

How many replicate tests are needed to test cookstove performance and emissions? – Three is not always adequate



Yungang Wang^{a,*}, Michael D. Sohn^a, Yilun Wang^b, Kathleen M. Lask^c,
Thomas W. Kirchstetter^{a,d}, Ashok J. Gadgil^{a,d}

^a Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA

^b ISO Innovative Analytics, San Francisco, CA 94111, USA

^c University of California, Berkeley, College of Engineering, Applied Science and Technology Program, Berkeley, CA 94720, USA

^d University of California, Berkeley, Department of Civil and Environmental Engineering, Berkeley, CA 94720, USA

ARTICLE INFO

Available online 13 March 2014

Keywords:

Cookstove
Berkeley–Darfur Stove
Variability
Confidence interval
Kolmogorov–Smirnov test
Bootstrap

ABSTRACT

Almost half of the world's population still cooks on biomass cookstoves of poor efficiency and primitive design, such as three stone fires (TSF). Emissions from biomass cookstoves contribute to adverse health effects and climate change. A number of improved cookstoves with higher energy efficiency and lower emissions have been designed and promoted across the world. During the design development, and for the selection of a stove for dissemination, the stove performance and emissions are commonly evaluated, communicated and compared using the arithmetic average of replicate tests made using a standardized laboratory-based test, commonly the water boiling test (WBT). However, the statistics section of the test protocol contains some debatable concepts and in certain cases, easily misinterpreted recommendations. Also, there is no agreement in the literature on how many replicate tests should be performed to ensure “confidence” in the reported average performance (with three being the most common number of replicates). This matter has not received sufficient attention in the rapidly growing literature on stoves, and yet is crucial for estimating and communicating the performance of a stove, and for comparing the performance between stoves. We illustrate an application using data from a number of replicate tests of performance and emission of the Berkeley–Darfur Stove (BDS) and the TSF under well-controlled laboratory conditions. Here we focus on two as illustrative: time-to-boil and emissions of PM_{2.5} (particulate matter less than or equal to 2.5 μm in diameter). We demonstrate that an interpretation of the results comparing these stoves could be misleading if only a small number of replicates had been conducted. We then describe a practical approach, useful to both stove testers and designers, to assess the number of replicates needed to obtain useful data from previously untested stoves with unknown variability.

© 2014 International Energy Initiative. Published by Elsevier Ltd. All rights reserved.

Introduction

About half of the world's population uses biomass as fuel for cooking (IEA, 2004). The smoke from biomass cooking fires was recently found to be the largest environmental threat to health in the world, and is associated with 4 million deaths each year (Lim et al., 2012). This exposure has also been linked to adverse respiratory, cardiovascular, neonatal, and cancer outcomes (Smith et al., 2004; Weinhold, 2011). A 2011 World Bank report notes significant contributions of biomass cooking to global climate change (World Bank, 2011). The contribution to climate change from black carbon (BC) emission from biomass cooking is a topic of growing interest, especially in terms of climate forcing and melting of glaciers (Hadley et al., 2010; Ramanathan and

Carmichael, 2008). Current biomass stoves lead to a large burden of disease, and contribute to adverse impacts on local and the global environment. Hence there is substantial interest in developing and disseminating fuel-efficient biomass stoves with reduced emissions (e.g. DOE, 2011). Launched in September 2010, the Global Alliance for Clean Cookstoves (GACC) “100 by 20” goal calls for 100 million homes to adopt clean and efficient stoves and fuels by 2020.

The “three-stone fire” (TSF) is a commonly prevailing cooking method for a large fraction of the population at the base of the economic pyramid. In quantifying the performance of an improved stove, the TSF is commonly used as the baseline. This least expensive class of stove is simply an arrangement of three large stones supporting a pot over an open and unvented biomass fire. A TSF is one of the two stoves we analyzed in this study. We also tested the performance and emissions of the Berkeley–Darfur Stove (BDS) as an exemplar of an improved fuel-efficient biomass cookstove. The BDS was developed at Lawrence Berkeley National Laboratory (LBNL) for internally displaced persons in

* Corresponding author. Tel.: +1 510 495 8065.
E-mail address: yungangwang@lbl.gov (Y. Wang).

Darfur, Sudan (<http://cookstoves.lbl.gov/darfur.php>). It is an all-metal precision-designed natural-convection stove, with design features co-developed by iterative feedback from Darfuri women cooks. The BDS by design accommodates Darfuri traditional round-bottom cooking pots and cooking techniques (Fig. 1).

A literature survey of recent laboratory cookstove testing in peer-reviewed journal articles shows widely different numbers of replicate tests (Bailis et al., 2007; Jetter and Kariher, 2009; Jetter et al., 2012; MacCarty et al., 2008, 2010; Roden et al., 2009; Smith et al., 2007). The number of replicates reported in these seven studies range from 1 to 23. However, six out of seven studies have reported results with only 3 or fewer replicates. One then can rightly ask: how many replicate tests do I need to test the performance and emissions of the stove? Answering this question is application specific, and requires greater specificity. For example, the question might be better phrased. For a water boiling test (WBT), how many replicates are needed to estimate the average “time to boil” to within 2 min and with 95% confidence? Or how many replicates are needed to confirm, with 95% confidence, that Stove “A” emits less $PM_{2.5}$ than Stove “B”? These questions exemplify perhaps the most frequently asked questions in planning stove experiments and interpreting their results.

There is no single or simple answer to the number of replicates needed to answer the above questions. The answer depends on the experimental design, how many parameters need to be estimated, and the resulting variability in the stove replicates. In this study, we investigate how to answer the above questions using data from the BDS and TSF water boiling experiments. We show how the number of replicates is linked to uncertainty and variability in the experiments and stove performance. We also show how many replicates are likely needed as various practical performance comparisons, such as “Does Stove A perform

better than Stove B?” and “What is the uncertainty in the expected performance of Stove A or Stove B?” Finally, we describe a practical approach to design an experiment to test the performance of a previously untested stove.

Problem statement and causes of variability

Appendix 6 of the WBT (version 3.0, <http://www.pciaonline.org/node/1048>) provides a detailed approach for comparing the performance of stoves. It describes a suite of test statistics and important considerations for interpreting test results. While comprehensive, the description contains some debatable concepts and in certain cases, easily-misinterpreted recommendations. For example, it affirms that “At least three tests should be performed on each stove” and provides a cogent explanation for it. It also discusses the importance of paying attention to the statistical significance of a series of comparison tests between the performances of two stoves. While both statements are correct, it is not surprising that stove testers misinterpret these comments as (i) “only three tests are needed” or (ii) a hypothesis test with strong p-value (assuming a Gaussian distribution) provides unarguable confirmation of stove performance or comparison results. In fact, neither interpretation is correct or claimed in the text. We reason further that elucidation of Appendix 6 is necessary, and a more transparent methodology would greatly benefit stove testers. We believe that a transparent methodology would be best accomplished by developing an approach that maps the trade space between sample size, variability, and confidence. We also believe it is important to show that alternative methods for comparing the performances of stoves are available and should be considered. This work thus builds and improves upon Appendix 6 by providing new methods of interpreting test results for stove testers.

