought of as independent with the treatment (T_i). Then, with the assumption that the support of the probability of treated and not-treated coincide, internal validity can be obtained.

Assumption 1. *Unconfoundedness.* $T_i \perp \{Y_i(1), Y_i(0)\} | X_i$.

Assumption 2. *Overlap.* $0 < Pr(T_i = 1 | X_i = x) < 1$

However, further assumptions are needed to obtain external validity.

Assumption 3. *External Unconfoundedness.* $D_i \perp (Y_i(1) - Y_i(0)) | X_i$

Assumption 4. *External Overlap.* $0 < Pr(D_i = 1 | X_i = x) < 1$

Assumption 3 means that the treatment effect and target selection are independent conditional on covariates. This assumption is violated if there is a systematic relationship between the treatment effect and site selection. Assumption 4 means that given covariates, the probabilities to be selected in the target population and in the other populations are both non-zero.

The problem we focus on is where assumption 1 does not hold. This is because when we use observational data, the treatment and potential outcomes might be affected by confounding factors that are not controlled with X. Thus, we use an instrumental variable for estimation. Let Z_i represent a dummy instrument variable and T_1 be the treatment status when $Z_i = 1$, and T_0 be the treatment status when $Z_i = 0$.

For the IV strategy, we adopt 4 assumptions from Abadie (2003).

Assumption 5. *Independence.* $\{Y_i(Z_i, T_i)$ *for every* Z_i *and* $T_i, T_0, T_1\} \perp Z_i | X_i$.

Assumption 6. *Exclusion.* $Y_i(Z_i = 0, T_i) | X_i = Y_i(Z_i = 1, T_i) | X_i \quad \forall T_i = 0, 1$

Assumption 7. *Nonzero Average Causal Effect.* $E[T_i(Z_i = 1) - T_i(Z_i = 0) | X_i] \neq 0$

Assumption 8. *Monotonicity.* $T_i(1) \geq T_i(0)$

Assumption 5 means that our instrumental variable is "as good as randomly assigned" conditional on covariates. Assumption 6 means that the instrumental variable, Z, does not directly affect the outcome variable. Assumption 7 means that conditional on covariates, instrumental variable, Z_i, and treatment, T_i, are correlated.

When we use binary instrument variables, we can think of four disjoint groups: always-takers, never-takers, defiers, and compliers. Always-takers are those who always take the treatment regardless of the instrument status while never-takers are those who always do not take the treatment regardless of the instrument status. Compliers are individuals who would receive the treatment only if they were assigned to the treatment group. Assumption 8 excludes the defiers, which is a group of people who are treated when they are not instrumented or not treated when they are instrumented.

Abadie (2003) shows that with assumptions 5, 6, 7, 8, and linear first stage, the two-stage least square estimator is the same as the treatment effect of compliers conditional on covariates ($E\left[Y(1) - Y(0)|X, T_1 > T_0\right]$).

Then, we estimate

$$\begin{aligned} &E[\tau|D_i = 0, T_1 > T_0] = \\ &E\left[\left[Y_i(1) - Y_i(0)|D_i = 1, X_i, T_1 > T_0\right]|D_i = 0\right] \end{aligned} \tag{1}$$

Equation 1 means that LATE for the target population can be estimated with data from other countries by taking the expectation of the estimated heterogeneous treatment effects. When we do IV estimations, the LATE estimates treatment effect among different sub-populations. If the sub-populations are systematically different between the target population and other populations, assumption 3 is violated. Then, the LATE estimated with data from other populations will be biased for LATE in the target population.

Case Study

Our case study focuses on the effect of the usage of solid fuel as a primary fuel for cooking on women's cooking hours. Approximately one-third of the population still relies on non-clean fuels for cooking energy in many developing countries, particularly in rural areas. The United Nations has established Sustainable Development Goal 7 (SDG7), which aims to ensure universal access to affordable, reliable, and modern energy services. Cooking energy is one of the primary targets of this policy (SDG, 2021). Therefore, it is crucial to evaluate the impact of solid fuel on cooking. The evaluation can be used as evidence for policy implementations.

Afridi et al. (2023) state that cooking time is the most time-intensive activity in home production which is disproportionally distributed to women. Moreover, in developing countries, many households continue to rely on solid fuels, such as wood, charcoal, and animal dung, along with traditional cookstoves for cooking (Krishnapriya et al., 2021).

However, there is limited evidence that shows the time-saving effect of the usage of non-solid fuels among women in developing countries. The time saved among women is expected to result in welfare gains, in addition to the health gains of using clean fuel (Verma and Imelda, 2022). For example, the saved time can be used as a time for education or work, which can increase women's empowerment.

Rather than solid fuel, there is previous research on the effect of improved cookstoves (ICS). These stoves incorporate advanced technology to make cooking more efficient, cleaner, and safer, despite still relying on solid fuels. The primary focus of the previous research is to discuss whether ICS decreases cooking time in comparison to traditional stoves. For instance, Krishnapriya et al. (2021) examine the effect of using an ICS on women's time allocation in six countries. The findings indicate that ICS reduces cooking time by an average of 34 minutes per day. However, the estimates vary across the six countries.[§] Instead of focusing on ICS, we investigate the impact of using solid fuels for cooking time in comparison to non-solid fuels.

The effect of solid fuel on cooking time among women is not clear. Afridi et al. (2023) use an RCT that nudged the treated into the usage of LPG (liquefied petroleum gas). They find a moderate time-saving effect of the usage of clean fuel on cooking time in the rural areas of the Indore district in India. Meanwhile, Williams et al. (2020) find a larger time-saving effect of LPG usage by using an RCT in Peru. Also, Su and Azam (2023) show that LPG usage decreases the cooking time using nationally representative India data. They also discovered evidence of a rebound effect, suggesting that the use of LPG for cooking may lead to a slight increase in the number of cooking events per day.

Afridi et al. (2023), Su and Azam (2023), and Williams et al. (2020) explore the relationship between fuel type and cooking time but with a specific emphasis on the adoption of LPG, which is one type of clean fuel. In contrast, our study considers solid fuel as the main variable of interest, encompassing information on LPG usage as well. Consequently, our study uses a broader definition than the aforementioned papers by focusing on solid fuels.

Although they share several common fuel types, we should be aware that solid fuel and non-clean fuel are distinct in their definitions. For example, kerosene is non-solid fuel but is categorized as non-clean fuel. Thus, the effect of non-clean fuel does not imply the effect of solid fuel. Further research is needed to see if the two different definitions lead to large differences in the results.

In this case study, we check if the estimated local average treatment effect from the IV estimation of one country can be closely predicted using data from other countries. The treatment effect of interest is whether the usage of clean fuel reduces the average cooking hours of women inside the household, and how much the usage of clean fuel reduces the average cooking hours of women. We categorize the main fuel used in cooking into a binary variable called "solid fuel" by following Kurata et al. (2020).[¶] If the household uses solid fuel, it is expected to increase the average cooking hours of women in that household on average.

