Int. J. Epidemiol. Advance Access published October 23, 2011

Published by Oxford University Press on behalf of the International Epidemiological Association
© The Author 2011; all rights reserved.
International Journal of Epidemiology 2011;1–9
doi:10.1093/ije/dyr150

Demonstrating bias and improved inference for stoves' health benefits

Valerie Mueller,[1]*,† Alexander Pfaff,[2,†] John Peabody,[3] Yaping Liu[4] and Kirk R. Smith[5]

[1]International Food Policy Research Institute, Development Strategy and Governance Division, Washington, DC, USA, [2]Duke University, Sanford School of Public Policy, Durham, NC, USA, [3]University of California, Department of Epidemiology & Biostatistics, San Francisco, CA, USA, [4]The First Hospital of Tsinghua University, Beijing, China and [5]Department of Environmental Health Sciences, University of California, School of Public Health, Berkeley, CA, USA

*Corresponding author. 2033 K Street, NW Washington, DC 20006 USA. E-mail: v.mueller@cgiar.org
†These authors contributed equally to this work

Accepted 14 September 2011

Background Many studies associate health risks with household air pollution from biomass fuels and stoves. Evaluations of stove improvements can suffer from bias because they rarely address health-relevant differences between the households who get improvements and those who do not.

Methods We demonstrate both the potential for bias and an option for improved stove inference by applying to household air pollution a technique used elsewhere in epidemiology, propensity-score matching (PSM), based on a stoves-and-health survey for China (15 counties, 3500 households).

Results Health-relevant factors (age, wealth, kitchen ventilation) do in fact differ considerably between the households with stove improvements and those without. We study the resulting bias in estimates of cleaner-stove impacts using a self-reported Physical Component Summary (PCS). Typical stoves-literature regressions with little control for non-stove factors suggest no benefits from a cleaner-fuel stove relative to a traditional biomass stove. Yet increasing controls raises the impact estimates. Our PSM estimates address the differences in health-relevant factors using 'apples to apples' comparisons between those with improved stoves and 'similar' households. This generates higher estimates of clean-stove benefits, which are on the order of one half the standard deviation of the PCS outcome.

Conclusions Our data demonstrate the potential importance of bias in household air pollution studies. This results from failure to address the possibility that those receiving improved stoves are themselves prone to better or worse health outcomes. It suggests the value of data collection and of study design for cookstove interventions and, more generally, for policy interventions within many health outcomes.

Keywords Cookstoves, improved stoves, household air pollution, propensity score matching, self-reported health, physical component summary

Downloaded from http://ije.oxfordjournals.org/ by guest on October 24, 2011

Introduction

Several studies link health risks with biomass energy use.[1] Most rely upon indirect measures of exposure, e.g. the fuel or stove used, for dose–response relationships.[2–10] It is difficult to track an individual's exposure, though it may help to know of behaviours that affect pollutant exposure for any given stove and fuel used (such as one's typical location relative to where cooking occurs) or to know about environmental variation relevant to exposure (e.g. local outdoor air pollution).

Another critical issue when assessing the impact of any cookstove improvement on health is that improved stoves may not be distributed randomly. We show this may be easier to address. Non-randomness could arise from government allocations of stove improvements to poorer or to politically connected or to high-pollution areas. Household choice can also drive the distribution.[11–12] For instance, those who like to stay nearer to their cooking may be more willing to invest in stoves.

Non-random distribution can bias estimates of the health impact of stove improvements. Health-relevant characteristics can differ between households with and without improved stoves, confounding inference. For instance, those receiving improved stoves could be poorer and, thus, when lacking controls for the effect of income on health, it is easy for the estimated impact of the stove to be biased down. Such differences can be addressed when estimating clean-stove impacts if a study collects data on characteristics associated with both health and those who receive the stove.

Recent efforts to address such non-randomness include randomized controlled trials.[13–15] Our results suggest benefit from randomization in reducing the bias when estimating the health benefits of interventions like improved stoves.[16–18] Yet, randomization is not always possible. However, by demonstrating that data describing those with and without improvements do help, we also suggest a way to improve impact estimates if randomization did not or will not occur.

Methods

Interventions

In the early 1980s, the Chinese government funded multiple programmes for stove improvements, including a National Improved Stove Program (NISP)[19] and others focused on cleaner stoves.[20] They included subsidies to households that depended on the stove and county. The NISP, for instance, supported 860 counties (of 2126 in China) with a first phase (until 1992) subsidizing dissemination. Counties had to apply and then were chosen based on criteria including energy shortages and the willingness to share the cost burden. This overlapped with other programmes.