The literature generally shows that even under carefully controlled conditions, stove test results show high test-to-test variability (coefficient of variation > 1.0 , e.g. Jetter et al., 2012). There are many possible causes of this variability even within a precisely defined test such as the latest WBT (version 4.2.2), and we list a few here. Stove efficiency and emissions are generally a function of thermal power, and owing to the discrete nature of fuel-feeding events, a stove’s thermal power invariably varies, also contributing to temporal variability within a test, which can translate into test-to-test variability. Despite due care, the ratio of bark to sapwood for various pieces of fuelwood can be different, and thus will have different burn characteristics. Furthermore, different pieces of fuelwood may have different surface to volume ratios, contributing to different rates of burning. Lastly, even reasonably experienced and careful stove testers demonstrate some variability in the way they tend the fire in the stove from test to test, and within a test (Granderson et al., 2009). All these (and other uncontrolled factors) together give rise to what we lump together as variability in the test-to-test replicate results for a stove under controlled laboratory conditions.

Approach

The question of “How many replicate tests do I need?” is not novel. It is a well-researched question in classical statistical theory, but has not received much attention from the stove research community. We briefly summarize here the statistical background relevant to answer the question.

Probability density function and cumulative distribution function

Technically, for a continuous random variable, the probability density function (PDF) describes the probability that a value will be within a certain range of the sample. However, as this range is evaluated by integrating, it can be chosen to be quite small, so for most practical purposes, the PDF may be considered the probability of obtaining a



Fig. 1. Schematic of the Berkeley-Darfur Stove. (1) A tapered wind collar that increases fuel-efficiency in the windy Darfur environment and allows for multiple pot sizes; (2) wooden handles for easy handling; (3) metal tabs for accommodating flat plates for bread baking; (4) internal ridges for optimal spacing between the stove and a pot for maximum fuel efficiency; (5) feet for stability with optional stakes for additional stability; (6) nonaligned air openings between the outer stove and inner fire box to accommodate windy conditions; and (7) small fire box opening to prevent using more fuel wood than necessary.

particular value (Ellison et al., 2009). Graphically, if the PDF is a curve, the cumulative distribution function (CDF) is the area under that curve. It is used to compute probability; the larger the included range, the greater the probability. Because of this, the CDF over the entire range is equal to 1. For a normal (or Gaussian) distribution, the CDF curve is a normal ogee curve, which is a smooth even S-shaped curve (Ellison et al., 2009). Skewing in the distribution away from the Gaussian will lead to one half of the S to be elongated or distorted.

Standard error and confidence interval for an average

The standard deviation refers to the variation of observations within individual experimental units, whereas the standard error refers to the random variation of an estimate (made with n replicates) from the mean value that will be obtained as the number of replicates increases. The standard deviation σ is calculated by:

$$\sigma = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2} \tag{1}$$

where $x_i = 1, 2, \dots, n$ are the individual measurements used to calculate the average. A convenient way to calculate the sample standard deviation is using the “STDEV” function in Excel. The standard error is the measure of the experimental error of an estimated statistic (e.g. the mean). For the sample average \bar{x} from n replicate tests, the standard error $\sigma_{\bar{x}}$ is $\frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of the n replicates. The standard error on the mean can be reduced by increasing the number of replicates. *Replication will not reduce the standard deviation but it will reduce the standard error.* In practical terms, this means that our goal is to achieve a standard error small enough to make convincing and useful conclusions. Additionally, in our experience, computing the variance can be problematic from very few replicates. It is mathematically correct that a variance can be computed from just three replicates. However, we have commonly found that three replicates resulted in a somewhat small variance, only to be often greater or much greater once we include the fourth and fifth samples. As a rule-of-thumb we are dubious of variances computed from fewer than five replicates.

The confidence interval indicates the reliability of an estimate made from a given number of replicates. The $(1 - \alpha)100\%$ confidence interval for the average \bar{x} has the form $\bar{x} \pm E$, where E is called the half-length, since a segment of the length of $2E$ centered on \bar{x} , provides the full confidence interval. E is related to α , σ , and n (the number of replicates) by the following equation.

$$E = \frac{Z_{\alpha/2} \sigma}{\sqrt{n}} \tag{2}$$

where $Z_{\alpha/2}$ is a dimensionless number that can be looked up in standard handbooks for various standard distributions (e.g. Berthouex and Brown, 2002). Transposing Eq. (2), the number of replicates that will produce this interval half-length is

$$n = \left(\frac{Z_{\alpha/2} \sigma}{E} \right)^2 \tag{3}$$

This assumes random sampling. It also assumes that n is large enough that the normal distribution can be used to define the confidence interval. To apply Eq. (3), we must specify E , α (or $1 - \alpha$), and σ . Values of $(1 - \alpha)$ that might be used are shown in the top row with corresponding values of Z in the bottom row of Table 1.

When the measurements are assumed to be normally distributed but the number of replicates is small (by small, textbooks suggest less

than 30) and the population standard deviation is unknown, a Student's t -distribution is used (Berthouex and Brown, 2002). To calculate the number of replicates n , the coefficient t_p is used in place of $Z_{\alpha/2}$ shown in Eq. (3). A selection of t -values is listed in Table 2. The t value decreases as n increases, but notice that there is little change once n exceeds 5. An exact solution of the number of replicates for small n (less than 30) requires an iterative solution, but a good approximate is obtained by using a rounded value of $t = 2.1$ or 2.2 , which covers a good working range of $n = 10$ to $n = 25$ ($p = 0.05$). When analyzing data we carry three decimal places in the value of t , but that kind of accuracy is misplaced. The greatest uncertainty lies in the value of the specified σ (refer to Eq. (2)), so we can conveniently round off t to one decimal place. Additional information about confidence interval estimation and experiment sizing can be found in Berthouex and Brown (2002), Spiegel et al. (2008), and Taylor (1997).

Bootstrapping

All the preceding discussion was predicated on the assumption of a Gaussian distribution of underlying population. What if the distribution is not Gaussian? Bootstrapping is a powerful statistical approach that allows estimation of the variability of many properties of the data without making any assumptions about the shape of the original distribution F . Efron (1979) provides an accessible explanation, with examples, of the bootstrap method. The key principle of bootstrapping is to simulate repeated observations from the unknown distribution F , using repeated sampling of the obtained single set of data. Bootstrapping can be implemented by constructing a number of resamples of the observed dataset. Each resample is obtained by random sampling with replacement from the original dataset (Varian, 2005). Increasing the number of resamples can reduce the impact of random sampling errors, but it cannot increase the amount of information existing in the original dataset (Efron and Tibshirani, 1993).