[§] Krishnapriya et al. (2021) also provide a comprehensive summary of previous research on the time-saving effects of ICS intervention.

[¶] We adopt the standard proposed by Kurata et al. (2020) to generate a dummy variable representing the use of solid fuels.

We want to empirically check if it is hard to predict LATE using other countries' estimates of LATE, which is caused by the fact that the compliers are different between different countries. This might be a result of multiple factors including historical, cultural, geographical, and weather differences between countries.

Internal Validity: IV estimation

Even if our main argument is external validity, it is important to obtain internal validity. We estimate the following equation:

$$y_i = \alpha_0 T_i + \beta X_i + \epsilon_i. \tag{2}$$

y_i is the outcome variable, the average cooking hours of women in household i. T_i is a dummy variable where $T_i = 1$ if the main source of cooking fuel is solid fuel and $T_i = 0$ otherwise. ϵ_i is the error term. X_i is the vector of control variables that includes the number of children (age under 5), the total number of household members, the number of females in the household, stove type (whether clean fuel stove), household structure, and wealth index. Also, we include the aggregate cooking solution tier variable provided by MTF data.

The OLS estimation may be biased because T_i can be endogenous, meaning $E[\epsilon_i|T_i, X_i] \neq 0$. To be specific, if the error term is as good as randomly distributed once conditional on fuel choice and covariates, OLS is an unbiased estimator. However, there may be some unobserved factors that could affect both the average cook time and solid fuel usage, making the solid fuel dummy variable endogenous. To solve this potential endogeneity issue, an instrument variable (IV) strategy is used. This involves using an instrumental variable that is correlated with solid fuel usage but affects average cook time only through the usage of solid fuel.

We adopt an instrument used by Kurata et al. (2020) which is a proxy for the availability of solid fuel in the neighborhood. The conjecture is that if other households in the neighborhood are mostly using solid fuel, it gives some suggestive evidence that solid fuel is more available in that neighborhood. Let *Neighborhoods* be the collection of neighborhoods and $J \in Neighborhood$ be an element of the collection. Then the instrument variable is the average of the usage of solid fuel of all of my neighbors in J:

$$Z_{is} = \frac{\sum_{j \neq i}^{|J|} T_j}{|J| - 1} \; for \; each \; i \in J. \tag{3}$$

To simplify the interpretation of the regression results, we create a new dummy instrument variable that is 1 if the neighborhood has higher availability than the country average ($\bar{Z}_s$).

$$Z_{is}^{dummy} = \begin{cases} 1 & \text{if } Z_{is} > \bar{Z}_s \\ 0 & \text{otherwise} \end{cases}$$

Data

We use Multi-Tier Framework Survey (MTF)[||] data to discuss the external validity issue of the impact of solid fuel on the average cooking time of women in households. We use MTF because it includes detailed information on cooking fuel usage in each household and time use related to cooking within households across several countries in Sub-Saharan Africa and Asia.

The MTF was created by Energy Sector Management Assistance Program (ESMAP) after the adoption of Sustainable Development Goals (SDGs). The survey provides a global standard to evaluate energy access and cooking solutions in a multi-dimensional approach, such as adequacy, availability, reliability, convenience, affordability, and safety, with tiers ranging from 0 to 5 rather than binary measures, indicating whether a household has electricity or not. (Bhatia and Angelou, 2015)

The main advantage of MTF data is that the data covers 13 countries from Africa and Asia which include Ethiopia, Honduras, Kenya, Cambodia, Liberia, Myanmar, Niger, Nigeria, Nepal, Rwanda, Sao Tome and Principe, Uganda, and Zambia. We use 10 countries, excluding three countries, Uganda, Nigeria, and Myanmar because these countries do not contain information about the cooking time of women which is our outcome variable.

The MTF data has rich information about the fuel type of each stove in the household. Consequently, we construct a dummy variable that indicates whether the main stove's fuel type is solid fuel or not. According to the definition of solid fuel from previous literature, animal waste, biomass, charcoal, coal, plant biomass, garbage, peat, pellets, sawdust, and wood are categorized as solid fuel while bio-gas, electricity, kerosene, LPG, natural gas, and solar are categorized as non-solid fuel. (Kurata et al., 2020)

While the MTF data contains information about time usage, there are certain limitations. The survey for MTF data has several questions for time allocation such as cooking time, cooking fuel collection time, and working time, etc. However, they don't have individual time usage data. Time usage is aggregated for the type of household members, men, women, boys, girls, and children (age under 5). Since we can observe how many women are in the household, we construct a variable that contains the average cooking time of women within each household.

Since the dependent variable is at the household level, the control variables are at the household level. To control the wealth level of the household, we construct a wealth index variable using Principal Component Analysis following Krishnapriya et al. (2021). Also, We construct a variable called ”household structure” to control the cooking environment and to have an alternative proxy of wealth. MTF has some information about the cooking environment such as the existence of ventilation and the location of the stove. However, those variables

[||] The datasets are available at: `https://datacatalog.worldbank.org/`

have a lot of missing values. Therefore, we construct a variable that proxies both cooking environment and houshold wealth by assuming that wealthier households have better house structure and better cooking environment. The house structure is a combination of two variables from the MTF data, which are wall and roof material. The usage of household material as a proxy for household wealth is not new (Bergeron et al., 2021). The detailed construction of the variable is in the code provided.

Predicting LATE

Before predicting LATE, we need an estimation method for heterogeneous treatment effects estimated with data from other countries. Let y_{is} denote the dependent variable which is the household, i, average time per woman used for cooking on a typical day in the country s. T_i is the treatment of interest. Let $\bar{X}_{D=1}$ represent the vector of the mean of in-sample covariates. $\tilde{X}_{is} = X_{is} - \bar{X}_{D=1}$ denotes the vector of demeaned covariates. The equation used for estimation is:

$$y_{is} = (\alpha \tilde{X}_i + \alpha_0)T_i + \sum_s \left(\beta_s \tilde{X}_i + \pi_s\right) + \epsilon_{is} \tag{4}$$

T_i might be an endogenous variable that is correlated with ϵ_{is}. We do IV estimation with instrument Z_{is}^{dummy} to solve this problem.

We predict LATE of a specific country using data from other countries. To extrapolate to the specific country, we assume a linear treatment effect.

Assumption 9. $E\left[\tau_i | X_i = x\right] = \alpha x + \alpha_0$

We denote the LATE of the country that will be predicted as $\tau_{D=0}$ and the LATE estimated with data from other countries as $\tau_{D=1}$. $\bar{X}_{D=0}$ represents the vector of mean covariates of the specific country we are targeting to predict. With the assumptions, we predict the average treatment effect as:

$$\hat{\tau}_{D=0} = \hat{\tau}_{D=1} + \hat{\alpha}(\bar{X}_{D=0} - \bar{X}_{D=1}). \tag{5}$$

Note that $\hat{\tau}_{D=1} = \hat{\alpha}_0$, and $\hat{\alpha}$ is from equation 4.

