Using Unobtrusive Sensors to Measure and Minimize Hawthorne Effects: Evidence from Cookstoves

Andrew M. Simons, Theresa Beltramo, Garrick Blalock, David I. Levine*

People act differently when they know they are being observed. This phenomenon, the Hawthorne effect, biases estimates of program impacts. Unobtrusive sensors substituting for human observation can remove this bias. To demonstrate this potential, we used temperature loggers to measure fuel-efficient cookstoves as a replacement for three-stone fires. We find a large Hawthorne effect: when in-person measurement begins participants increase fuel-efficient stove use approximately three hours/day (54%) and reduce three-stone fire use by approximately two hours/day (32%). When in-person measurement ends, participants reverse those changes. Our results reinforce concerns about Hawthorne effects, and demonstrate that sensors can sometimes provide a solution.

Keywords: observation bias, Hawthorne effect, sensors, improved cookstoves, monitoring and evaluation, impact evaluation

* Simons and Blalock: Charles H. Dyson School of Applied Economics and Management, Cornell University (email: ams727@cornell.edu and garrick.blalock@cornell.edu). Beltramo: Impact Carbon, San Francisco (email: tbeltramo@impactcarbon.org). Levine: Haas School of Business, University of California, Berkeley (email: levine@haas.berkeley.edu).

Acknowledgments: This study was funded by the United States Agency for International Development under Translating Research into Action, Cooperative Agreement No. GHS-A-00-09-00015-00. The recipient of the grant was Impact Carbon who co-funded and managed the project. Juliet Kyaesimira and Stephen Harrell expertly oversaw field operations and Amy Gu provided excellent research support. We thank Impact Carbon partners Matt Evans, Evan Haigler, Jimmy Tran, Caitlyn Toombs, and Johanna Young; U.C. Berkeley Household Energy, Climate, and Health Research Group partners including Kirk Smith, Ilse Ruiz-Mercado, and Ajay Pillarisetti; Berkeley Air partners including Dana Charron, David Pennise, Michael Johnson, and Erin Milner; the USAID TRAction Technical Advisory Group, and seminar participants at Cornell University, UCLA, and Oxford University for valuable comments. Data collection was carried out by the Center for Integrated Research and Community Development (CIRCODU), and the project's success relied on expert oversight by CIRCODU's Director General Joseph Ndemere Arineitwe and field supervisors Moreen Akankunda, Innocent Byaruhanga, Fred Isabirye, Noah Kirabo, and Michael Mukembo. We thank the Atkinson Center for a Sustainable Future at Cornell University, the Institute for the Social Sciences at Cornell University and the Cornell Population Center for additional funding of related expenses. The findings of this study are the sole responsibility of the authors, and do not necessarily reflect the views of their respective institutions, nor USAID or the United States Government.

Introduction

The validity of empirical research depends on the quality of the underlying data. Unlike the physical sciences, for which data often is generated in controlled laboratory settings, the social sciences construct variables involving human behaviors that make ensuring high data quality a challenge. Respondents often do not answer surveys candidly (Bertrand and Mullainathan 2001) and the act of surveying can change later behaviors of those being surveyed (Zwane et al. 2011). These drawbacks to surveys have been one factor contributing to a push for more experiments in social science research (Falk and Heckman 2009; Banerjee and Duflo 2009; Duflo, Glennerster, and Kremer 2008). While much has been learned from social science experiments they are prone to issues less prevalent in the physical sciences, such as observation bias or Hawthorne effect.

We explore an emerging class of technology—small, inexpensive, and unobtrusive sensors—as a remedy to the Hawthorne effect. A growing variety of sensors have become available to researchers. GPS trackers and motion detectors, for example, allow non-obtrusive measurement of subject location and body movements (Ermes et al. 2008). Medical doctors wear badges with sensors that detect the scent of alcohol used in hand sanitizers to alert the doctor and/or patient if the doctor has not washed his or her hands recently (Smith 2014). Loop detectors installed in the lanes of freeways allow monitoring of congestion and driver behavior (Bento et al. 2014).

The degree to which these sensors interfere with subjects' behavior can vary widely. In some cases, individuals may self-select to be observed to intentionally motivate a behavioral response. For example, long-distance bikers and runners can opt into programs that will report the location, time, and speed of excursions to a website that others can monitor (Mueller et al. 2010). Such schemes are intended to use peer observation as a motivation aid. In other cases, such as room occupancy detectors that control lighting and climate control, the sensor may be far more unnoticed (Buchanan, Russo, and Anderson 2014).

