Seiss, Mark, Ralph Hall, and Eric Vance. 2014. “The Importance of Cleaning Data During Fieldwork: Evidence from Mozambique.” *Survey Practice* 7 (4). https://doi.org/10.29115/SP-2014-0017.

ARTICLES

The Importance of Cleaning Data During Fieldwork: Evidence from Mozambique

Mark Seiss[1], Ralph Hall[1], Eric Vance[1]

[1] Virginia Tech

Keywords: sanitation, monitoring and evaluation, impact evaluation, computer assisted personal interviewing (capi), water, household survey, on-the-ground statistician, data editing, africa

https://doi.org/10.29115/SP-2014-0017

Survey Practice

Vol. 7, Issue 4, 2014

In many small-scale surveys with limited resources, data editing is usually conducted by a statistician after data collection has concluded. Including the statistician and the data editing process in the data collection phase of the survey has many benefits. This paper describes a procedure for survey implementation of small-scale surveys in which the statistician identifies and edits the data as it is collected. We implemented this procedure during a household survey conducted in Maputo, the capital of Mozambique, and detailed data on the editing process was recorded. This article analyzes this data to gain insight into the effects on the collected data. The results of the analysis indicate that the edited data may be of higher quality than data without edits. We also identify areas of improvement in the procedure for future household surveys.

Introduction

The Stanford program on Water, Health, and Development funded by the Woods Institute for the Environment[1] has established a research program on non-network water and sanitation in developing countries. The Maputo project in Mozambique, Africa, was one of three projects initiated under this program. In Maputo, the objective was to evaluate the impact of new regulations that legalize the resale of water by households with a water network connection to the population without access to this network. Prior to legalization, many residents of Maputo with access to the water network illegally resold water to their neighbors. The change in policy went into effect in September 2010, which allowed a unique opportunity to conduct a baseline and follow-up study of households affected by the legislation. This paper discusses data collection in Maputo during the follow-up stage conducted by the faculty and graduates in the Laboratory for Interdisciplinary Statistical Analysis (LISA)[2] and the Urban Affairs and Planning (UAP) program at Virginia Tech in collaboration with colleagues at Stanford University.

Advances in portable computing technology have facilitated the integration of data collection and data editing. Computer-assisted personal interviewing (CAPI) is the term given to the use of computing technology for data collection during personal interviews. De Waal, Pannekoek, and Scholtus (2011) discuss one of the main advantages of CAPI as the capability of

[1] http://woods.stanford.edu/research/centers-programs/water-health-development/

[2] http://www.lisa.stat.vt.edu/

researchers to start the data editing process during data collection by informing the surveyor of a possible error at the time the data value was recorded. The surveyor may confirm the data point was correctly entered or correct an erroneously collected data point.

For many small-scale surveys, the survey software that implements real-time editing and the necessary technology to support it may not be available due to limited resources and other constraints. The CAPI instrument used in the household survey described in this paper allowed for the almost instantaneous access to collected data, but was not capable of flagging possible errors during interviews. We offer an alternative approach to data collection for these small-scale surveys that offers many of the benefits of the real-time editing procedures in spite of these limitations. The proposed procedure requires statisticians to be involved with the cleaning and editing of data during the data collection process by reviewing the collected data and flagging any suspicious values. The small-scale of the survey mitigates concerns of a time-consuming data editing process due to the small number of surveys requiring review on a daily basis. In the proposed procedures, the statistician only edits potential errors after consulting the surveyor who collected the data point. This restriction minimizes the risk of "over editing" (De Waal, Pannekoek, and Scholtus 2011). In many small-scale surveys, the surveyor is able to recall the correct data value, removing subjectivity from the editing process and the risk of "creative editing" (De Waal, Pannekoek, and Scholtus 2011).