Kolmogorov–Smirnov test

The Kolmogorov–Smirnov (K–S) test quantifies whether two cumulative distribution functions (CDFs) are from the same population. It does so by exploring the maximum distance between the two CDFs. Corder and Foreman (2009) provide a good summary of the K–S test. The null hypothesis of a K–S test poses that the two samples are from the same population, and the research hypothesis poses either that they generally differ, leading to a two-tailed probability estimate, or that they differ in a specific direction, leading to a one-tailed estimate (Wall and Jenkins, 2003). The K–S test can be used to compare a sample distribution and a reference distribution or to compare two sample distributions. We will apply this test to help us explore how many replicates are needed to confirm whether the performance of two stoves is indistinguishable.

The K–S test is a nonparametric statistical test and is only limited by the condition that it must be applied to continuous distributions. Unlike the t -test and other parametric tests, which require assuming Gaussian distribution, continuity is the primary requirement for application of K–S test making it a very useful tool with unknown distributions. Also for small and medium samples, it is more effective to use the K–S test over other nonparametric “goodness-of-fit” tests, such as the chi-square test or the Wilcoxon test. The different research hypotheses of the K–S test also provide directional flexibility which the chi-square test cannot provide (Wall and Jenkins, 2003).

Methods

Laboratory testing

Laboratory tests of the BDS and TSF were performed at the LBNL cookstove testing facility. Concentrations of PM_{2.5} (particulate matter

Table 1
Summary of Z values.

$1 - \alpha = 0.99$ $z = 2.56$	$1 - \alpha = 0.95$ $z = 1.96$	$1 - \alpha = 0.90$ $z = 1.64$
-----------------------------------	-----------------------------------	-----------------------------------

Table 2
Student's t-distribution critical values.

n (Number of replicates)	n – 1 (Degrees of freedom)	$t_{.995}$ (one sided) or	$t_{.975}$ (one sided) or	$t_{.95}$ (one sided) or
		$t_{.99}$ (two sided)	$t_{.95}$ (two sided)	$t_{.90}$ (two sided)
1	–	–	–	–
2	1	63.657	12.706	6.314
3	2	9.925	4.303	2.920
4	3	5.841	3.182	2.353
5	4	4.604	2.776	2.132
6	5	4.032	2.571	2.015
7	6	3.707	2.447	1.943
8	7	3.500	2.365	1.895
9	8	3.355	2.306	1.860
10	9	3.250	2.262	1.833
11	10	3.169	2.228	1.812
12	11	3.106	2.201	1.796
13	12	3.054	2.179	1.782
14	13	3.012	2.160	1.771
15	14	2.977	2.145	1.761
16	15	2.947	2.132	1.753
17	16	2.921	2.120	1.746
18	17	2.898	2.110	1.740
19	18	2.878	2.101	1.734
20	19	2.861	2.093	1.729
21	20	2.845	2.086	1.725
22	21	2.831	2.080	1.721
23	22	2.819	2.074	1.717
24	23	2.807	2.069	1.714
25	24	2.797	2.064	1.711
26	25	2.787	2.060	1.708
27	26	2.779	2.056	1.706
28	27	2.771	2.052	1.703
29	28	2.763	2.048	1.701
30	29	2.756	2.045	1.699

less than or equal to 2.5 μm in diameter), carbon monoxide/carbon dioxide (CO/CO_2), BC, and several other co-pollutants emitted from the BDS and TSF were simultaneously measured. The DustTrak measures the amount of light scattered by particles and relates that to their mass. It is calibrated for a National Institute of Standards and Technology (NIST) certified PM standard composed of soil from Arizona. Since the amount of light scattered by particles is specific to their morphology and chemical composition, in this study a calibration specific to wood smoke was developed, per the manufacturer's recommendation, by comparing $\text{PM}_{2.5}$ concentrations measured with the DustTrak after adjusting for secondary dilution to those measured gravimetrically. However, the DustTrak data are not as reliable and consistent as gravimetric results.

The CO/CO_2 concentrations were measured in a single instrument by nondispersive infrared absorption spectroscopy (NDIR analyzer, CAI 600 series). A cookstove smoke-specific calibration was developed for the BC aethalometer measurements. The results were compared with particle light-absorption coefficients measured with a photoacoustic absorption spectrometer (PAS) and elemental carbon concentrations measured using a thermal-optical analysis method. The moisture content of each piece of fuel wood was measured using a moisture meter (Delmhorst, J-2000). Soft (pine and fir) and hard (oak) woods were used in an equal number of tests with both stove types. Soft wood pieces were saw-cut to approximately 15 cm long with a square cross-section of approximately 4 cm^2 and hard wood pieces were hatchet-cut to a similar size but irregular shape. The variability in the laboratory test results could probably be further reduced by using consistent quality wood with more consistent dimensions.

The BDS and TSF were compared using a modification of the WBT V3.0 protocol. The WBT is intended to provide a method to compare the performance and emissions of different stoves in completing a defined standardized task (Bailis et al., 2007). In our modified protocol, a fire is ignited and maintained by periodic feeding of fuelwood to bring 2.5 L of water in a 2.3 kg metal Darfur pot (without pot lids) to boil and subsequently maintain it on simmer for 15 min, whereupon the

fire is extinguished and the mass of remaining fuelwood is measured. The WBT suggests a default test volume of water of 5 L. We chose to test with 2.5 L of water, because it reflects the actual volume of food stove users prepare at a time. Our previous testing results show no significant difference of time to boil between cold start and hot start for both the BDS and TSF. Therefore, only one high-power phase (cold start) was included in each test. Note that the International Organization for Standardization (ISO) International Workshop Agreement (IWA) metrics average high-power (cold start and hot start) values (<http://www.pciaonline.org/files/ISO-IWA-Cookstoves.pdf>). When three WBT replicate tests are performed, n is equal to 6.

One of the main metrics in our modified WBT test is the time to boil. In an important report by the United States Agency for International Developing (USAID, 2008), authors state, "Fuel-efficient stoves can deliver numerous benefits to end-user households, including fuel and time savings." This underlines what we found in our work in Darfur, time savings are indeed important to the users. Moreover, we learned from our field partners that the most attractive feature of the BDS is that the stove could take their drinking water to boiling in less than 5 min. The refugee women in Darfur IDP camps have named the BDS in Arabic "Kanun Khamsa Dagaig" (i.e., "the 5-min stove"), indicating this as the single most important feature of the BDS from their perspective. Therefore, we believe that "time to boil" is an important testing matrix from the user perspective and consequently, it is important for us to examine for both the BDS and TSF.