We compare the predicted LATE to the LATE obtained using data from the target country. For distinction, let us call the LATE obtained using data from the target country as actual LATE. The equation used for estimation is:

$$y_i = \alpha_0 T_i + \beta X_i + \epsilon_i. \tag{6}$$

The estimation method is IV estimation with the same instrument variable as before. Let $\hat{\tau}_{D=0}^{actual} = \alpha_0$ be the estimated actual LATE from equation 6.

Test

We perform a simple test to check if the predicted LATE is different from the actual LATE. The test is based on large sample asymptotic properties. Let

$\hat{\Omega} = \hat{\tau}_{D=0} - \hat{\tau}^{actual}_{D=0}$ be the difference in the predicted LATE and actual LATE. The null and alternative hypothesizes are:

$$H_0 : \hat{\Omega} = 0$$
$$H_1 : \hat{\Omega} \neq 0$$

The test statistic is:

$$t = \frac{\hat{\Omega}}{\sqrt{\hat{Var}(\hat{\Omega})}}. \tag{7}$$

where $\hat{Var}(\hat{\Omega}) = \hat{Var}(\hat{\tau}^{actual}_{D=0}) + \hat{Var}(\hat{\tau}_{D=0})$. The covariance between the two estimators are 0 because we assume that the distributions of covariates are independent between countries.

Results

OLS and IV regressions

Table 2 and 3 summarizes the OLS and IV estimation results. The standard errors are calculated as robust-standard errors. The coefficient of interest is the coefficient of "solid fuel". The IV results show that, except for São Tomé and Príncipe, the effect of using solid fuel increases the average cooking hours of women in the household.

When we compare the IV estimate and OLS estimate of the effect of solid fuel, the IV estimate of the coefficient of interest is larger than the OLS estimate in absolute terms for all countries. The effect of solid fuel on the average cooking hours of women in the household becomes positively significant in Kenya, Cambodia, Nepal, and Zambia when we use IV compared to OLS. When we use OLS, there were negatively significant effects of solid fuel on the average cooking hours of women in the household in Liberia and Niger but they become insignificant once we use IV. However, in São Tomé and Príncipe, the effect of solid fuel on the average cooking hours of women is negatively significant when we use IV estimation. This means that those households that use solid fuel spend less cooking time per household member than those that do not use solid fuel on average in São Tomé and Príncipe.

When we do prediction in the next section, we only make predictions for the countries that have a positive significant coefficient in Table 3.

Prediction

Table 1 summarizes the results. The rows with IPW in Table 1 represent the prediction with inverse probability weights. We estimate $D_i = 0$, which means that the observation is from the target country, with logistic regression using the covariates. The predicted value, $\hat{p}_{D=0}$, is the probability of the observation to be drawn from the target country given the covariates. Then, we generate weights as $\hat{w} = (1 - D) + D \times \frac{\hat{p}_{D=0}}{1-\hat{p}_{D=0}}$. When we do prediction, all the observations we use for estimation have $D = 1$. Thus, the weights are $\hat{w} = \frac{\hat{p}_{D=0}}{1-\hat{p}_{D=0}}$. The intuitive meaning of the process is to give higher weights to observations that are similar to the target population. The rows with ”post lasso” in Table 1 represent the prediction after performing the lasso estimation. Lasso estimation is performed to select important covariates for the prediction.

The main findings are as follows. Out of 6 countries, 3 countries, which are Ethiopia, Honduras, and Cambodia, consistently fail to reject the null hypothesis of difference in predicted LATE and actual LATE at the 10% significance level under different specifications. Although there is a network of assumptions imposed in the estimation, the failure of rejection of the null hypothesis gives suggestive evidence that the compliers are not systematically different in the target country and the 9 other countries. However, we need to be careful with the claim. Looking at the 4th row in Table 1, the differences in the predicted LATE and actual LATE for Ethiopia, Honduras, and Cambodia are over 40%. This suggests that the failure of rejection of the null hypothesis might be from imprecise estimation resulting from IV estimation method.

Nepal and Zambia consistently reject the null hypothesis at the 5% significance level, meaning that there is statistical evidence that predicted LATE and actual LATE are different. The results give us suggestive evidence that the compliers might be systematically different for those countries. The results of Kenya change depending on the specifications of the test. The actual LATE and predicted LATE are mostly similar for Kenya. However, if we predict after selecting variables with the lasso, the null hypothesis is not rejected at the 1% significance level.

The results are similar with or without IPW adjustment. The t-statistics are slightly lower in absolute values for Nepal and Zambia but the null hypothesis is rejected at the 10% significance level. The Ethiopia prediction is further away from the actual LATE if we use IPW and have large standard errors. This means that the prediction is highly imprecise for Ethiopia if we use IPW. One caution is that we could not trim the dataset to have only $0.1 \leq \hat{p} \leq 0.9$ for better overlap as in Allcott (2015) and Imbens et al. (2009) There were rank problems with the covariates. Moreover, the results do not change much if we predict after selecting variables with the lasso. This is because lasso mostly excludes 1 or 2 variables out of 11 covariates for each country.[**]

Overall, the predicted LATE is not significantly different from the actual LATE for Ethiopia, Honduras, and Cambodia while is significantly different for Nepal and Zambia. This gives some evidence that, in our case study, external validity is not seriously threatened when we incorporate IV estimation. However, 2 out of 6 countries have statistically different predicted and actual LATE. These results give suggestive evidence that the compliers in Nepal and Zambia are

[**] In the appendix, we incorporate aggregate variables as covariates to account for differences in countries when we predict heterogeneous treatment effects.

systematically different from other countries, indicating that systematically different sub-population is still a non-trivial factor for external validity.

Table 1. Prediction table: The effect of solid fuel on cooking time

Dependent: Average cooking time						
Country	Ethiopia	Honduras	Kenya	Cambodia	Nepal	Zambia
Actual LATE	1.573	2.005	1.365	0.637	11.321	3.115
	(0.447)	(0.635)	(0.513)	(0.155)	(4.433)	(0.700)
Predicted LATE	0.902	1.152	1.257	0.898	1.515	1.295
	(0.234)	(0.205)	(0.205)	(0.264)	(0.208)	(0.197)
Predicted with IPW	0.353	1.090	1.093	0.801	2.624	1.553
	(3.366)	(0.249)	(0.204)	(0.608)	(0.460)	(0.304)
t-statistics						
Country	Ethiopia	Honduras	Kenya	Cambodia	Nepal	Zambia
$\frac{predicted-actual}{actual} \times 100$	-43%	-43%	-8%	41%	-87%	-58%
t-statistics	-1.328	-1.278	-0.197	0.851	-2.210	-2.504
t-statistics with IPW	-0.359	-1.341	-0.494	0.260	-1.951	-2.049
Post lasso						
Actual LATE: post lasso	1.498	1.937	0.598	0.623	10.810	3.115
	(0.403)	(0.611)	(0.203)	(0.150)	(4.153)	(0.700)
Predicted: post lasso	1.427	1.003	1.232	0.879	1.413	1.295
	(0.204)	(0.207)	(0.170)	(1.009)	(0.176)	(0.197)
Predicted: lasso, IPW	0.248	0.958	1.040	0.705	2.380	1.553
	(4.857)	(0.239)	(0.175)	(1.056)	(0.403)	(0.304)
$\frac{predicted-actual}{actual} \times 100$: post lasso	-5%	-48%	106%	41%	-87%	-58%
t-statistics: post lasso	-0.158	-1.448	2.396	0.251	-2.261	-2.504
t-statistics: lasso, IPW	-0.256	-1.492	1.653	0.077	-2.021	-2.049