Data Editing Methodology

The Maputo project selected eight peri-urban neighborhoods around Maputo in which to conduct interviews. Six of the neighborhoods, referred to as old neighborhoods, were the same neighborhoods interviewed during the baseline survey. Prior to the follow-up survey, the project further funded data collection in two more neighborhoods, referred to as new neighborhoods, which were not part of the baseline survey. In total, 1,864 households were interviewed; 1,369 households were in the old neighborhoods and 495 in the new neighborhoods.

For the Maputo follow-up survey, five graduate students from Stanford University and Virginia Tech and 23 household surveyors from Maputo implemented the survey. Three Stanford students were on-the-ground survey logistics coordinators, in charge of the day-to-day activities and supervising the surveyors. Two Virginia Tech students were statisticians providing support from the United States to the three on-the-ground survey logistics coordinators. The following sections provide a general description of the duties of the team members. Figure 1 describes the flow of data and the duties performed by the survey team members over a three day period. Since it takes 3 days to process the data collected from a given day, the review of the data from one day overlaps with the data review from the following two days.

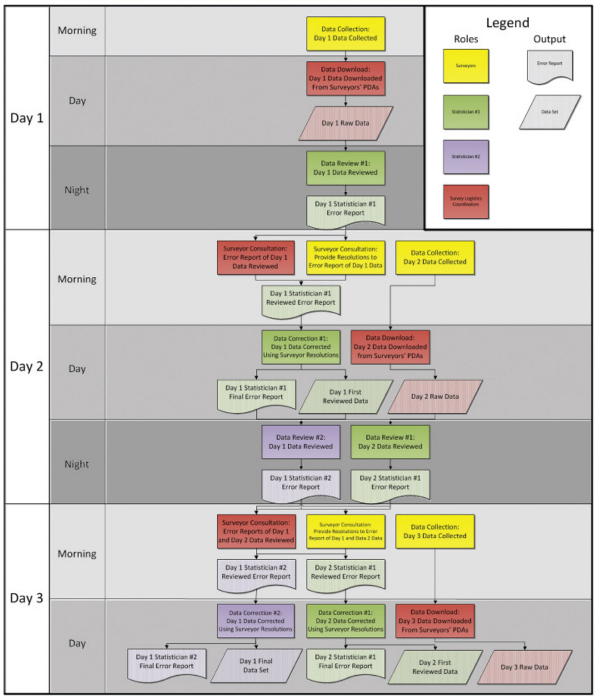

Day 1
Morning
Day
Night
Data Collection: Day 1 Data Collected
Data Download: Day 1 Data Downloaded From Surveyors' PDAs
Day 1 Raw Data
Data Review #1: Day 1 Data Reviewed
Day 1 Statistician #1 Error Report
Legend
Roles
Output
Surveyors
Statistician #1
Statistician #2
Survey Logistics Coordinators
Error Report
Data Set
Day 2
Morning
Day
Night
Surveyor Consultation: Error Report of Day 1 Data Reviewed
Surveyor Consultation: Provide Resolutions to Error Report of Day 1 Data
Data Collection: Day 2 Data Collected
Day 1 Statistician #1 Reviewed Error Report
Data Correction #1: Day 1 Data Corrected Using Surveyor Resolutions
Data Download: Day 2 Data Downloaded from Surveyors' PDAs
Day 1 Statistician #1 Final Error Report
Day 1 First Reviewed Data
Day 2 Raw Data
Data Review #2: Day 1 Data Reviewed
Data Review #1: Day 2 Data Reviewed
Day 1 Statistician #2 Error Report
Day 2 Statistician #1 Error Report
Day 3
Morning
Day
Surveyor Consultation: Error Reports of Day 1 and Day 2 Data Reviewed
Surveyor Consultation: Provide Resolutions to Error Report of Day 1 and Data 2 Data
Data Collection: Day 3 Data Collected
Day 1 Statistician #2 Reviewed Error Report
Day 2 Statistician #1 Reviewed Error Report
Data Correction #2: Day 1 Data Corrected Using Surveyor Resolutions
Data Correction #1: Day 2 Data Corrected Using Surveyor Resolutions
Data Download: Day 3 Data Downloaded From Surveyors' PDAs
Day 1 Statistician #2 Final Error Report
Day 1 Final Data Set
Day 2 Statistician #1 Final Error Report
Day 2 First Reviewed Data
Day 3 Raw Data

Figure 1 Data cleaning process.