NISP's second phase (1990–95) replaced subsidized stove distribution with incentives to rural energy industries. This overlapped with a programme started by the Ministry of Health in the 1990s to improve kitchens, e.g. in ventilation, in poor regions to target fluorosis. NISP's third phase included a response to the 1998 flooding of the Yangtze River. The Yangtze River Valley Environmental Protection Project was intended to reduce soil erosion by supporting reforestation in the region and it included promotion of improved stoves and coal to reduce fuel wood demand.

Data collection and measurement

We use a cross-sectional survey of ~3500 households in three provinces of China (Shaanxi, Hubei and Zhejiang) collected during 2001–03.[19] It includes adult health outcomes (ages $\geqslant$18 years), household demographics, fuel use and stove use by type and characteristic. We focus on stoves used primarily for cooking, as they generate most exposure to air pollutants: traditional biomass stoves (16%); improved biomass (47%); coal (32%); and clean-fuel (6%). Traditional biomass (TB) and improved biomass (IB) use wood, crop residues or dung. IB have at least a flue and a grate, whereas Clean stoves include electric, liquified petroleum gas or biogas. In terms of poor health outcomes, these stoves have been ranked: Coal, TB, IB and Clean.[21]

Self-reported health outcome

We analyse a measure of self-reported general health which is based upon 12 questions about physical and mental distress, the 'SF-12'.[22] We use the Physical Component Summary (PCS), an index based on the answers. The SF-12 is relatively easy to add to a standard household survey, as it does not require additional measurement techniques. A natural concern is that self-reports are subject to measurement error though recent work correlates them with a suite of illnesses.[23–25]

The questions in the SF-12 aim to elicit a sense of physical and mental distress. We focus on the physical distress in constructing the PCS, which has a standardized score from 1 to 100.[22] Scores better than 50 indicate above-average health, whereas lower than 50 means below average. The standard deviation (SD) of the PCS index is 10, which facilitates the interpretation of the index.

Statistical methods

Regression average treatment effect

We begin with regressions like in the stoves literature that estimate average treatment effect (ATE). We include indicator variables for the stoves plus relevant covariates such as kitchen ventilation, measured by kitchen openings, as well as the presence of an additional open-air kitchen.[26] Time near to the cookstove is captured by the number of minutes spent

Downloaded from http://ije.oxfordjournals.org/ by guest on October 24, 2011

cooking in a given day.[8–9,27] Characteristics that influence health, such as age (continuous or as 26–40, 41–55 and >55 years, with 18–25 years as the reference category), gender (indicator for male), income (continuous or as an indicator for over 12 000 Yuan), asset wealth (indicators for owning a washing machine and for owning a television) and whether one smokes are included. Finally, we use indicators for one's region (provinces or smaller counties) to control for unobserved influences that are common for everyone in the region in question but can vary across larger areas, e.g. variable county policies.

Table 3 presents five regressions of this sort. The first includes all of the users of all four types of stoves in our sample, with stove indicators (IB, Coal, Clean; TB is the reference) plus the household and individual covariates. The second regression adds province indicators, and the third replaces those with county indicators. The fourth regression follows the third but limits the sample to the Clean and TB stove users in the counties with sufficient numbers of each of those to compare. The fifth extends the fourth. For additional control for the influences of covariates, the continuous variables (age, income, cooking time) are included as linear and quadratic terms. We focus on the health impact of Clean versus TB stoves, the same comparison as in matching.

Interacting treatment with covariates

For two different reasons, we are interested in whether the stove impact varies with covariates. Methodologically, if impact does vary then matching's estimates of average treatment effect on the treated (ATT) can differ from the ATE (the average effect in the whole sample) if the stove distribution is non-random. For health policy reasons, we may care about the effects on specific subpopulations (e.g. young/old, rich/poor). Thus, within the regression results of Table 3, we also interact the treatment indicator with all the covariates: $\mathrm{PCS} = \gamma + \alpha T + \beta x + \delta(x - \mathrm{X})T + \varepsilon$, where T and x refer to the treatment and covariates and X is the sample average for x (we bootstrap this interaction regression 1000 times in order to compute Table 3 standard errors).[28] This implies estimated stove impacts which vary by observation, permitting calculation of various averages.