Stove testers control the fuel feeding rate that determines the time to boil. Two trained stove testers were employed for all the tests in this study. The average fuel burning rates for the BDS and TSF are 12.2 ± 0.9 g/min and 13.8 ± 1.3 g/min (mean \pm 1SD), respectively. These values indicate that the fire tending skill of the two testers is very consistent. Please note in other areas of the world where fuel is more abundant and inexpensive compared to Darfur, users often sacrifice fuel consumption for time savings. As shown in Fig. 1, the BDS has a small fire box opening to prevent using more fuel wood than necessary. The TSF has no such restriction, so it can achieve a higher fuel burning

rate than the BDS, therefore, the TSF could have a shorter time to boil if fuel consumption is not an issue. The detailed testing methodology and results are given by Kirchstetter et al. (2010).

Data analysis

Stove performance is strongly influenced by the skill of the person tending the stove. Dozens of tests were practiced by trained stove testers on both the TSF and BDS, and these data were discarded before performing the tests to produce the data reported in this paper. This ensured that the variability observed in the test results was not being primarily influenced by increasing the skill of the tester in tending the stove. There were 20 and 21 tests completed for the TSF and BDS for data analysis, respectively. All instrumentations discussed above operated properly during these 41 tests. The statistical analysis was performed using Statistical Analysis System (SAS Institute Inc., version 9) and R (<http://www.r-project.org/>).

Results and discussion

Data overview

The stove performance and emission results of 21 BDS tests and 20 TSF tests are comprehensively presented in Kirchstetter et al. (2010). The moisture content and dry mass of the soft and hard woods were similar to each other and were the same for the TSF and BDS tests. The completion of tests with softwood (10 tests) required about 90% of the time duration and 90% of the wood mass compared to those with hard wood (10 tests).

The data of time to boil and $PM_{2.5}$ emission factor (g/g of fuel consumed) for the TSF and BDS are selected for the statistical analysis in this study. We understand that $PM_{2.5}$ emissions per energy delivered to the cooking pot (g/MJ delivered) is an important metric of cookstove performance, because it is based on the fundamental desired output – cooking energy – that enables valid comparisons between all stoves and fuels (Smith et al., 2000). Also cooking energy tends to have less variation than time to boil, so it might require a smaller number of replicates. However, the data for the mass of water evaporated and the mass of fuel consumed during cold start were not collected when these tests were conducted. Thus, a shortcoming of this study is that it is not possible to calculate the emission factors based on energy delivered to the pot.

The histogram plots of these data are shown in Figs. 2 and 3. The CDF plots for the same data are shown in Figs. 4 and 5. On average, cooking tests with the BDS were completed in 74% of the time for the TSF (30.3 min vs. 41.0 min). There was less variation in time to boil with the BDS, as indicated by a narrower spread in the CDF curves for the BDS compared to the TSF (Fig. 4). The average $PM_{2.5}$ emission factor for the BDS tests was 80% of that for the TSF (3.1 g/kg-wood burned vs. 3.9 g/kg-wood burned). $PM_{2.5}$ shows large test-to-test variability. The distributions of the BDS and TSF $PM_{2.5}$ data overlap substantially, but the question to answer is whether data from the BDS and TST tests show performance data that are different and discernible.

Number of replicate tests to estimate the mean

We now discuss the number of replicate tests needed to estimate the experiment mean within a user-defined level of confidence. For example, suppose the analyst desires to compute the expected boil time of the BDS within a range of plus or minus 2 min. Suppose also that the analyst desires the certainty of that estimate to be 95%. In other words, the analyst is saying, “I would like to know the number of replicate tests needed to compute the average time to boil the BDS within a range of 4 min, and I want to know that range with a confidence of 95%.” Fig. 6 shows the number of replicates needed for three probability levels (0.1, 0.05, and 0.01), which correspond to confidences of 90%, 95%,

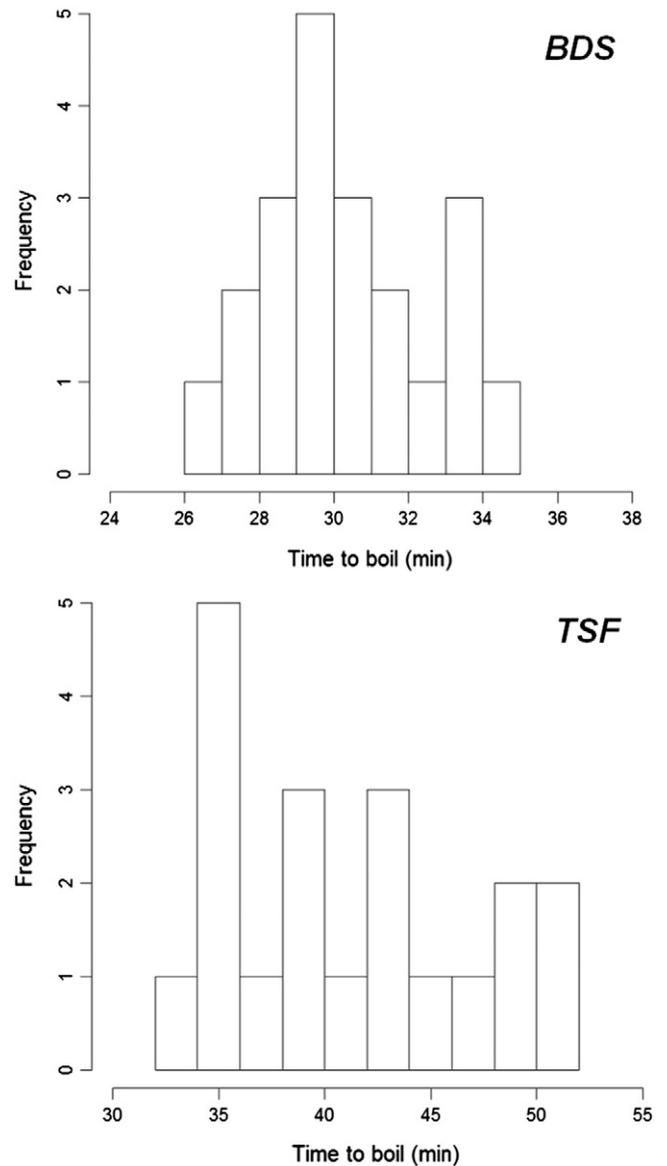


Fig. 2. Histogram of time to boil data for the BDS and the TSF.

and 99%, respectively. We compute the number of replicates using Eq. (3). The x -axis represents the number of replicates ranging from 1 to 25. The y -axis represents the width of the confidence interval about the mean, which is twice the E value in Eq. (2). As can be seen in the figure, the smaller the confidence interval about the mean desired, the larger the number of replicates required.