Surveyors

The 23 surveyors traveled from house to house within the neighborhoods collecting data using The Survey System (TSS)[3] software on Hewlett-Packard (Palo Alto, CA, USA) iPAQ personal digital assistant (PDA) devices, working part-time as their schedules allowed. Initially, the survey logistics coordinators trained 13 surveyors on the survey contents. As the survey progressed in the six old neighborhoods from the baseline study, surveyors worked more and more sporadically as their schedules became more demanding. To ensure the two new neighborhoods were surveyed within the project timeframe, the surveyor logistics coordinators trained ten more surveyors. While the new surveyors worked only in the new neighborhoods, some of the original surveyors also worked in the new neighborhoods once the old neighborhoods were completed.

Survey Logistics Coordinators

The survey logistics coordinators consulted the surveyors on the errors found by the statisticians. Due to the short turnaround period, the surveyors were able to resolve errors in most cases, either by memory or hand written notes. The other duties performed by these coordinators were as follows: designing the questionnaire, training surveyors, planning the schedule for the survey, transporting surveyors to and from interview sites, and providing on-the-

[3] http://www.surveysystem.com/

ground support to the surveyors while in the field. Two survey logistics coordinators managed all field activities, while the third coordinator was a global positioning system (GPS) specialist supporting logistics.

Statisticians

The two statisticians at Virginia Tech performed independent reviews of the collected data on a daily basis. The second statistician acted as a second level of review, performing another thorough review of the data independent of the first statistician. The statisticians conducted all data reviews in the United States performing the following tasks.

1. Check the data for consistency. The data consistency check focused on critical variables and checked the structure of the data. First, several different questions were used to obtain the same data point. The statistician checked these values against one another to ensure consistency. Any inconsistencies in the values collected for the same data point lead to questions about the validity of either response. Additionally, these inconsistencies may propagate into the final summary tables, undermining the quality of the data. Second, the answers to certain questions activated later sections of the survey. The statistician checked consistency between answers to the setup questions and the sections activated. Inconsistencies between the answers to setup questions and activated sections indicate that there was an error in the survey logic.

2. Flag suspicious data entries. The PDAs contained small screens that required styluses to input the data. Occasionally, the surveyor erred in recording the data. Left unchecked, this potentially incorrect value would alter the reported values in the final summary tables for the community and other tabulations that include this data point.

3. Flag data omissions. There were cases in which the respondent did not know the answer to the question or refused to answer. The survey logistics coordinators instructed the surveyors to record these values as some variation of "9999", but sometimes the surveyors left the field blank. There were also cases in which a respondent answered zero for a question, but the surveyor left the field blank. It is impossible for analysts to distinguish between these two situations. When analyzing the data, treating these values as all zero may introduce a bias into any statistics calculated. Treating these values as all missing may reduce the sample size. As part of the data review, the statistician flagged any fields incorrectly left empty to determine whether the value should be missing or zero.

4. Tabulate important indices. In addition to the value-by-value review of the data, the statistician tabulated important indices to provide an early indication of the results and ensured the surveyors as a whole

were correctly collecting the data.[4]

In addition to these tasks, the statisticians periodically generated tabulations of important indices for each surveyor to determine if there were any systematic differences between the data collected by a given surveyor and the other surveyors. The following describes reasons why a systematic difference may be found between surveyors in a CAPI study:

1. Surveyor misunderstanding. Despite the extensive training given to the surveyors prior to survey implementation, the surveyor may not understand exactly what data should be collected for a given question.
2. Interview miscommunication. Miscommunication between the surveyor and the respondent may result in the collection of incorrect data. Using the tabulations provided by the statisticians, the survey logistics coordinators, using their knowledge of the water sources used the neighborhood, identified households where this miscommunication may have occurred and resolved the potential error with the surveyor.
3. Intentional survey error. Due to the logic of the survey, some answers entered for particular questions may cause the PDA to skip one or many questions in the survey. Tabulations by the statisticians assisted the survey logistics coordinators in identifying surveyors that may be falsifying data to intentionally shorten the length of the interview.