The ATE is simply α, the average of those varied estimates across the full sample, whereas ATT focuses on the covariate space where stove improvements went (averaging impact estimates over the treated observations: $\alpha + (\Sigma_i T_i)^{-1}[\Sigma_i\ T_i(x_i - X)\delta]$ [28]). Sixth and seventh regressions in Table 3 report average impacts for the observations within two of our age categories, to compare the younger with the older, as only age's interaction with the stove treatment consistently has lower P values.

Using treatment propensity

Our final regression estimates provide a bridge between our regressions and our matching effort by using as a weight for each observation a summary index of all of its covariates, the propensity to be treated or probability that given those covariate values, a given household receives a stove. This is produced by first estimating a probit regression for whether a Clean stove was received, using our covariates. For instance, wealth might raise or lower the chance of receiving a stove. The predicted probabilities of being treated, for each individual, are referred to as 'propensity scores'. Eighth and ninth regressions in Table 3 use those as weights, thus focusing inference on the covariate space where improved stoves were provided. The ninth regression also includes in the covariates the linear and quadratic terms included in the fifth regression, as additional controls.

Propensity score matching

Propensity Score Matching (PSM[29,30]) is typically used to estimate the average treatment effect on the treated (ATT) accounting for non-random treatment by comparing households with Clean stoves (treated) to the 'most similar' users of TB stoves (untreated). Matching has been applied in health[31–34] but has not been applied within stove evaluation. Our goal is to demonstrate that it could help significantly to reduce bias in estimated stove impacts due to nonrandom stove allocation.

PSM uses the propensity score to measure similarity across the households. We compare the households with Clean stoves to the TB households with the most similar propensity scores. For those two groups, the matching ATT equals: $(1/N)\ [\sum_t H_t - \sum_u (w/n)H_u]$ where: H is outcome; t and u indicate treated and untreated; N is the number treated, i.e. number of Clean stove users; w represents the number of times that an observation in the untreated or control group has been matched and n is the number of untreated controls that is used to match with each treated case.

Table 4 presents 10 variations upon this approach, i.e. five for each of two specifications distinguished by using covariates either as in most of Table 3 (Specification A) or with the linear and quadratic terms for age, income and cooking time (Specification B). We present the results from these matching comparisons but also discuss the impact estimates from regressions, using only the treated and matched untreated observations, that further control for covariate differences between the treated and matched untreated groups, which remain even for the most similar pairs.

For each of the two specifications, Table 4 presents five matching variants. First is the kernel approach that uses all of the untreated cases but places higher weight upon the untreated with most similar propensity scores. This has the worst balances, i.e. the

Downloaded from http://ije.oxfordjournals.org/ by guest on October 24, 2011

least similar matches. We focus then on four variants of nearest-neighbour matching, comparing just a few most similar untreated observations with each treated case. The second and seventh rows use the two best matches (i.e. $m=2$) whereas the third and eighth rows use the three best matches. The latter provides more data but the additional untreated match is less similar. The fourth and ninth rows add to the second and seventh, a 'caliper' to drop the treated cases for which there are no untreated case with a propensity score within 0.1, whereas the 5th and 10th rows apply the same approach for a refinement upon the third and eighth rows.

Results

Initial differences in stove users

Table 1 shows stoves are not used equally across counties. TB stoves are not present in a number of counties but in three counties they make up over 70% of the stoves. IB stoves, over 75% in eight counties, are not in Yanchuan or Hancheng (the latter has almost all Coal). Clean stoves are not present in some counties but amount to 17% of the stoves in Changyang as well as 31% of the stoves in Kaihua. We address these differences.

Table 1 also shows county averages for observed individual and household characteristics. The youngest age category varies from 25% to >50% of the sample and the oldest goes from <10% to over 33%. Across counties, the fraction earning over 12 000 Yuan varies from 1% to 85%. Concerning ventilation, the fraction with a single kitchen opening varies from <5% to >60% and the same is true for the fraction with more than two kitchen openings. Based on these important variations, we believe controlling for county could matter empirically. In addition to differences in observed characteristics and in stoves used, that are discussed above, there may also be unobserved characteristics that vary across counties and which matter for health.

Matching reduces differences

Table 2 conveys success in constructing similar groups to compare, in terms of not only the counties but also all of the other observable factors. It reports the means for the treated, the untreated and the matched untreated using the two most similar matches (i.e. second row specification in Table 4), as well as the absolute standardized differences of the means plus tests for differences in means.