Conclusion

Using the case study of the impact of solid fuel on women's average cooking time in households, we test the external validity in the IV context. There are two main findings from our case study. First, we find evidence that the sub-population that responds to the IV is an important factor for external validity issues. To be specific, if the sub-populations are systematically different between different populations, then predicting LATE for the target population with data from other populations might be biased. Second, we find that the external validity issue is not severe in our case study. Out of 6 countries, the results of Ethiopia, Honduras, and Cambodia consistently show that the predicted LATE and the actual LATE are not significantly different under different specifications. On the other hand, the predicted LATE and the actual LATE are statistically different for Nepal and Zambia under different specifications. This highlights the presence of potential external validity issues when using IV estimation, even though 3 out of 6 countries show similar estimates for the predicted and actual LATE. It is important to acknowledge that the limitations of this research lie in the challenge of precisely identifying the specific sub-population that responds to the instrumental variable. Further investigations into the external validity within the IV framework, as well as other quasi-experimental settings, are warranted to deepen our understanding in this area.

OLS and IV estimation results

Table 2. OLS

	(1) Ethiopia	(2) Honduras	(3) Kenya	(4) Cambodia	(5) Leberia	(6) Niger	(7) Nepal	(8) Rwanda	(9) São Tomé and Príncipe	(10) Zambia
Solid Fuel	0.1931***	0.4959***	0.1486	0.0396	-0.1142**	-0.4312**	0.0066	-0.1516	-0.0140	0.1021
	(0.047)	(0.120)	(0.098)	(0.027)	(0.040)	(0.142)	(0.198)	(0.161)	(0.088)	(0.060)
CS aggregate tier	0.0453	0.0564	0.0102	-0.0540***	0.0443*	0.0744	-0.0027	-0.0973***	-0.0185	0.0751*
	(0.028)	(0.037)	(0.036)	(0.010)	(0.017)	(0.082)	(0.018)	(0.022)	(0.029)	(0.035)
HH average women age	0.0026	-0.0028	0.0103***	0.0003	-0.0040*	-0.0015	0.0052**	-0.0016	-0.0021	0.0027
	(0.003)	(0.003)	(0.003)	(0.001)	(0.002)	(0.007)	(0.002)	(0.002)	(0.004)	(0.002)
Sex of head	-0.0180	0.0967	-0.0618	-0.0416	0.0147	-0.4451**	-0.0727	0.0634	-0.0101	-0.0582
	(0.053)	(0.076)	(0.059)	(0.022)	(0.033)	(0.152)	(0.038)	(0.038)	(0.062)	(0.049)
HH max educ	0.0149	-0.1828**	-0.1074*	-0.0480*	0.0313	0.0000	0.0554	0.0772*	-0.0071	0.0415
	(0.049)	(0.068)	(0.051)	(0.022)	(0.032)	(.)	(0.043)	(0.032)	(0.072)	(0.047)
HH number of children	0.1008**	0.1091*	0.1109**	0.0123	0.0154	0.0352	0.0133	-0.0033	0.1014*	0.1021***
	(0.032)	(0.046)	(0.034)	(0.015)	(0.018)	(0.059)	(0.023)	(0.020)	(0.047)	(0.028)
HH number of people	-0.1431***	-0.2155***	-0.1553***	-0.0692***	-0.1337***	-0.1398***	-0.0962***	-0.0197*	-0.2128***	-0.0654***
	(0.012)	(0.018)	(0.013)	(0.006)	(0.006)	(0.015)	(0.008)	(0.009)	(0.020)	(0.010)
HH share of women	-1.7615***	-2.7447***	-2.3257***	-0.9553***	-2.0444***	-3.4266***	-0.8209***	-0.6160***	-2.4513***	-2.7436***
	(0.135)	(0.231)	(0.151)	(0.065)	(0.091)	(0.317)	(0.108)	(0.109)	(0.213)	(0.184)
Non-clean stove	-0.2249***	0.1303*	-0.0366	-0.0168	0.6690	0.7193***	0.0735	-0.1499***	0.1013	0.3425*
	(0.064)	(0.066)	(0.076)	(0.025)	(0.414)	(0.179)	(0.192)	(0.033)	(0.107)	(0.135)
Rural	0.3570***	0.1551*	0.2330***	-0.0024	-0.0544	-0.0475	-0.0359	-0.0274	0.0116	-0.0763
	(0.066)	(0.077)	(0.053)	(0.023)	(0.039)	(0.122)	(0.030)	(0.030)	(0.061)	(0.059)
House structure	0.0445	-0.0356	0.1437***	-0.0774**	0.0024	0.1399*	0.0043	0.0520	0.0042	-0.0462
	(0.041)	(0.072)	(0.037)	(0.024)	(0.025)	(0.055)	(0.026)	(0.032)	(0.074)	(0.032)
Wealth index	-0.0113	0.0044	-0.0030	-0.0049	-0.0016	-0.0027	-0.0046	-0.0045	-0.0090	0.0400**
	(0.013)	(0.018)	(0.011)	(0.005)	(0.012)	(0.010)	(0.008)	(0.007)	(0.011)	(0.012)
cons	2.1310***	3.2512***	2.2778***	1.6633***	2.0380***	3.6126***	1.2258***	1.8202***	3.4586***	1.6438***
	(0.160)	(0.271)	(0.217)	(0.080)	(0.423)	(0.398)	(0.132)	(0.199)	(0.249)	(0.216)
N	3392	2189	3417	3100	3092	717	5602	2964	1559	2808
Adjusted R-square	0.0944	0.1657	0.1096	0.1412	0.2546	0.2188	0.0362	0.0337	0.1219	0.1141
F	30.4395	37.2190	36.0427	43.4697	88.9843	19.2346	18.5433	9.6027	19.0185	31.1227