As the 0.05 probability in Fig. 6 shows, if the width of the confidence interval for the mean time to boil is 4 min at the probability of 0.05, 7 replicates are required. Note that σ for the underlying distribution in Eq. (2) is calculated based on the original 21 replicate tests. If only two replicates are conducted, the width of the confidence interval about the mean is 38 min at the probability of 0.05 (191 min for the probability of 0.01, 19 min for the probability of 0.10). When the number of replicates increases to 5, the width shrinks to 5.3 min at the probability of 0.05 (8.8 min for the probability of 0.01, 4.1 min for the probability of 0.10). The width of the confidence interval about the mean is relatively stable when the number of replicates is greater than 15. A similar trend is observed for the BDS $PM_{2.5}$ emission factor data. The width of the confidence interval about the mean BDS $PM_{2.5}$ emission factor is enormous for $n < 5$, and becomes steady when $n > 10$.

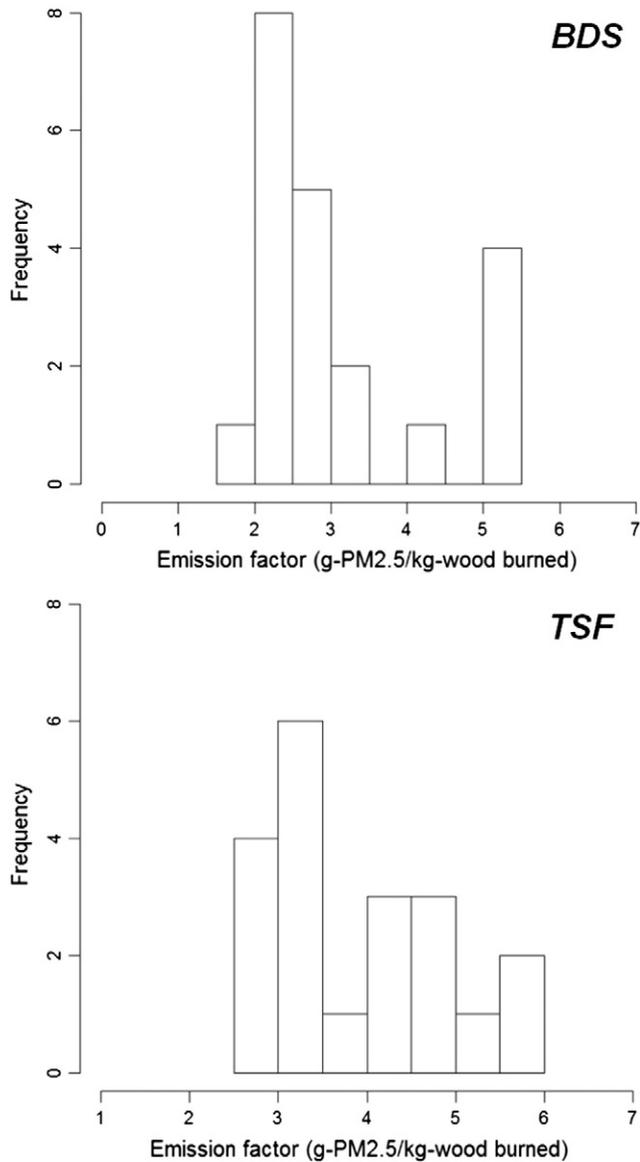


Fig. 3. Histogram of PM_{2.5} emission factor data for the BDS and the TSF.

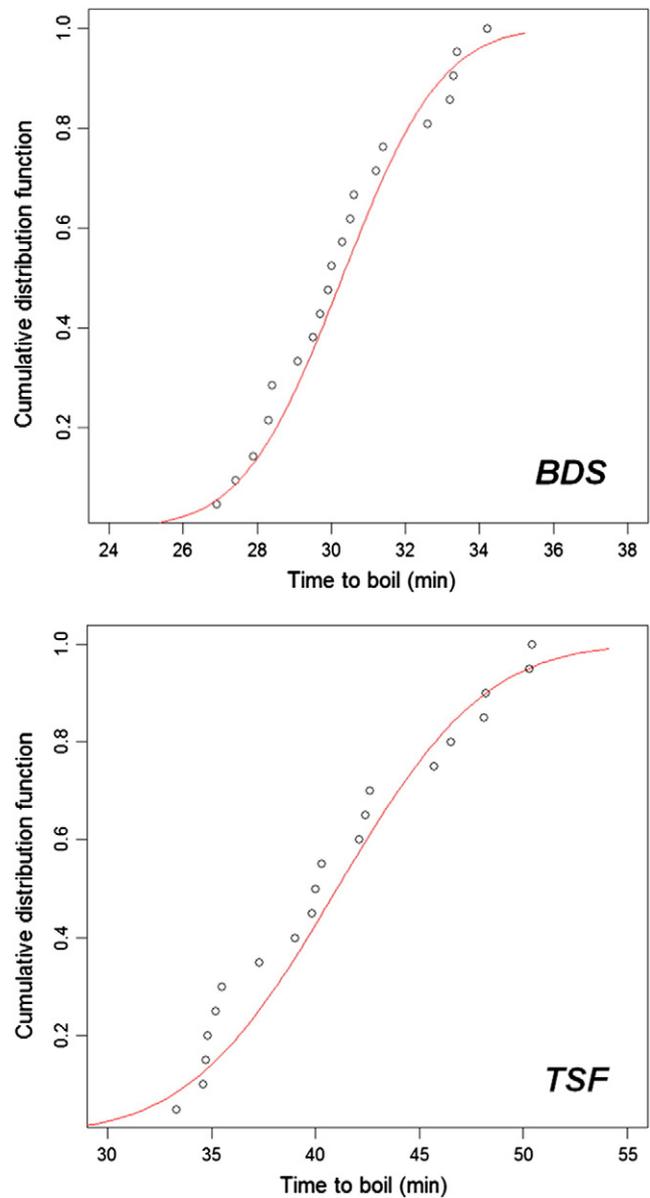


Fig. 4. Cumulative distribution function (CDF) of time to boil data for the BDS and the TSF.

Number of replicate tests to compare two stoves

We now discuss how many replicate tests are needed to confirm whether the performance of two stoves is indistinguishable, within a level of confidence. In essence, we test whether the underlying statistical distribution of the two stoves for the mean boil time or emission factor are the same. Fig. 7 shows the probability as a function of the number of replicates calculated using the K–S test.