These tabulations allowed the survey logistics coordinators to either retrain the surveyors if the errors were unintentional or discipline the surveyors if the errors were intentional.

Analysis

Of the 1,864 households, 900 (48 percent) contained at least one data value that required editing by the statistician following the review of an error report with the surveyor. Between the two groups of surveyors, the second surveyor group committed a larger number of errors per survey than the first surveyor group. This discrepancy may be attributed to limitations of the second survey group and their training. There were fewer people available to train the surveyors in the second group due to ongoing management of the first group of surveyors in the old neighborhoods. Half of the first group conducted a similar survey in 2010, while all surveyors in the second group were new to this

[4] In a previous impact evaluation, the research team shared summary statistics generated by the statistician in the field with all research partners via a weekly update. This action proved extremely valuable, since it provided some transparency to the research and helped the research team gain the trust of those entities under evaluation. These entities, in turn, were more forthcoming with information about the project.

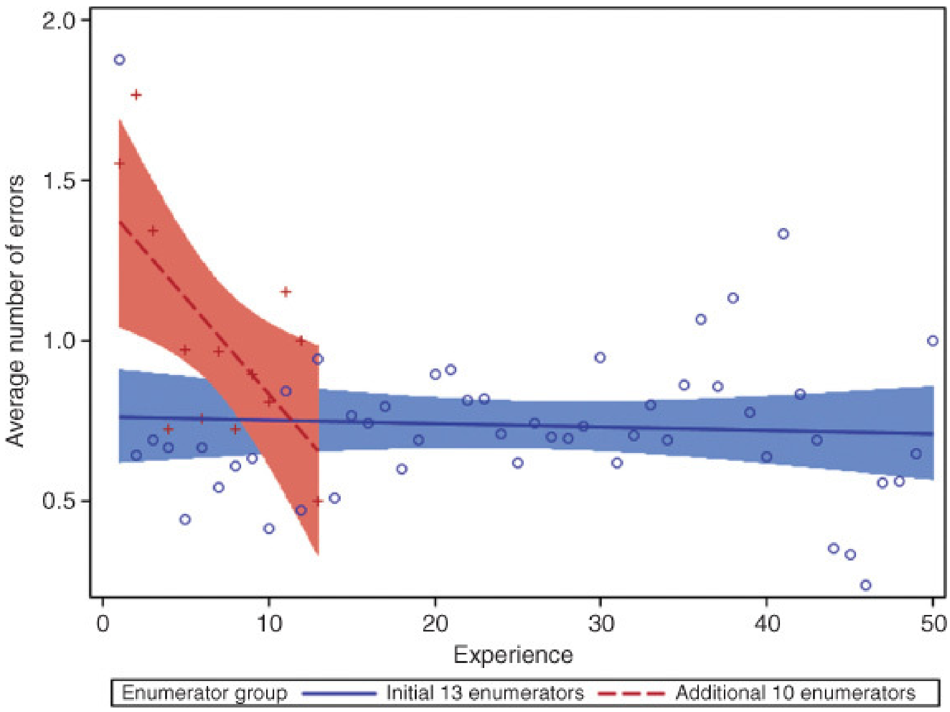


Figure 2 Relationship between experience and errors committed.

type of survey. While the second group contained experienced surveyors, they proved difficult to train on a survey that was different from the surveys they had previously administered.