We demonstrate a technique to remedy the Hawthorne effect that uses unobtrusive temperature sensors to conduct a small-scale evaluation of technology adoption within a set of households in Uganda. We use minimally invasive temperature sensors to measure usage of the new technology, fuel-efficient cookstoves, and the old technology, traditional three-stone fires.[1] We then compare usage in periods when observers visit the households each day with periods when no observers are present. A major challenge for direct observational studies is that they alter participants' behavior (as noted in the cookstove literature (Ezzati, Saleh, and Kammen 2000; Smith-Sivertsen et al. 2009) and in social sciences more broadly (Schwartz et al. 2013; Leonard and Masatu 2010; Das, Hammer, and Leonard 2008; Leonard 2008; Levitt and List 2007; Leonard and Masatu 2006)). We find a large

[1] A three-stone fire is simply three large stones, approximately the same height, on which a cooking pot is balanced over a fire.

Hawthorne effect: households increase the use of the fuel-efficient stove and decrease the use of three-stone fires when observers are present.

In our setting, the daily observers measured wood use and household exposure to particulate matter. Thus, those observations will be affected by changes in cooking patterns due to the presence of observers. Once the magnitude of this Hawthorne effect is known, we can adjust the hours of stove use per day when observers were present, and estimate unbiased impacts of how fuel-efficient stoves affect wood use and exposure to particulate matter.

Methods

We executed a series of randomized control trials in rural areas of the Mbarara District in southwestern Uganda from February to September 2012. Upon arriving in a new parish, staff displayed the fuel-efficient stove (Envirofit G-3300)[2] and offered it for sale to anyone who wanted to purchase at 40,000 Ugandan Shillings (approximately USD $16, see (Levine et al. 2013) for an overview of the sales contract). Consumers who wanted to buy the stove were randomly assigned into two groups (early buyers, late buyers). We asked both early buyers and late buyers if they would agree to have a temperature data logger placed on their traditional stoves immediately. Then approximately two weeks later the early buyers received their first Envirofit stove, and approximately four to five weeks after that the late buyers received their first Envirofit stove. In each parish, more than twelve households agreed to join the study; therefore among those that agreed, we randomly selected twelve households per parish for the usage study.

Approximately four weeks after late buyers received their Envirofits, both groups were surprised with a second Envirofit stove. Common cooking practices in the area require two simultaneous cooking pots (for example rice and beans, or *matooke* (starchy cooking banana) and a sauce); therefore, because the Envirofit is sized for one cooking pot, we gave a second Envirofit to enable normal cooking behavior as much as possible. The experimental analysis of how the new stoves affected stove usage and other outcomes (wood use, exposure to particulate matter, and so forth) are not the focus on this paper and are analyzed elsewhere. The study tracked stove usage both before and after the purchase of a fuel-efficient stove in fourteen rural parishes in Mbarara (168 total households).

To track usage, we used small, inexpensive and unobtrusive sensors: stove use monitors (SUMs) that record stove temperatures without the need for an observer to be present.[3] Using SUMs to log stove temperatures was suggested by (Ruiz-

[2] Product manufacturer states that the Envirofit G-3300 reduces smoke and harmful gasses by up to 80%, reduces biomass fuel use by up to 60% and reduces cooking time by up to 50%. For more details see: http://www.envirofit.org/products/?sub=cookstoves&pid=10.

[3] The SUMs used for our project, iButtons™ manufactured by Maxim Integrated Products, Inc., are small stainless steel temperature sensors about the size of a small coin and the thickness of a watch battery which can be affixed to any stove type. Our SUMs record temperatures with an accuracy of +/- 1.3 degrees C up to 85°C. For additional details see the product description website at: http://berkeleyair.com/services/stove-use-monitoring-system-sums/ The SUMs cost approximately

Mercado et al. 2008). We installed SUMs on two Envirofits and the primary three-stone fire (a few households had a second SUM on the secondary three-stone fire, we examine those data in robustness checks).

We also performed standard kitchen performance tests (KPT) (Bailis, Smith, and Edwards 2007) in each household to measure the quantity of fuel wood used, record detailed food diaries, and measure household air pollution. The KPT lasts approximately a week and involves daily visits by a small team of researchers weighing wood, monitoring household air particulate monitors, and collecting survey data on stove usage over the last 24 hours and related topics. Comparing stove usage calculated from the temperature data collected by the SUMs in the week while KPT measurement teams are present versus stove usage in the week before and after the measurement week provides a test of a Hawthorne effect.

We use an algorithm to convert temperature data into daily minutes of stove use (Simons et al. 2014).[4]

Specification

Assign the subscripts *t=-1* to the week prior to measurement week, *t=0* to the measurement week, and *t=1* to the week after the measurement week. Let the coefficient on stove type $s = TSF$ for three-stone fire or ENV for Envirofit, and $Adjacent\_Week$ be a dummy variable for an adjacent week (*t=-1* or *t*=1). The regression is modeled using Ordinary Least Squares (OLS) as:

$$H_{it}^{S} = B^{S} * Adjacent\_Week + I_i + e_{it} \quad (1)$$

where H_{it}^{S} is the total hours cooked per day on stove type *s* for household *i* during the week, I_i is fixed effects for the individual household (controls for household level characteristics that don't change over these three weeks like family size, income, housing, etc.), and e_{it} is an error term. The coefficient B^{S} is the estimate of how different (in hours cooked per day) the average adjacent week is compared to a measurement week on stove type *s*. Standard errors are clustered at the household.