The higher P values for the differences in the means of the matched untreated and treated shows that matching moved towards 'apples to apples' comparison. Standardized differences fall for most of the covariates and for a few that rise, often the initial differences are small.[35] Looking at the counties, the P values are quite low to start for three of four, then rise a lot with matching.

Perhaps the most meaningful such reductions in differences are for ventilation and income. Pre-matching, the treated or Clean households have less kitchen ventilation, as seen in the higher fraction with one

Table 1 Stove and user characteristics (fraction within county)

	Stoves					Stove user characteristics						
						Age (years)			Income	One	Two	>Two
County	*N*	TB (563)	IB (1668)	Coal (1150)	Clean (206)	26–49	41–55	>55	>12 000 Yuan	kitchen opening	kitchen openings	kitchen openings
Fuping	275	0.00	0.00	0.99	0.01	0.58	0.27	0.08	0.13	0.15	0.40	0.43
Heyang	217	0.03	0.01	0.96	0.00	0.62	0.21	0.08	0.05	0.13	0.77	0.07
Lintong	332	0.07	0.65	0.23	0.05	0.47	0.29	0.15	0.15	0.21	0.32	0.43
Yanchuan	256	0.71	0.00	0.29	0.00	0.38	0.32	0.24	0.01	0.05	0.89	0.04
Hancheng	196	0.00	0.00	0.96	0.04	0.50	0.32	0.09	0.13	0.54	0.19	0.03
Suizhou	263	0.02	0.92	0.03	0.03	0.55	0.27	0.14	0.21	0.46	0.28	0.17
Changyang	236	0.03	0.75	0.05	0.17	0.55	0.27	0.13	0.20	0.42	0.38	0.17
Tongcheng	116	0.28	0.10	0.57	0.04	0.52	0.34	0.05	0.24	0.44	0.40	0.16
Xiantao	245	0.02	0.60	0.30	0.08	0.58	0.24	0.12	0.21	0.22	0.37	0.37
Yicheng	336	0.00	0.47	0.51	0.02	0.50	0.29	0.10	0.21	0.13	0.65	0.22
Anji	242	0.09	0.88	0.00	0.04	0.26	0.51	0.19	0.31	0.63	0.31	0.04
Kaihua	164	0.01	0.68	0.00	0.31	0.34	0.45	0.18	0.29	0.51	0.34	0.08
Xianju	163	0.00	0.94	0.00	0.06	0.52	0.25	0.21	0.30	0.18	0.33	0.44
Chunan	221	0.10	0.81	0.00	0.09	0.25	0.37	0.36	0.17	0.35	0.32	0.31
Tongxiang	325	0.79	0.18	0.00	0.03	0.38	0.32	0.24	0.85	0.07	0.16	0.75

Downloaded from http://ije.oxfordjournals.org/ by guest on October 24, 2011

Table 2 Balance improvements through matching

Stove Subgroup Observations Variable	**Clean** All 59 Mean	**TB** All 358 Mean	ASD[a]	$P>\|t\|$	**TB** Matched, two most similar 118 Mean	ASD[a]	$P>\|t\|$
Age range (years)							
26–40	0.56	0.39	0.35	0.01	0.50	0.12	0.52
41–55	0.25	0.32	0.14	0.32	0.22	0.08	0.67
>55	0.14	0.25	0.28	0.06	0.22	0.22	0.23
Male	0.37	0.30	0.15	0.28	0.36	0.02	0.93
Income > 12 000 Yuan	0.51	0.67	0.33	0.02	0.47	0.07	0.72
Washing machine	0.58	0.57	0.02	0.90	0.54	0.07	0.71
Television	0.97	0.94	0.11	0.49	0.92	0.20	0.32
Cook time (minutes/day)	100	100	0.00	0.99	114	0.24	0.19
Smoker	0.31	0.21	0.23	0.09	0.30	0.02	0.92
Kitchen openings							
1	0.37	0.18	0.45	0.00	0.31	0.16	0.44
2	0.31	0.23	0.18	0.19	0.39	0.19	0.34
>2	0.29	0.59	0.63	0.00	0.28	0.02	0.92
Kitchen open-air = Y	0.03	0.01	0.20	0.04	0.02	0.12	0.56
County							
Lingtong	0.27	0.06	0.57	0.00	0.35	0.21	0.37
Tongcheng	0.08	0.09	0.03	0.86	0.04	0.15	0.35
Anji	0.15	0.06	0.31	0.01	0.09	0.19	0.33
Chunan	0.32	0.06	0.69	0.00	0.38	0.16	0.50