On the *x*-axis is the number of replicates. For every replicate number, we generated 50,000 bootstrap samples using the original 21 replicate tests for the BDS and 50,000 bootstrap samples using the original 20 TSF replicate tests. For each pair of samples, we compute the probability (*p* value) that they come from the same distribution. We then compute the ratio, or probability, of the number of pairs that come from the same distribution divided by 50,000 with a confidence of 95%. The *y*-axis shows the resulting probability. When the number of replicates is greater than 6, the probability that the BDS and the TSF time to boil data are from two different distributions is greater than 95%. For the PM_{2.5} emission factor data, 30 replicates are required to ensure that at least there is 95% chance that the BDS and the TSF samples are drawn from two different distributions.

A practical approach to assess the number of replicate tests

The difficulty with estimating the number of replicate tests needed for one particular stove is the lack of prior knowledge about the expected σ of the planned experiments. We knew the σ for the above demonstration because we had already conducted 21 replicates.

In the absence of the σ , the experiment designer must speculate on the variance. We recommend reviewing the literature of similar stoves to pose a notional variance. In the absence of such data, then the designer must use any other information as a starting point, such as the variance computed from the TSF and BDS replicates reported here. Note that the σ values for the BDS and TSF for time-to-boil are 2.1 min and 5.6 min, respectively, and for emission factor for PM_{2.5} they are 1.2 g/kg-wood and 1.0 g/kg-wood, respectively. The σ values calculated for all measured variables are summarized in Table 3.

Note the wide difference in the three-stone-fire and the Berkeley-Darfur Stove. The former is a setup with three stones with irregular shape, and the dimensions and shape and spacing of the stones can vary from test to test. The results reported in the literature have generally been with consistent dimensions, shape, and spacing of the stones

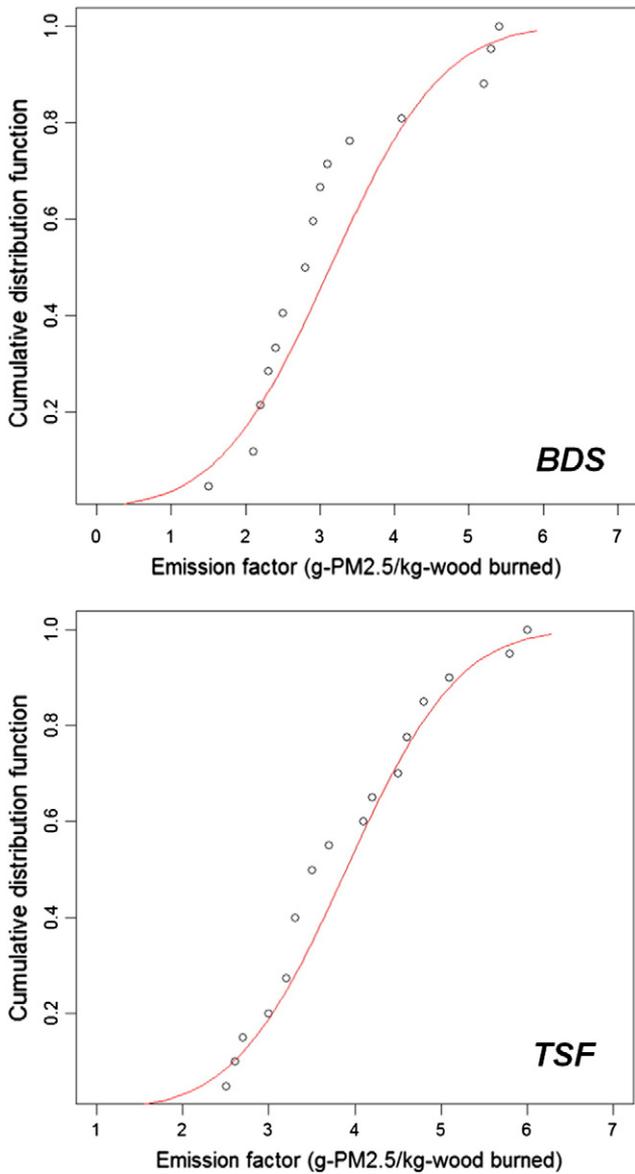


Fig. 5. Cumulative distribution function (CDF) of PM_{2.5} emission factor data for the BDS and the TSF.

(or bricks). This factor may have caused more variation in our TSF results compared to literature values. In contrast, the BDS is precisely an engineered metal stove of fixed dimensions. The remarkable point is that while there is a difference in the σ values for the time to boil, there is not a large difference in the σ values for emission factors of the TSF and BDS despite the significant design difference. So, we recommend starting conservatively, with the notional σ similar to the value for the TSF. If the designer's stove or testing conditions are likely to show less variation, then perhaps start with a notional variance that is 10% less. Conversely, our BDS experiment was conducted in a controlled laboratory setting. If the designer expects greater variation in the experiment (say, owing to variable field conditions), then begin with a notional variance of 10, 50, or even 100% greater. For example, when testing fan-assisted stoves, which burn engineered wood pellets, we might start with a notional variance that is 10% smaller than was found here owing to the uniform nature of the engineered fuel pellets. On the other hand, when testing an open fire (the fire is open completely to the ambient environment), we might begin with a notional variance that is twice that of the TSF (three stones or bricks are places