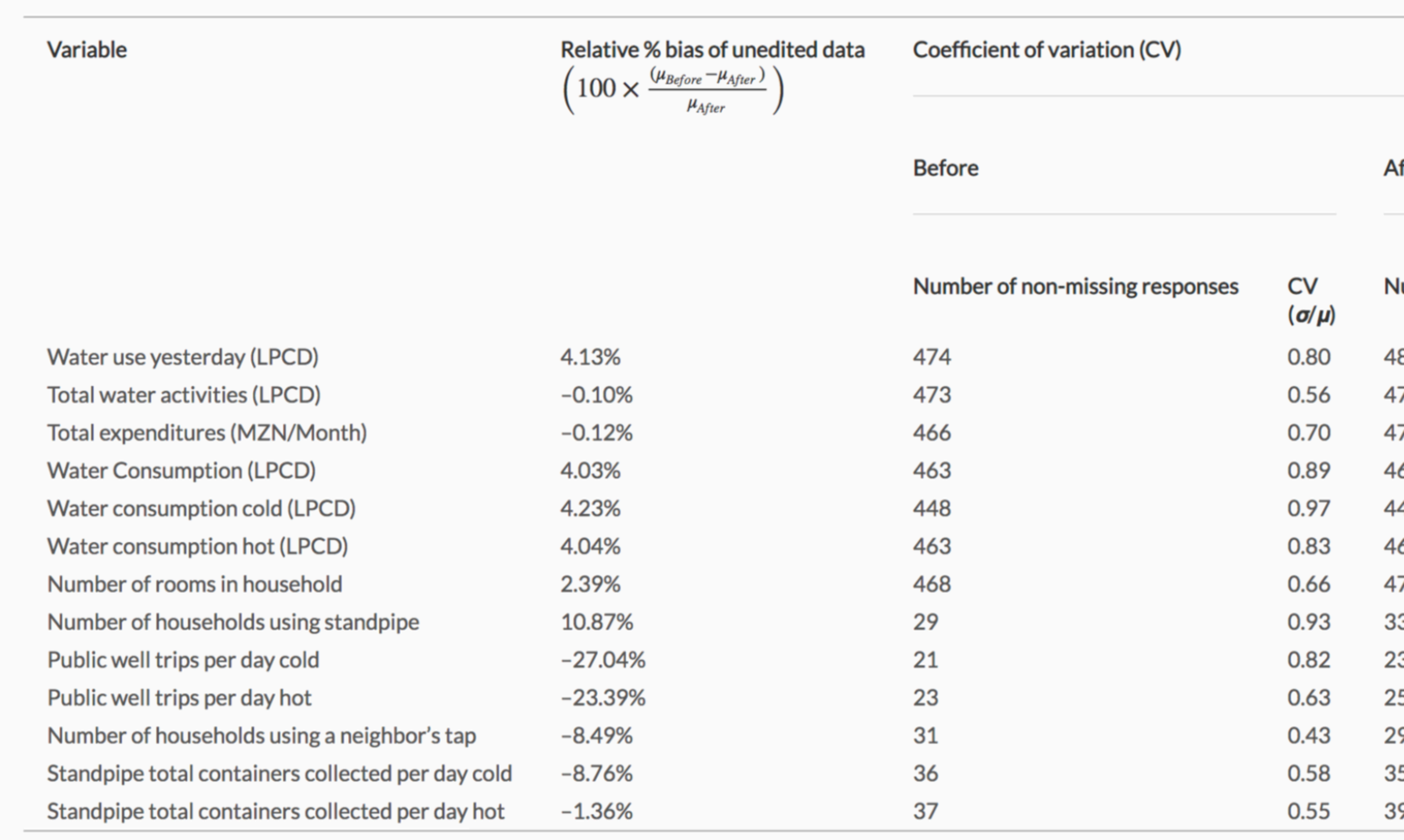
Figure 2 provides a scatterplot of the relationship between the amount of experience a surveyor had and the number of errors he or she committed. Each data point provides the average number of errors committed by surveyors in the listed survey group and the amount of experience (in days worked) the surveyor had at the time of the survey. The scatterplot also fits a regression line to the data points for each survey group, along with 95 percent confidence limits around the regression line. The regression line for the original thirteen surveyors is flat, indicating there was no relationship between the amount of experience the original surveyors had at the time of the interview and the number of errors they committed. The regression line for the additional ten surveyors shows a negative slope, indicating a decrease in the number of errors a surveyor committed as he or she gathered more experience implementing the survey. These two findings suggest that the additional ten surveyors benefitted more from the review process than the original thirteen surveyors.