[a]Absolute Standardized Difference (see, e.g. Stuart[36]) indicates the difference between these subsample means as a fraction of the square root of the average of the sample variances in the treated and non-treated groups.

kitchen opening and lower fraction with more than two openings. Clean stove owners also have lower average income which, like fewer kitchen openings, could mask gains from the stove. These differences bias downward Clean impacts as less healthy people use Clean. Matching greatly reduces these critical differences, as seen in the considerably higher *P* values.

One other important difference is reduced but from a starting point with the opposite bias. Clean stove owners are considerably younger before matching, which unlike the income and the ventilation differences could bias estimates in typical regressions towards a finding of benefits. Since matching reduces this difference, the effect of matching below is not due to this influence.

Clean stove impacts

Table 3 presents regression estimates of the health impact of Clean stoves relative to TB stoves. Every column includes the covariates, and the variations in additional efforts to control for the potentially confounding influences of non-stove factors are indicated. Comparing the second and third specifications with the first confirms the importance of regional controls, as the impact coefficient rises and the *P* value falls considerably just from adding the indicators.

Given the apparent importance of counties, within Tables 1 and Table 3 columns (1)–(3), the fourth and fifth specifications restrict the sample to those counties with sufficient numbers of both Clean and TB stoves to permit comparison. The fifth specification adds non-linear (i.e. linear and quadratic) terms for continuous age, income and cooking time. As within columns (1)–(3), the increases in controls again both raise the impact coefficients and lower the *P* value, more than doubling the stove-impact coefficient from column (3) and more than halving *P* values, e.g. down to 0.05 within column (5).

Sixth and seventh specifications in Table 3 introduce another form of additional controls, i.e. the interactions of treatments with characteristics, yielding an impact estimate for each observation. As only age's interactions dependably have low *P* values, columns (6) and (7) show average impacts for

Downloaded from http://ije.oxfordjournals.org/ by guest on October 24, 2011

Table 3 Regression estimates for Clean-stove impacts on self-reported health

Specification	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Clean effect	−0.05	0.56	0.95	1.98	2.57	3.31	3.37	2.97	4.35
For whom?	Sample	Sample	Sample	Sample	Sample	Age 26–40 years	Age >55 years	Weighted	Weighted
95% CI	(−1.39 to 1.29)	(−0.77 to 1.89)	(−0.62 to 2.53)	(−0.57 to 4.52)	(−0.05 to 5.19)	(0.96–5.65]	(0.64–6.10)	(0.30–5.63)	(1.04–7.65)
P value	0.94	0.41	0.24	0.13	0.05	0.01	0.02	0.03	0.01
Sample									
Stoves	All	All	All	Clean,TB	Clean,TB	Clean,TB	Clean,TB	Clean,TB	Clean,TB
Counties	All	All	All	Select[a]	Select[a]	Select[a]	Select[a]	Select[a]	Select[a]
Including									
All observables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Spatial dummies	None	Province	County	County	County	County	County	County	County
Quadratics[b]	No	No	No	No	Yes	No	No	No	Yes
Interactions[c]	No	No	No	No	No	Yes	Yes	No	No
PS weights[d]	No	No	No	No	No	No	No	Yes	Yes
N	3587	3587	3587	417	417	417	417	417	417
Adjusted R^2	0.11	0.13	0.14	0.15	0.15	0.16	0.16	0.31	0.35

[a]For within-county comparisons here, we selected only the counties with more than a trivial number of each of the Clean and TB stove types.
[b]Age, income and cooking time are included using both their linear and their quadratic terms (adding cubics and quartics does not add further).
[c]To allow for heterogeneous treatment effects, the Clean stove indicator is interacted with each of the observable covariates that we include here.
[d]To reflect which groups were actually treated, i.e. received Clean stoves, here we weight all observations using the probability of being treated.