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3. IV

	(1) Ethiopia	(2) Honduras	(3) Kenya	(4) Cambodia	(5) Leberia	(6) Niger	(7) Nepal	(8) Rwanda	(9) São Tomé and Príncipe	(10) Zambia
Solid fuel	1.5727***	2.0054**	1.3654**	0.6370***	-0.2493	-58.2543	11.3215*	-13.6066	-2.1164***	3.1150***
	(0.447)	(0.635)	(0.513)	(0.155)	(0.144)	(176.424)	(4.433)	(10.088)	(0.602)	(0.700)
CS aggregate tier	0.1228**	0.2801**	0.3397*	-0.0226	0.0567**	-7.9981	0.2276**	-0.3212	-0.2641***	0.1109*
	(0.038)	(0.101)	(0.139)	(0.013)	(0.020)	(24.497)	(0.082)	(0.169)	(0.076)	(0.053)
HH average women age	0.0075*	-0.0034	0.0080**	-0.0002	-0.0045*	-0.0418	0.0057**	-0.0042	-0.0069	0.0063*
	(0.004)	(0.003)	(0.003)	(0.001)	(0.002)	(0.161)	(0.002)	(0.005)	(0.005)	(0.003)
Sex of head	-0.0522	0.0896	-0.0342	-0.0216	0.0156	0.1760	-0.1180**	0.0668	0.0245	0.0044
	(0.056)	(0.073)	(0.063)	(0.023)	(0.028)	(3.207)	(0.042)	(0.062)	(0.071)	(0.066)
HH max educ	-0.0502	-0.1758*	-0.1035	-0.0340	0.0393	0.0000	0.0532	0.0794	0.0178	-0.1204
	(0.062)	(0.070)	(0.056)	(0.025)	(0.035)	(.)	(0.056)	(0.046)	(0.085)	(0.074)
HH number of children	0.1066**	0.1222**	0.1138**	0.0130	0.0131	-0.8205	0.0331	-0.0453	0.1177*	0.1488***
	(0.038)	(0.047)	(0.037)	(0.016)	(0.018)	(2.636)	(0.030)	(0.047)	(0.053)	(0.041)
HH number of people	-0.1150***	-0.2306***	-0.1678***	-0.0765***	-0.1352***	-0.0088	-0.1113***	-0.0199	-0.2483***	-0.0496***
	(0.015)	(0.018)	(0.011)	(0.007)	(0.007)	(0.468)	(0.012)	(0.014)	(0.023)	(0.013)
HH share of women	-1.9572***	-2.7246***	-2.2034***	-0.9891***	-2.0396***	-10.2826	-0.8158***	-0.7572**	-2.3363***	-2.7541***
	(0.163)	(0.220)	(0.152)	(0.065)	(0.093)	(21.339)	(0.123)	(0.251)	(0.246)	(0.235)
Non-clean stove	-0.1053	-0.0652	0.2265	-0.0468	0.8144***	36.3137	-10.0027*	-0.2337**	-1.3069**	-2.2640***
	(0.074)	(0.105)	(0.131)	(0.029)	(0.221)	(109.133)	(4.053)	(0.085)	(0.428)	(0.640)
Rural	1.0757***	-0.1203	0.1312	0.0477	-0.1172	-14.6585	-0.1356**	-0.0293	-0.2659*	1.3027***
	(0.249)	(0.146)	(0.068)	(0.027)	(0.073)	(44.669)	(0.049)	(0.041)	(0.107)	(0.333)
House structure	0.1021*	0.1528	0.1919***	-0.1167***	0.0221	1.5732	-0.0383	0.0339	-0.0787	-0.5370***
	(0.050)	(0.102)	(0.045)	(0.032)	(0.034)	(4.556)	(0.035)	(0.058)	(0.093)	(0.126)
Wealth index	0.0222	0.0298	0.0034	-0.0034	0.0051	-0.0432	0.0337*	-0.1098	0.0131	-0.0326
	(0.017)	(0.020)	(0.010)	(0.005)	(0.014)	(0.158)	(0.015)	(0.085)	(0.015)	(0.023)
_cons	0.7177	1.4824	0.5916	1.5920***	1.9618***	36.9303	-0.0310	15.7617	6.1015***	2.2822***
	(0.488)	(0.759)	(0.719)	(0.096)	(0.200)	(100.301)	(0.466)	(10.422)	(0.787)	(0.359)
N	3392	2189	3417	3100	3092	717	5602	2964	1559	2793
F-statistic (1st)	12.4993	5.9222	6.2535	18.2313	0.9058	35.7677	11.7725	4.8859	16.6925	39.2671
P-value (1st)	0.0004	0.0150	0.0124	0.0000	0.3413	0.0000	0.0006	0.0272	0.0000	0.0000
F-statistic	44.3564	71.5192	82.1128	101.7880	125.6447	0.1061	14.2120	2.8095	40.9662	36.5367

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Derivation of $Cov(\hat{\tau}^{actual}_{D=0}, \hat{\tau}_{D=0}) = 0$

Let $\mathbb{1}$ be a $1 \times k$ vector of ones. Let X be the $n \times k$ covariate matrix and Z be the $n \times k_1$ matrix of instruments in which k_1 is the number of valid instruments. $Y = X\beta + \epsilon$ in which ϵ is the error term. The general expression for the 2SLS estimators is:

$$\beta_{2sls} = (X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z(Z'Z)^{-1}Z'Y) = \beta + (\hat{X}'\hat{X})^{-1}\hat{X}'\epsilon,$$

in which $\hat{X} = Z(Z'Z)^{-1}Z'X$. $\hat{\tau}^{actual}_{D=0}$ is one element of β.

The predicted ATE is:

$$\hat{\tau}_{D=0} = \hat{\tau}_{D=1} + \hat{\alpha}(\bar{X} - \bar{X}_{D=1}).$$

Note that $\hat{\tau}_{D=1}$ and $\hat{\alpha}$ are estimates using data from other countries rather than the target population. We assume that the covariates are independently distributed between countries. Then,

$$\begin{aligned}
& cov\left[(\hat{\tau}_{D=1} + \hat{\alpha}(\bar{X} - \bar{X}_{D=1}))\mathbb{1}, \beta + (\hat{X}'\hat{X})^{-1}\hat{X}'\epsilon\right] \\
&= cov\left[\hat{\alpha}\bar{X}\mathbb{1}, (\hat{X}'\hat{X})^{-1}\hat{X}'\epsilon\right] \\
&= E\left[(\hat{\alpha}\bar{X}\mathbb{1} - E\left[\hat{\alpha}\bar{X}\mathbb{1}\right])(\hat{X}'\hat{X})^{-1}\hat{X}'\epsilon)'\right] \\
&= E\left[E\left[(\hat{\alpha}\bar{X}\mathbb{1} - E\left[\hat{\alpha}\bar{X}\mathbb{1}\right])(\hat{X}'\hat{X})^{-1}\hat{X}'\epsilon)'|X, Z\right] = 0\right] \\
&= 0
\end{aligned}$$

Estimation method

$$\begin{aligned}
& E\left[\tau_i | D_i = 0, T_1 > T_0\right] = \\
& E\left[\tau_i | D_i = 1, T_1 > T_0\right] = \\
& E\left[\left[Y_i(1) - Y_i(0) | D_i = 1, X_i, T_1 > T_0\right] | D_i = 0\right]
\end{aligned}$$

The first equality follows from assumption 3. The second equality follows by definition. The two-stage-least square (2SLS) estimation method follows Abadie (2003) theorem 3.1 with an assumption of the linear first stage:

$$P(Z = 1 | X = x) = x'\pi.$$

According to Abadie (2003), with assumptions 5, 6, 7, 8, and the linear first stage, 2SLS estimator for treatment effect is the same as $E[Y(1) - Y(0)|X, T_1 > T_0]$. The theorem can be applied to both conditional on $D_i = 1$ and $D_i = 0$.