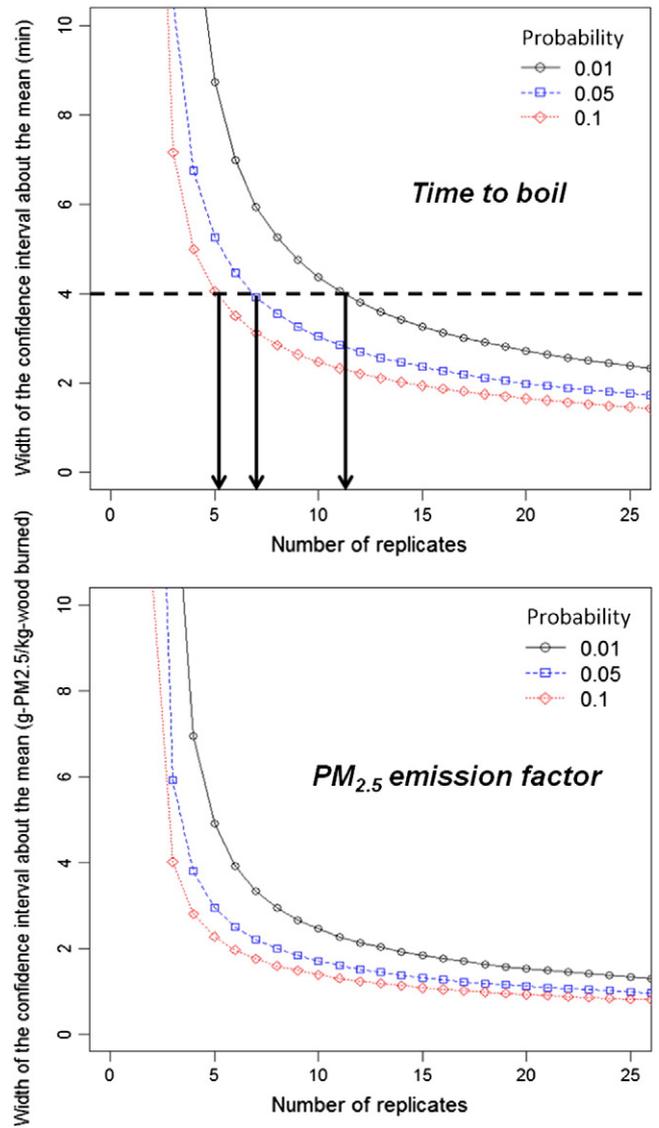


Fig. 6. The width of the confidence interval about the mean as a function of the number of replicate tests at three probability levels (0.1, 0.05, and 0.01) for the BDS time to boil and PM_{2.5} emission factor data. For example, if the width of the confidence interval for the mean time to boil is 4 min at probability levels of 0.1, 0.05, and 0.01, 5, 7 and 12 replicates are required, respectively, as indicated by the black horizontal dash line and the black vertical arrows.

between the fire and the ambient to provide the support to the cooking pot and some insulation of the fire) laboratory tests reported here.

With a notional variance, the designer would proceed with Eq. (3) to compute the number of replicates needed based on the desired size of confidence interval (E) and the level of confidence desired (α). Remember also that test conditions change, instruments malfunction, and interpretation of tests differs (such as the precise time of onset of hard boil, or the precise duration that water simmer, can be questionable). These factors should also be considered beyond what is computed from the above statistics to arrive at the number of replicate tests. More replicate tests should be planned than required by the statistical estimation to compensate for these unusual occurrences. This also increases the margin of safety in case the variability in the underlying distribution, represented by the standard deviation (σ) in Eq. (2), is larger than anticipated. A conservative margin of 100% is recommended based on our abundant stove laboratory testing experience.

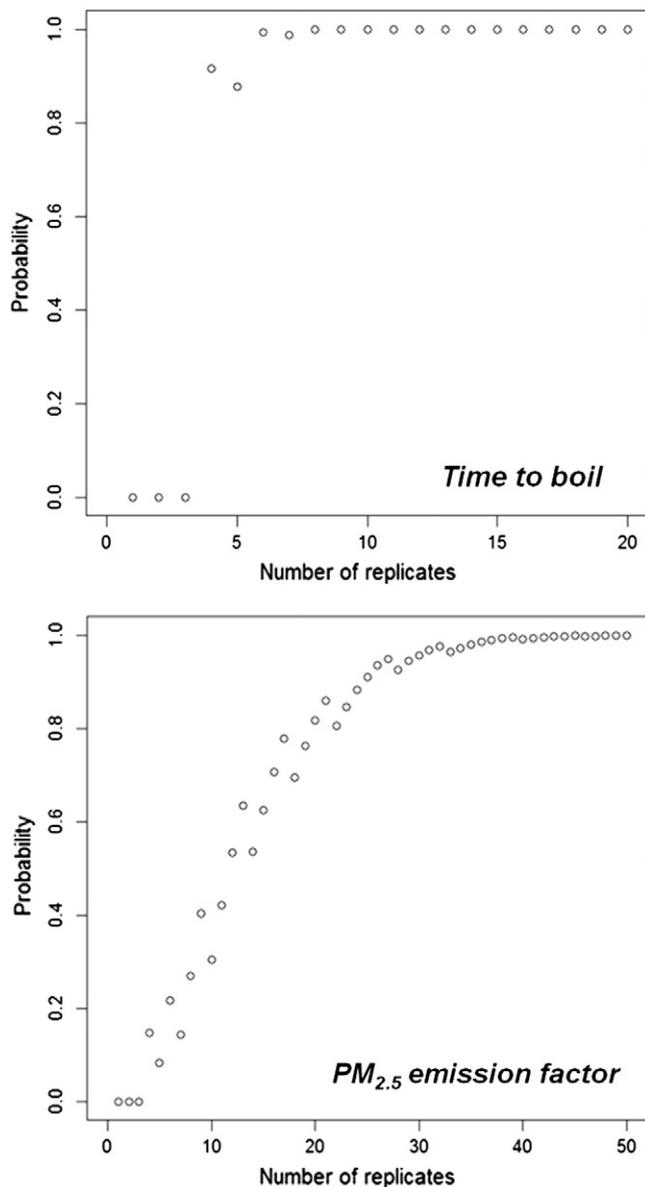


Fig. 7. Kolmogorov–Smirnov test result showing the probability of the BDS and the TSF bootstrap samples are drawn from two different distributions as a function of the number of replicate tests for the time to boil and $PM_{2.5}$ emission factor data.

With the number of replicate tests determined, the experimenters conduct the tests. With these data now in hand, the experimenters can calculate the actual, observed variance computed from the experiment. This value should be used to estimate the analysis results. One might need to conduct additional replicate tests to achieve the desired confidence interval and desired level of confidence in the mean estimation from the test results.

Table 3

Summary of the standard deviation values (σ) of all measured variables for the BDS ($n = 21$) and the TSF ($n = 20$).

	TSF	BDS
Time to boil (minute)	5.6	2.1
Dry wood burned (g)	75.4	33.6
CO emission factor (g/kg-wood)	6.8	5.8
$PM_{2.5}$ emission factor (g/kg-wood)	1.0	1.2
BC emission factor (g/kg-wood)	0.3	0.5

Conclusions

Our results show moderate inherent variability (coefficient of variation up to 0.4) among the TSF's and BDS' time to boil and $PM_{2.5}$ emission measurements based on the modified WBT protocol. We demonstrate using these data as examples that some stove laboratory testing results could be misleading if only a small number of replicate tests were conducted. However, there are costs associated with increasing the number of replicates. The average value of any measured parameter should be always reported together with the number of replicates conducted and the uncertainty (e.g. standard deviation or confidence interval). Cautions must be exercised in the interpretation of results based on only a few replicates. We then describe a practical approach to calculate the number of replicate tests needed to obtain useful data from previously untested stoves.

The implications of these results include the following: (1) In the stove design and laboratory testing phase, researchers need to conduct a relatively large number of replicate tests to ensure with some confidence that the improvements of stove performance and emission levels are truly achieved. (2) In the stove field testing phase, even more tests are required because of the less controlled testing environment and the associated larger inherent variability within the replicates. (3) In the stove dissemination and adoption phase, decision makers and policy analysts should take into consideration the variability and confidence intervals of the laboratory and field testing results prior to any decisions.