Tables 1 and 2 provide a comparison of the relative bias of the mean estimates, the number of nonmissing responses, and the coefficient of variation of important indices in the Maputo data before and after data editing. The tables show that the data cleaning procedures generally had a nontrivial effect, particularly for data collected in the new neighborhoods. The magnitudes of most relative biases in the new neighborhoods are greater than 4 percent, reaching as high as 27 percent. The coefficient of variation of these variables

Table 1 Variable summaries from old neighborhoods.

Variable	Relative % bias of unedited data $\left(100 \times \frac{(\mu_{Before}-\mu_{After})}{\mu_{After}}\right)$	Coefficient of variation (CV)				
		Before		After		Relative % difference $\left(100 \times \frac{(CV_{Before}-CV_{After})}{CV_{After}}\right)$
		Number of non-missing responses	CV (σ/μ)	Number of non-missing responses	CV (σ/μ)	
Water use yesterday (LPCD)	1.47%	1365	0.63	1369	0.57	10.53%
Total water activities (LPCD)	−0.36%	1363	0.53	1365	0.52	1.92%
Total expenditures (MZN/Month)	−0.05%	1348	0.66	1360	0.67	−1.49%
Water consumption (LPCD)	1.13%	1322	0.71	1331	0.68	4.41%
Water consumption cold (LPCD)	1.11%	1287	0.78	1296	0.73	6.85%
Water consumption hot (LPCD)	1.14%	1322	0.69	1331	0.63	9.52%
Number of rooms in household	0.23%	1354	0.37	1366	0.37	0.00%
Number of households using standpipe	0.23%	180	1.40	181	1.40	0.00%
Public well trips per day cold	−6.63%	34	0.74	37	0.72	2.78%
Public well trips per day hot	−8.15%	35	0.69	38	0.63	9.52%
Number of households using a neighbor's tap	2.54%	267	0.50	265	0.42	19.05%
Standpipe total containers collected per day cold	4.85%	167	0.76	171	0.52	46.15%
Standpipe total containers collected per day hot	0.95%	171	0.51	175	0.52	−1.92%

generally decreased. Note that only observed values were used for these calculations and that “Hot” and “Cold” refer to the months in which the water was collected.



Variable	Relative % bias of unedited data $\left(100 \times \frac{(\mu_{Before} - \mu_{After})}{\mu_{After}}\right)$	Coefficient of variation (CV)		
		Before		Af
		Number of non-missing responses	CV (σ/μ)	Nu
Water use yesterday (LPCD)	4.13%	474	0.80	48
Total water activities (LPCD)	−0.10%	473	0.56	47
Total expenditures (MZN/Month)	−0.12%	466	0.70	47
Water Consumption (LPCD)	4.03%	463	0.89	46
Water consumption cold (LPCD)	4.23%	448	0.97	44
Water consumption hot (LPCD)	4.04%	463	0.83	46
Number of rooms in household	2.39%	468	0.66	47
Number of households using standpipe	10.87%	29	0.93	33
Public well trips per day cold	−27.04%	21	0.82	23
Public well trips per day hot	−23.39%	23	0.63	25
Number of households using a neighbor's tap	−8.49%	31	0.43	29
Standpipe total containers collected per day cold	−8.76%	36	0.58	35
Standpipe total containers collected per day hot	−1.36%	37	0.55	39

Table 2 Variable summaries from new neighborhoods.