Prediction with aggregate variables

In this section, we include aggregate variables in the set of covariates following Hotz et al. (2005) in the attempt to adjust for the difference between countries.

Incorporating aggregate variables change equation (4) and the heterogeneous treatment effect accounts for country differences. Aggregate variables include the average GDP per capita for 5 years (from 2010 to 2015) and average precipitation for 5 years (from 2010 to 2015). GDP per capita is at the national level and average precipitation is at the regional level. The regions, which are called level 1, are divided by World Bank and we follow that division. In the estimation of actual LATE, the GDP per capita is excluded due to the perfect collinearity issue. But in the estimation of predicted LATE, the GDP per capita shifts $(\alpha\tilde{X}_i + \alpha_0)T_i$ in equation (4) which will change the prediction of LATE in equation (5).

A noticeable difference with Table 1 is that the tests reject the null hypothesis at the 1% significance level for Kenya consistently under different specifications. For Zambia, the tests fail to reject the null hypothesis at the 10% significance level consistently. For Nepal, if we implement IPW, the tests fail to reject the null hypothesis at the 10% significance level. But without IPW, the test rejects the null hypothesis at the 10% significance level. The standard errors of the estimators are higher when we include aggregate variables.

Table 4. Prediction table with aggregate variables: The effect of solid fuel on cooking time

Dependent: Average cooking time						
Country	Ethiopia	Honduras	Kenya	Cambodia	Nepal	Zambia
Actual LATE	1.580	1.219	1.404	0.633	11.321	2.345
	(0.448)	(0.831)	(0.512)	(0.157)	(4.433)	(0.645)
Predicted LATE	2.080	0.620	3.887	1.299	2.596	3.327
	(0.677)	(1.107)	(0.506)	(0.397)	(0.366)	(0.633)
Predicted with IPW	-5.061	0.939	4.167	0.408	4.404	5.312
	(7.371)	(1.042)	(0.620)	(0.827)	(1.051)	(2.419)
t-statistics						
Country	Ethiopia	Honduras	Kenya	Cambodia	Nepal	Zambia
$\frac{predicted-actual}{actual} \times 100$	32%	-49%	-177%	105%	-77%	42%
t-statistics	0.617	-0.433	3.448	1.562	-1.961	1.088
t-statistics with IPW	-0.899	-0.211	3.438	-0.267	-1.518	1.185
Post lasso						
Actual LATE: post lasso	1.498	1.175	0.613	0.615	10.810	2.345
	(0.401)	(0.782)	(0.203)	(0.151)	(4.153)	(0.645)
Predicted: post lasso	1.938	1.441	3.416	1.126	2.422	3.327
	(0.440)	(0.761)	(0.170)	(1.017)	(0.334)	(0.633)
Predicted: lasso, IPW	-5.261	0.212	3.838	0.115	4.267	5.312
	(8.280)	(0.983)	(0.400)	(1.131)	(1.000)	(2.419)
$\frac{predicted-actual}{actual} \times 100$	29%	23%	457%	83%	-78%	42%
t-statistics: post lasso	0.738	0.244	6.247	0.496	-2.013	1.088
t-statistics: lasso, IPW	-0.815	-0.767	5.777	-0.439	-1.532	1.185

Prediction using household weights

In this section, we use household weights to better represent the sample as the population of each country. If we use the household weights, 4 countries have a positive and significant effect of solid fuel on the average time of cooking among women in the household (table 5): Ethiopia, Honduras, Kenya, and Zambia. Table 6 shows the comparison of the predicted and actual LATE. The results show that the predicted LATE is close to the actual LATE. All countries fail to reject the null hypothesis of equality of the predicted and actual LATE and the difference between the two is smaller than 30%.

Table 5. IV estimation using household weights

	(1) Ethiopia	(2) Honduras	(3) Kenya	(4) Cambodia	(5) Leberia	(6) Niger	(7) Nepal	(8) Rwanda	(9) São Tomé and Príncipe	(10) Zambia
solid_fuel1	2.280***	2.894***	2.963**	0.221	-0.185	1.222	13.71	7.456	-1.352	1.703*
	(0.452)	(0.779)	(1.000)	(0.247)	(0.149)	(0.795)	(9.152)	(5.696)	(0.867)	(0.783)

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6. Prediction table with household weights: The effect of solid fuel on cooking time

Dependent: Average cooking time				
Country	Ethiopia	Honduras	Kenya	Zambia
Actual LATE	2.280	2.894	2.963	1.703
	(0.452)	(0.779)	(0.100)	(0.783)
Predicted LATE	1.830	2.139	2.756	2.110
	(0.657)	(0.485)	(0.648)	(0.441)
t-statistics				
Country	Ethiopia	Honduras	Kenya	Zambia
$\frac{predicted-actual}{actual} \times 100$	-20%	-26%	-7%	24%
t-statistics	-0.564	-0.823	-0.174	0.453