Downloaded from http://ije.oxfordjournals.org/ by guest on October 24, 2011

Table 4 Matching estimates for Clean-stove impacts on self-reported health

Specification	ATT (95% CI[a])	*P* value
A (all observables)		
(01) Propensity Score Matching (kernel density algorithm)	4.43 (0.07–8.80)	0.05
(02) Propensity Score Matching (nearest neighbour algorithm, $m=2$)	4.86 (−0.56 to 10.27)	0.08
(03) Propensity Score Matching (nearest neighbour algorithm, $m=3$)	3.37 (−1.28 to 8.01)	0.16
(04) Propensity Score Matching (nearest neighbour algorithm, $m=2$, with caliper = 0.1)	5.04 (0.20–9.88)	0.04
(05) Propensity Score Matching (nearest neighbour algorithm, $m=3$, with caliper = 0.1)	4.13 (−0.48 to 8.74)	0.08
B (continuous age, income, cooking time using linear and squared terms)		
(06) Propensity Score Matching (kernel density algorithm)	7.54 (−0.18 to 15.27)	0.06
(07) Propensity Score Matching (nearest neighbour algorithm, $m=2$)	7.36 (−0.12 to 14.83)	0.05
(08) Propensity Score Matching (nearest neighbour algorithm, $m=3$)	5.03 (−1.48 to 11.55)	0.13
(09) Propensity Score Matching (nearest neighbour algorithm, $m=2$, with caliper = 0.1)	5.52 (−1.30 to 12.35)	0.11
(10) Propensity Score Matching (nearest neighbour algorithm, $m=3$, with caliper = 0.1)	4.73 (−1.45 to 10.90)	0.13

[a]Bootstrapped standard errors with 1000 repetitions.

households in two age categories. They are similar and are somewhat higher than column (5)'s estimate. These no longer represent the whole sample in columns (1)–(3) or whole restricted sample in columns (4) and (5), whereas columns (8) and (9) return to the full restricted sample using propensity-score weights to focus on the covariate space where the Clean stoves went. All of columns (6)–(9) have higher impact coefficients and lower *P* values even than column (5), demonstrating that increasing controls raises impact estimates.

Table 4 furthers that trend with matching to better achieve 'apples to apples' comparisons, albeit for the covariate space where Clean stoves went, i.e. the most similar owners of TB stoves. As noted, (01) and (06) have the worst balances, coming least close to good Clean–TB matches, whereas the balance for (02) is good (Table 2) and those of (03)–(05) are quite similar to (02). That supports a summary of Specification A as significant impact coefficients ranging around 4. We note that even for good average balances, as in Table 2 for (02), individual matched pairs are not identical, so we also run post-matching Ordinary Least Squares regressions, using only the treated and matched untreated observations. They support a summary of Specification A as coefficients around 3.5.

Table 4's Specification B increases the controls further by matching on the linear and the quadratic terms for continuous age, income and cooking time. In this case, the $m=2$ specification (07), like (06), does not do well in balancing asset indicators of wealth, whereas the $m=3$, i.e. (08), has good balances of covariate means, as in Table 2, and (09) and (10) are similar. These suggest the summary of Specification B as impact coefficients ranging around 5, and adding the post-matching Ordinary Least Squares (OLS) regressions supports the summary of Specification B being coefficients >4.

Discussion

We demonstrated significant potential for bias within a typical stove regression analysis due to limited controls for confounding influences of non-stove factors (here, Clean stove owners were poorer and had worse kitchen ventilation, biasing the estimates). This has implications for public health research in terms of both data collection and study design.

For example, this suggests the value of randomization of stove treatments if it is feasible, since that should eliminate any association of treatment with other factors that influence health. Recent studies have made this point and in a few places have randomized stove interventions,[13–18] yet there is significant ongoing debate on pros and cons of randomized control trials (RCTs)[37–39] and our sense is that this is not now the dominant approach within the whole stoves community.

Our results from greater efforts to control in regressions and from moving into matching also suggest significant value from investments in data collection to permit increases in controls. Increased household surveys alone can reduce confounding influences upon estimated impacts. Matching will not always change results and we have shown it is not the only approach to this. Further, matching is not a panacea, of course. Estimates may suffer from omitted variable biases, though difference-in-difference matching estimators with panel data can control for individual-specific, stable characteristics correlated with covariates but unobservable to the researcher.[40]

Downloaded from http://ije.oxfordjournals.org/ by guest on October 24, 2011

Putting these results and implications in context, the strengths of this study are our novel approach for the stoves literature, a large sample size and the availability of important covariates within the data set. Limitations include the use of cross-sectional data and just a single season, as well as an inability in our data to identify and try to control for effects of being a passive smoker. Summarizing, despite data limitations we believe the potential for significant bias in the typical stove regressions has been demonstrated, as has the value of various increased efforts to control.