Discussion

The data cleaning procedure outlined in this article provides an alternative to fully automated data editing for small-scale surveys in which resources are limited. In cases such as the Maputo survey, time constraints and resources made it impossible to implement a fully automated data editing procedure. Given these limitations, our proposed data cleaning procedures provide many of the benefits associated with fully automated data editing.

The communication between team members associated with the proposed procedures provides greater ability to identify and correct problems with the survey. The statisticians identified potential errors in data collection within a time frame that enabled the correct data point to be obtained or the collected data point to be confirmed. For the majority of errors, surveyors recalled the interview and provided the intended response. In some cases, the recorded value was the intended response despite being classified as an outlier by the statisticians. With traditional data cleaning after the data entry process is complete, consulting the surveyors would either not be possible since the surveyors may not be available or not productive since the surveyors would be unable to recollect the intended response. The only recourse in these situations would be to either leave the suspicious data points in the data set or to delete them. In the first scenario, the data entry errors would remain. In the second scenario, legitimate data points would be deleted. Both scenarios potentially introduce bias into the mean and standard error estimates, as confirmed in Tables 1 and 2.

We discuss the other benefits of the data cleaning procedures below.

1. Improved survey management. With statisticians as part of the team, the survey logistics coordinators could concentrate on implementing the survey, while the statisticians provided editing of all incoming data.
2. Real-time monitoring of surveyors and survey results. As a consequence of more thoroughly reviewed data, tabulations can be made to evaluate the performance of the surveyors as a whole and each individual surveyor. If it is determined that certain indices are suspicious for all surveyors, the survey logistics coordinators can retrain the surveyors early in the survey implementation. The tabulations evaluating individual surveyors can also show whether one or more surveyors are collecting systematically different data. This may be a result of misinterpretation during training, which the survey logistics coordinator can correct early in the survey implementation. It may also be a result of intentionally falsified data, in which case the survey coordinators can take appropriate disciplinary action.
3. Statisticians as subject matter experts. The statisticians for the Maputo follow-up survey were graduate level students, who had experience with similar types of household surveys undertaken in Africa. These two statisticians were able to use this past survey experience, experience with statistical software packages, and general statistical knowledge to analyze and present the data in ways scientists from other disciplines would not be able. These findings reinforce the assertion by Bethlehem (1997) that data editing improves when the statistician becomes a subject matter expert rather than just a data analyst.
4. Extended training of surveyors. The overall data review and cleaning process described in this paper led to several benefits that emerged from the constant level of communication required among members of the research team. Within the early days of the fieldwork, the surveyors became acutely aware of the importance of completing accurate surveys. Further, the regular feedback they received through the error review process effectively continued their training through the entire fieldwork. This approach stands in stark contrast to the more traditional method of training surveyors in one intensive period before the fieldwork, which we argue is an inadequate process based on our experience. Without this level of data review, a number of significant data collection errors would have persisted throughout

the fieldwork, which would ultimately have affected the research findings.

Acknowledgements

The research described in this paper was undertaken as part of a Stanford University research program on non-network water and sanitation in developing countries funded by the Woods Institute for the Environment. The authors would like to thank Dr. Jenna Davis, Dr. Valentina Zuin, and Maika Elizabeth Nicholson of Stanford University for the opportunity to collaborate on the project and their comments; Marcos Carzolio of Virginia Tech for his participation in the project; and Dr. J.P. Morgan of Virginia Tech and Dr. David Banks of Duke University for their comments.

REFERENCES

Bethlehem, J.G. 1997. "Integrated Control Systems for Survey Processing." In *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. Leeuw, C. Dippo, N. Schwarz, and D. Trewin, 371–92. New York: John Wiley.

De Waal, T., J. Pannekoek, and S. Scholtus. 2011. "Handbook of Statistical Data Editing and Imputation." http://www.Wiley.com.