Figure 1. Average Women Cooking Hours

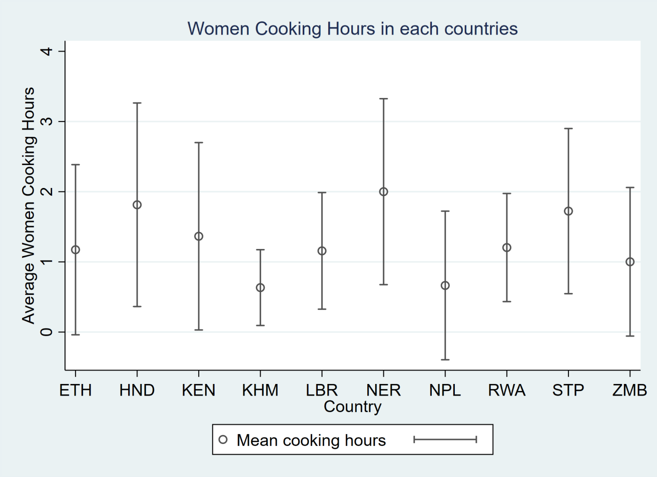

Figure 2. Distribution of solid fuel by country

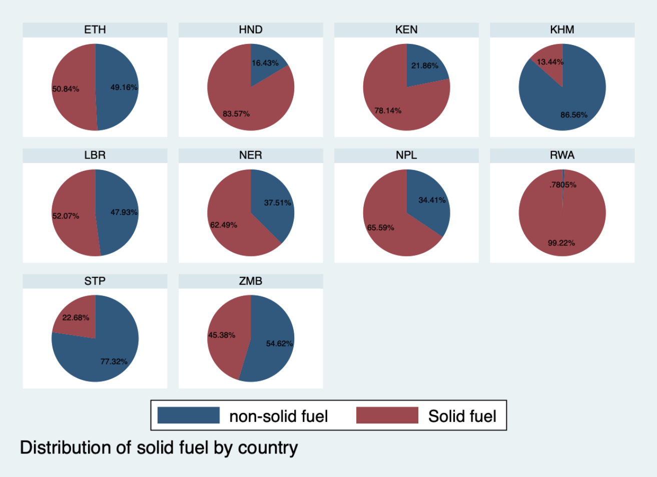

Table 7. Caption

	(1) ETH mean	 sd	(2) HND mean	 sd	(3) KEN mean	 sd	(4) KHM mean	 sd	(5) NPL mean	 sd	(6) ZMB mean	 sd
Animal Waste/Dung	0.021	0.144	0.000	0.000	0.000	0.000	0.000	0.000	0.021	0.142	0.000	0.000
Biogas	0.000	0.010	0.000	0.000	0.002	0.047	0.002	0.048	0.021	0.142	0.000	0.000
Biomass briquette	0.000	0.000	0.000	0.000	0.002	0.042	0.000	0.000	0.000	0.009	0.000	0.000
Charcoal	0.214	0.410	0.005	0.069	0.159	0.365	0.047	0.211	0.000	0.018	0.374	0.484
Coal/Lignite	0.000	0.008	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Crop Residue/Plant Biomass	0.014	0.119	0.000	0.020	0.001	0.039	0.002	0. 047	0.016	0.125	0.000	0.000
Electricity	0.039	0.194	0.001	0.036	0.001	0.037	0.018	0.133	0.00 0	0.009	0.149	0.357
Kerosene	0.000	0.000	0.000	0.000	0.069	0.254	0.000	0.000	0.000	0.000	0.000	0.000
LPG/Cooking Gas	0.001	0.027	0.211	0.408	0.138	0.345	0.308	0.462	0.264	0.441	0.001	0.037
Piped natural gas	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.019
Processed biomass (Pellets)/Wood chips	0.000	0.000	0.001	0.035	0.000	0.016	0.002	0.047	0.000	0.000	0.000	0.000
Sawdust	0.000	0.020	0.000	0.012	0.000	0.016	0.000	0.000	0.000	0.022	0.001	0.025
Solar	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.000	0.000
Wood	0.000	0.000	0.781	0.414	0.627	0.484	0.000	0.000	0.678	0.467	0.000	0.000
Wood collected	0.600	0.490	0.000	0.000	0.000	0.000	0.561	0.496	0 .000	0.000	0.465	0.499
Wood purchased	0.110	0.313	0.000	0.000	0.000	0.000	0.059	0.236	0 .000	0.000	0.009	0.093
Observations	3734		2300		3808		3171		5928		2913	

	(1) ETH mean	 sd	(2) HND mean	 sd	(3) KEN mean	 sd	(4) KHM mean	 sd	(5) NPL mean	 sd	(6) ZMB mean	 sd
Clean fuel stove	0.040	0.196	0.213	0.409	0.141	0.348	0.328	0.469	0.284	0.451	0.151	0.358
Improved biomass stove	0.183	0.387	0.376	0.484	0.103	0.304	0.352	0.478	0.005	0.071	0.004	0.065
Kerosene stove	0.000	0.000	0.000	0.000	0.069	0.254	0.000	0.000	0.000	0.000	0.000	0.000
Three stone stove	0.642	0.479	0.077	0.267	0.616	0.486	0.047	0.213	0.150	0.357	0.474	0.499
Traditional stove	0.135	0.341	0.335	0.472	0.070	0.255	0.272	0.445	0.561	0.496	0.371	0.483
Observations	3741		2302		3807		3170		5928		2913	

Table 8. Caption

Table 9. Summary statistics of control variables

	mean	sd		mean	sd		mean	sd
ETH			HND			KEN		
CS aggregate tier	2.24	1.12	CS aggregate tier	2.59	1.00	CS aggregate tier	1.78	1.38
HH average women age	25.32	9.29	HH average women age	28.62	13.14	HH average women age	26.34	11.05
Sex of head	0.70	0.46	Sex of head	0.75	0.43	Sex of head	0.77	0.42
HH max educ	0.60	0.49	HH max educ	0.52	0.50	HH max educ	0.69	0.46
HH number of children	0.59	0.80	HH number of children	0.64	0.80	HH number of children	0.58	0.80
HH number of people	4.77	2.02	HH number of people	4.74	2.06	HH number of people	4.48	2.11
HH share of women	0.37	0.21	HH share of women	0.36	0.17	HH share of women	0.38	0.20
Non-clean stove	0.55	0.50	Non-clean stove	0.44	0.50	Non-clean stove	0.75	0.43
Rural	0.42	0.49	Rural	0.64	0.48	Rural	0.60	0.49
House structure	0.34	0.50	House structure	1.40	0.50	House structure	1.47	0.67
Wealth index	0.07	1.90	Wealth index	-0.19	1.98	Wealth index	0.04	2.32
Abadie kappa	0.22	1.04	Abadie kappa	0.58	0.82	Abadie kappa	0.60	0.89
Availability of solid fuel	0.53	0.50	Availability of solid fuel	0.66	0.47	Availability of solid fuel	0.70	0.46
Observations	3,769		Observations	2,677		Observations	3,833	
KHM			LBR			NER		
CS aggregate tier	2.66	1.24	CS aggregate tier	1.69	0.79	CS aggregate tier	2.31	0.96
HH average women age	30.65	11.24	HH average women age	24.92	8.40	HH average women age	23.88	7.95
Sex of head	0.70	0.46	Sex of head	0.75	0.43	Sex of head	0.89	0.31
HH max educ	0.72	0.45	HH max educ	0.72	0.45	HH max educ	1.00	0.00
HH number of children	0.56	0.72	HH number of children	0.71	0.89	HH number of children	0.85	0.99
HH number of people	4.57	1.93	HH number of people	5.58	2.40	HH number of people	5.72	3.39
HH share of women	0.40	0.19	HH share of women	0.36	0.19	HH share of women	0.33	0.18
Non-clean stove	0.25	0.43	Non-clean stove	1.00	0.03	Non-clean stove	0.59	0.49
Rural	0.50	0.50	Rural	0.48	0.50	Rural	0.20	0.40
House structure	1.84	0.39	House structure	1.41	0.67	House structure	1.24	0.86
Wealth index	0.05	2.22	Wealth index	0.03	1.36	Wealth index	1.93	4.47
Abadie kappa	0.38	0.90	Abadie kappa	0.70	0.66	Abadie kappa	-0.22	5.58
Availability of solid fuel	0.37	0.48	Availability of solid fuel	0.51	0.50	Availability of solid fuel	0.75	0.43
Observations	3,171		Observations	3,485		Observations	3,783	
NPL			RWA			STP		
CS aggregate tier	2.64	2.08	CS aggregate tier	1.69	0.73	CS aggregate tier	2.14	1.10
HH average women age	30.27	10.54	HH average women age	25.00	10.64	HH average women age	24.96	11.37
Sex of head	0.80	0.40	Sex of head	0.75	0.43	Sex of head	0.62	0.49
HH max educ	0.87	0.34	HH max educ	0.49	0.50	HH max educ	0.76	0.43
HH number of children	0.47	0.73	HH number of children	0.78	0.87	HH number of children	0.68	0.78
HH number of people	4.79	1.99	HH number of people	4.80	1.92	HH number of people	4.34	1.86
HH share of women	0.39	0.16	HH share of women	0.34	0.19	HH share of women	0.35	0.19
Non-clean stove	0.64	0.48	Non-clean stove	0.57	0.50	Non-clean stove	0.88	0.33
Rural	0.45	0.50	Rural	0.51	0.50	Rural	0.50	0.50
House structure	1.67	0.55	House structure	1.31	0.47	House structure	1.18	0.39
Wealth index	0.11	2.41	Wealth index	0.01	2.30	Wealth index	0.33	2.71
Abadie kappa	0.60	0.92	Abadie kappa	0.49	0.80	Abadie kappa	0.40	0.91
Availability of solid fuel	0.63	0.48	Availability of solid fuel	0.69	0.46	Availability of solid fuel	0.38	0.49
Observations	5,928		Observations	3,098		Observations	1,721	
ZMB			Total					
CS aggregate tier	2.11	1.13	CS aggregate tier	2.20	1.39			
HH average women age	25.13	10.80	HH average women age	27.07	10.86			
Sex of head	0.70	0.46	Sex of head	0.74	0.44			
HH max educ	0.69	0.46	HH max educ	0.70	0.46			
HH number of children	0.66	0.87	HH number of children	0.62	0.81			
HH number of people	5.56	2.38	HH number of people	4.88	2.16			
HH share of women	0.29	0.14	HH share of women	0.36	0.18			
Non-clean stove	0.88	0.33	Non-clean stove	0.65	0.48			
Rural	0.49	0.50	Rural	0.49	0.50			
House structure	1.29	0.87	House structure	1.35	0.72			
Wealth index	-0.20	2.12	Wealth index	0.08	2.28			
Abadie kappa	0.33	1.01	Abadie kappa	0.47	1.25			
Availability of solid fuel	0.43	0.49	Availability of solid fuel	0.57	0.50			
Observations	2,935		Observations	34,400				