Acknowledgements

We thank Travis Riddell, Ben Arnold and Jack Colford for extensive feedback, Subhrendu Pattanayak for a helpful discussion, and participants at the AERE Environment and Health workshop and an AERE session in the ASSA conference for their many useful comments.

Conflict of interest: None declared.

KEY MESSAGES

- Stove improvements rightfully receive attention for their potential to improve health.
- Evaluation of their impacts, in light of stoves allocations, has received less attention.
- We demonstrate that controlling for the differences in health-relevant characteristics between households with improved stoves and those without them affects evaluation.
- Such results support randomization in stoves allocation, as one way to study impacts. However, with or without that design, they also show real value from data collection.

References

[1] Smith KR, Mehta S, Maeusezahl-Feuz M. Indoor smoke from household solid fuels. In: Ezzati M, Rodgers A, Lopez AD *et al*. (eds). *Comparative Quantification of Health Risks: Global and Regional Burden of Disease Due to Selected Major Risk Factors*. Geneva, Switzerland: World Health Organization, 2004, pp. 1435–93.

[2] Orozco-Levi M, Garcia-Aymerich J, Villar J *et al*. Wood smoke exposure and risk of chronic obstructive pulmonary disease. *Eur Respir J* 2006;**27:**542–46.

[3] Triche EW, Belanger K, Bracken MB *et al*. Indoor heating sources and respiratory symptoms in nonsmoking women. *Epidemiology* 2005;**16:**377–84.

[4] Chapman RS, He X, Blair AE *et al*. Improvement in household stoves and risk of chronic obstructive pulmonary disease in Xuanwei, China: a retrospective cohort study. *BMJ* 2005;**11:**150–60.

[5] Mishra V, Xiaolei D, Smith K, Lasten M. Maternal exposure to biomass smoke and reduced birth weight in Zimbabwe. *Ann Epidemiol* 2004;**14:**740–47.

[6] Bruce N, McCracken J, Albalak R *et al*. Impact of improved stoves, house construction, and child location on levels of indoor air pollution exposure in young Guatemalan children. *J Expo Anal Environ Epidemiol* 2004;**14:**S26–33.

[7] Boy E, Nigel B, Hernan D. Birth weight and exposure to kitchen wood smoke during pregnancy in rural Guatemala. *Environ Health Persp* 2002;**110:**109–14.

[8] Ezzati M, Kammen D. Indoor air pollution from biomass combustion and acute respiratory infections in Kenya: an exposure-response study. *Lancet* 2001;**358:**619–24.

[9] Ezzati M, Kammen D. Quantifying the effects of exposure to indoor air pollution from biomass combustion on acute respiratory infections in developing countries. *Environ Health Persp* 2001;**109:**481–88.

[10] Smith KR, Samet JM, Romieu I, Bruce N. Indoor air pollution in developing countries and acute lower respiratory infections in children. *Thorax* 2000;**55:**518–32.

[11] Pattanayak SK, Pfaff A. Behavior, environment and health in developing countries: evaluation and valuation. *Ann Rev Resource Econ* 2009;**1:**183–217.

[12] Pfaff A., Chaudhuri S., Nye H. Household production & environmental Kuznets Curves: examining the desirability and feasibility of substitution. *Envir Resource Econ* 2004;**27:**187–200.

[13] Smith-Sivertsen T, Diaz E, Pope D *et al*. Effect of reducing indoor air pollution on women's respiratory symptoms and lung function: the RESPIRE Randomized Trial, Guatemala. *Am J Epidemiol* 2009;**170:**211–20.

[14] McCracken JP, Smith KR, Diaz A, Mittleman MA, Schwartz J. Chimney stove intervention to reduce long-term wood smoke exposure lowers blood pressure among Guatemalan women. *Environ Health Persp* 2007;**115:** 996–1001.

[15] Smith KR, McCracken JM, Weber MW *et al*. RESPIRE: A randomised controlled trial of the impact of reducing household air pollution on childhood pneumonia in Guatemala. *The Lancet* in press, 2011.

[16] Smith KR, Samet JM, Romieu I, Bruce N. Indoor air pollution in developing countries and acute lower respiratory infections in children. *Thorax* 2000;**55:**518–32.