Acknowledgments

This work was performed at the Lawrence Berkeley National Laboratory, operated by the University of California, under DOE Contract DE-AC02-05CH11231. We gratefully acknowledge the partial support for this work from LBNL's LDRD funds, and DOE's Biomass Energy Technologies Office. The data used in this work were collected during research supported with grant number 500-99-013 from the California Energy Commission (CEC). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the CEC. Kathleen M. Lask was supported by the National Defense Science and Engineering Graduate (NDSEG) Fellowship and the National Science Foundation Graduate Research Fellowship. The authors gratefully acknowledge Douglas Sullivan, Jessica Granderson, Chelsea Preble, Odelle Hadley and Philip Price of Lawrence Berkeley National Laboratory for their support of this project, as well as the many students, interns, and researchers who, before us, contributed to the development of the Berkeley–Darfur stove. The authors are very grateful to have the paper manuscript reviewed by the journal reviewers. The paper quality is substantially improved owing to their careful review.

Disclaimer

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, or the Regents of the University of California.

References

- Bailis R, Berrueta V, Chengappa C, Dutta K, Edwards R, Masera O, et al. Performance testing for monitoring improved biomass stove interventions: experiences of the household energy and health project. *Energy Sustain Dev* 2007;11(2):57–70.
- Berthouex PM, Brown LC. *Statistics for environmental engineers*. 2nd ed. Lewis Publishers; 2002.
- Corder G, Foreman D. *Nonparametric statistics for non-statisticians: a step-by-step approach*. Wiley; 2009.
- DOE (Department of Energy). Biomass cookstoves technical meeting: summary report, Alexandria, VA. http://www1.eere.energy.gov/biomass/pdfs/cookstove_meeting_summary.pdf, 2011. [Accessed February 4, 2013].
- Efron B. Bootstrapping methods: another look at the jackknife. *Ann Stat* 1979;7(1):1–26.
- Efron B, Tibshirani R. *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall/CRC 0-412-04231-2; 1993.
- Ellison S, Barwick V, Duguid Farrant T. *Practical statistics for the analytical scientist: a bench guide*. 2nd ed. Royal Society of Chemistry; 2009.
- Granderson J, Sandhu JS, Vasquez D, Ramirez E, Smith KR. Fuel use and design analysis of improved woodburning cookstoves in the Guatemalan Highlands. *Biomass Bioenergy* 2009;33:306–15.
- Hadley OL, Corrigan CE, Kirchstetter TW, Cliff SS, Ramanathan V. Measured black carbon deposition on the Sierra Nevada snow pack and implication for snow pack retreat. *Atmos Chem Phys* 2010;10:7505–13.
- IEA (International Energy Agency). *Energy and development. World energy outlook 2004*. Paris: IEA Publications; 2004.
- Jetter J, Kariher P. Solid-fuel household cook stoves: characterization of performance and emissions. *Biomass Bioenergy* 2009;33:294–305.
- Jetter J, Zhao Y, Smith KR, Khan B, Yelverton T, DeCarlo P, et al. Pollutant emissions and energy efficiency under controlled conditions for household biomass cookstoves and implications for metrics useful in setting international test standards. *Environ Sci Technol* 2012;46:10827–34.
- Kirchstetter T, Preble C, Hadley O, Gadgil A. Quantification of black carbon and other pollutant emissions from a traditional and an improved cookstove. Lawrence Berkeley National Laboratory (LBNL) report, number: LBNL-6062E; 2010 [Available: http://gadgillab.berkeley.edu/wp-content/uploads/2010/11/TWK_improved-cookstove_f_13-6-10.pdf. Accessed December 5, 2013].
- Lim SS, Vos T, Flaxman AD, Danaei G, Shibuya K, Adair-Rohani H, et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 2012;380:2224–60.
- MacCarty N, Ogle D, Still D, Bond T, Roden C. A laboratory comparison of the global warming impact of five major types of biomass cooking stoves. *Energy Sustain Dev* 2008;12(2):56–65.
- MacCarty N, Still D, Ogle D. Fuel use and emissions performance of fifty cooking stoves in the laboratory and related benchmarks of performance. *Energy Sustain Dev* 2010;14:161–71.
- Ramanathan V, Carmichael G. Global and regional climate changes due to black carbon. *Nat Geosci* 2008;1:221–7.
- Roden CA, Bond TC, Conway S, Benjamin A, Pinel O, MacCarty N, et al. Laboratory and field investigations of particulate and carbon monoxide emissions from traditional and improved cookstoves. *Atmos Environ* 2009;43:1170–81.
- Smith KR, Uma R, Kishore VVN, Lata K, Joshi V, Zhang J, et al. Greenhouse gases from small-scale combustion devices in developing countries; EPA/600/R-00/052. Washington, DC: U.S. Environmental Protection Agency; 2000.
- Smith KR, Mehta S, Maeusezahl-Feuz M. Indoor smoke from household solid fuels. In: Ezzati M, Rodgers AD, Lopez AD, Murray CLJ, editors. *Comparative quantification of health risks: global and regional burden of disease due to selected major risk factors*. Geneva, Switzerland: World Health Organization; 2004.
- Smith KR, Dutta K, Chengappa C, Gusain PPS, Berrueta V, Masera O, et al. Monitoring and evaluation of improved biomass cookstove programs for indoor air quality and stove performance: conclusions from the household energy and health project. *Energy Sustain Dev* 2007;11(2):5–18.
- Spiegel MR, Lipschutz S, Liu J. *Mathematical handbook of formulas and tables*. 3rd ed. McGraw-Hill; 2008.
- Taylor JR. *An introduction to error analysis*. 2nd ed. University Science Books/University Science Books; 1997.
- USAID (United States Agency for International Development). Fuel-efficiency stove programs in IDP settings – summary evaluation report, Darfur, Sudan. Available: http://pdf.usaid.gov/pdf_docs/PDADM099.pdf, 2008. [December, accessed January 6, 2014].
- Varian H. Bootstrap tutorial. *Math J* 2005;9:768–75.
- Wall JV, Jenkins CR. *Practical statistics for astronomers*. Cambridge University Press. Cambridge University Press; 2003.
- Weinhold B. Indoor PM pollution and elevated blood pressure: cardiovascular impact of indoor biomass burning. *Environ Health Perspect* 2011;119(10):A442.
- World Bank. Household cookstoves, environment, health and climate change: a new look at an old problem. Washington, DC: The World Bank; 2011 [Available: <http://climatechange.worldbank.org/sites/default/files/documents/Household%20Cookstoves-web.pdf>. Accessed December 5, 2013].