References

Abadie A (2003) Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics* 113(2): 231–263.

Afridi F, Debnath S, Dinkelman T and Sareen K (2023) Time for clean energy? cleaner fuels and women's time in home production. *The World Bank Economic Review* 37(2): 283–304.

Allcott H (2015) Site selection bias in program evaluation. *The Quarterly Journal of Economics* 130(3): 1117–1165.

Angrist JD and Pischke JS (2009) *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Angrist JD and Pischke JS (2010) The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives* 24(2): 3–30. DOI:10.1257/jep.24.2.3. URL https://www.aeaweb.org/articles?id=10.1257/jep.24.2.3.

Bergeron A, Tourek G and Weigel J (2021) The State Capacity Ceiling On Tax Rates: Evidence From Randomized Tax Abatements In The Drc. CEPR Discussion Papers 16116, C.E.P.R. Discussion Papers. URL https://ideas.repec.org/p/cpr/ceprdp/16116.html.

Bhatia M and Angelou N (2015) Beyond connections .

Hotz VJ, Imbens GW and Mortimer JH (2005) Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics* 125(1): 241–270. DOI:https://doi.org/10.1016/j.jeconom.2004.04.009. URL https://www.sciencedirect.com/science/article/pii/S0304407604000806. Experimental and non-experimental evaluation of economic policy and models.

Imbens G, Crump R, Hotz VJ and Mitnik O (2009) Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96(1): 187–199.

Imbens GW (2010) Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). *Journal of Economic Literature* 48(2): 399–423. DOI:10.1257/jel.48.2.399. URL https://www.aeaweb.org/articles?id=10.1257/jel.48.2.399.

Krishnapriya P, Chandrasekaran M, Jeuland M and Pattanayak SK (2021) Do improved cookstoves save time and improve gender outcomes? evidence from six developing countries. *Energy Economics* 102: 105456.

Kurata M, Takahashi K and Hibiki A (2020) Gender differences in associations of household and ambient air pollution with child health: Evidence from household and satellite-based data in bangladesh. *World Development* 128: 104779. DOI:https://doi.org/10.1016/j.worlddev.2019.104779.

Leamer EE (1983) Let's take the con out of econometrics. *The American Economic Review* 73(1): 31–43. URL http://www.jstor.org/stable/1803924.

SDG T (2021) The energy progress report. *IEA: Paris, France* .

Stuart EA and Rhodes A (2017) Generalizing treatment effect estimates from sample to population: A case study in the difficulties of finding sufficient data. *Evaluation Review* 41(4): 357–388. DOI:10.1177/0193841X16660663. URL https://doi.org/10.1177/0193841X16660663. PMID: 27491758.

Su Q and Azam M (2023) Does access to liquefied petroleum gas (lpg) reduce the household burden of women? evidence from india. *Energy Economics* 119: 106529.

Tipton E (2013) Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics* 38(3): 239–266. URL http://www.jstor.org/stable/41999424.

Tipton E (2014) How generalizable is your experiment? an index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics* 39(6): 478–501. URL https://doi.org/10.3102/1076998614558486.

Tipton E, Hallberg K, Hedges LV and Chan W (2017) Implications of small samples for generalization: Adjustments and rules of thumb. *Evaluation review* 41(5): 472–505.

Tipton E and Peck LR (2017) A design-based approach to improve external validity in welfare policy evaluations. *Evaluation Review* 41(4): 326–356. DOI:10.1177/0193841X16655656. URL https://doi.org/10.1177/0193841X16655656. PMID: 27474752.

Verma AP and Imelda (2022) Clean Energy Access: Gender Disparity, Health and Labour Supply. *The Economic Journal* 133(650): 845–871. DOI:10.1093/ej/ueac057. URL https://doi.org/10.1093/ej/ueac057.

Williams KN, Kephart JL, Fandiño-Del-Rio M, Simkovich SM, Koehler K, Harvey SA and Checkley W (2020) Exploring the impact of a liquefied petroleum gas intervention on time use in rural peru: A mixed methods study on perceptions, use, and implications of time savings. *Environment International* 145: 105932. DOI:https://doi.org/10.1016/j.envint.2020.105932. URL https://www.sciencedirect.com/science/article/pii/S0160412020318870.