[17] Bruce NL, Rehfuess E, Mehta S, Hutton G, Smith K. Indoor air pollution. In: Jamison DT *et al*. (eds). *Disease Control Priorities in Developing Countries*. 2nd edn. Washington D.C.: World Bank, New York: Oxford University Press, 2006.

[18] Duflo E, Greenstone M, Hanna R. Cooking stoves, indoor air pollution and respiratory health in rural Orissa. *Econ Polit Weekly* 2008;**43:**71–76.

Downloaded from http://ije.oxfordjournals.org/ by guest on October 24, 2011

[19] Sinton JE, Smith KR, Peabody JW *et al*. An assessment of programs to promote improved household stoves in China. *ESD* 2004;**8:**33–52.
[20] Xiaohua W, Jingfei L. Influence of using household biogas digesters on household energy consumption in rural areas: a case study in Lianshui County in China. *Renew Sust Energ Rev* 2005;**9:**229–36.
[21] Peabody J, Riddell T, Smith KR *et al*. Indoor air pollution in rural China: cooking fuels, stoves, and health status. *Arch Environ Occup Health* 2005;**60:**1–10.
[22] Ware JE, Kosinski M, Keller SD. *SF-12: How to Score the SF-12 Physical and Mental Health Summary Scales*. Boston, MA: The Health Institute, New England Medical Center, 1995.
[23] DeSalvo K, Jones T, Peabody J *et al*. Health care expenditure prediction with a single item, self-rated health measure. *Med Care* 2009;**47:**440–47.
[24] Butrick E, Peabody JW, Solon O, Desalvo KB, Quimbo SA. A comparison of objective biomarkers with a subjective health status measure among children in the Philippines. *Asia-Pacific J Public Health*. doi:10.1177/1010539510390204 [Epub 15 December 2010].
[25] Peabody J, Nordyke R, Tozija F *et al*. Quality of care and its impact on population health: a cross-sectional study from Macedonia. *Soc Sci Med* 2006;**62:**2216–24.
[26] Dasgupta S, Huq M, Khaliquuzaman M, Pandey K, Wheeler D. Indoor air quality for poor families: new evidence from Bangladesh. *Indoor Air* 2006;**16:**426–44.
[27] Ezzati M, Saleh H, Kammen D. The contributions of emissions and spatial microenvironments to exposure to indoor air pollution from biomass combustion in Kenya. *Environ Health Persp* 2000;**108:**833–39.
[28] Wooldridge J. *Econometric Analysis of Cross-Section and Panel Data*. Cambridge: Massachusetts Institute of Technology Press, 2002.
[29] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;**70:**41–55.
[30] Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc* 1984;**79:**516–24.
[31] Scott JC. *Water Supply, Sanitation, and Gastrointestinal Illness: Estimating the Risk of Disease Using Cross-sectional Data and Marginal Structural Models*. Saarbrucken: VDM, 2008.
[32] Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. *Am J Epidemiol* 2006;**163:**1149–56.
[33] Cheng YW, Hubbard A, Caughey A, Taylor I. The association between persistent fetal occiput posterior position and perinatal outcomes: an example of propensity score and covariate distance matching. *Am J Epidemiol* 2010;**171:**656–63.
[34] Lunt M, Solomon D, Rothman K *et al*. Different methods of balancing covariates leading to different effect estimates in the presence of effect modification. *Am J Epidemiol* 2009;**169:**909–17.
[35] Abadie A, Imbens G. On the failure of the bootstrap for matching estimators. *Econometrica* 2008;**76:**1537–57.
[36] Stuart E. Matching methods for causal inference: a review and a look forward? *Stat Sci* 2010;**1:**1–21.
[37] Banerjee AV, Duflo E. The experimental approach to development economics. *Annu Rev Econ* 2010;**1:**151–78.
[38] Heckman JJ. Randomization and social policy evaluation. In: Manski C, Garfinkel I (eds). *Evaluating Welfare and Training Programs*. Cambridge, MA: Harvard University Press, 1992.
[39] Deaton A. Instruments, randomization, and learning about development. *J Econ Lit* 2010;**48:**424–55.
[40] Smith J, Todd P. Does matching overcome lalonde's critique of nonexperimental estimators? *J Econometrics* 2005;**125:**305–53.

Downloaded from http://ije.oxfordjournals.org/ by guest on October 24, 2011