To test the weeks separately, we use a slightly different specification. Let H_{it-1}^{S} be a dummy variable equal to 1 for the week before the measurement week (when *t=-1*) and 0 otherwise, and let H_{it+1}^{S} be a dummy variable equal to 1 for the week after the measurement week (when *t=1*) and 0 otherwise. Then the regression is modeled using OLS as:

USD$16 each and could record temperature data for 24 hours a day for six weeks in a household before needing minimal servicing from a technician. After the data download they can be reset and re-used.

[4] Overnight, while most participants report sleeping, SUMs record the residual heat absorbed in the large stones of the three stone fires and/or from coals banked overnight. Therefore our algorithm overestimates overnight cooking of three stone fires. We adjust for this in the subsequent analysis. For further discussion and a description of the technical adjustment see Harrell et al. (2014).

$$H_{it}^S = \gamma_1^S * H_{it-1}^S + \gamma_2^S * H_{it+1}^S + I_i + e_{it} \quad (2)$$

where I_i is household fixed effects and γ_1^S is the estimate of the difference (in hours cooked per day) of the week before the measurement week compared to the measurement week. The coefficient of γ_2^S is the estimate of the difference cooked in the week after the measurement week compared to the measurement week. Standard errors are clustered at the household.

Results

In the week before the observers arrived (when *t=-1*), primary three-stone fires were used 5.99 hours per day (95% CI = [4.77 to 7.21]) and combined usage on Envirofits was 5.53 hours per day (95% CI = [4.36 to 6.71]). Regression results are in Table 1. On average, usage of the Envirofit stoves is 2.97 hours higher during the measurement week than during the adjacent weeks (95% CI = [1.79 to 4.15], $p<0.01$, column 3). This increase is matched by a reduction of 1.78 hours in usage of the three-stone fire (95% CI = [0.86 to 2.70], $p<0.01$, col. 1).

In columns 2 and 4 we relax the assumption that stove usage is the same in the week prior to and the week after our measurement period. Contrasted with the measurement week, households use their primary three-stone fire 1.17 hours per day more in the prior week (95% CI = [0.10 to 2.24], $p<0.05$, col. 2) and 2.37 hours more in the following week (95% CI = [1.12 to 3.62], $p<0.01$). These coefficients are jointly significantly different than zero ($p<0.01$), but not statistically significantly different from each other ($p=0.10$).

The total usage of Envirofits follows a mirror image (col. 4), and is 2.58 hours per day lower in the week prior to measurement week than in measurement week (95% CI = [1.21 to 3.94], $p<0.01$) and 3.30 hours per day lower the following week (95% CI = [2.04 to 4.57], $p<0.01$). These coefficients are jointly significantly different from zero ($p<0.01$), but not statistically significantly different from each other ($p=0.20$).

Adjusting for the Hawthorne effect

Because the kitchen performance test is widely used to measure the effects of new cookstoves on fuel usage and household air pollution (Smith et al. 2007), estimates of how new stoves affect fuel use and carbon emissions may be substantially biased. The same bias can arise in studies, such as ours, that measure household air pollution or health effects with repeated household visits.

The field of epidemiology has efficacy trials (testing the effects of an intervention under ideal conditions) and effectiveness trials (testing the effects of an intervention under realistic conditions) (Flay 1986). In the context of cookstoves, the kitchen performance test provides a valid measure of how the new stove affects wood usage during the measurement week (efficacy); however, we need to adjust for the gap in usage between measurement weeks and weeks when no observers are influencing behaviors to generalize to weeks without daily visits (that is, to estimate effectiveness).

Consider the following illustrative example. Assume that households without Envirofits use three-stone fires twelve hours per day and households with Envirofits use three-stone fires nine hours a day and Envirofits four hours a day. During the KPT, households with Envirofits use three-stone fires seven hours a day and Envirofits six hours per day. Assume that three-stone fires create one unit and Envirofits create a half unit of pollution per hour.

With these illustrative assumptions, pollution per day declines from twelve units prior to the introduction of an Envirofit to eleven units of pollution (9 + (½*4) = 11) once the Envirofit is present, a decline of 8.33%. However, if instead we used the data from the kitchen performance test, we would estimate a decline in pollution from twelve units to ten units (7 + (½*6) = 10), a decline of 16.67%, or twice the true decline. While these figures are merely illustrative, they show the importance of adjusting data to minimize bias caused by Hawthorne effects.

Discussion

We demonstrate a technique to measure the magnitude of a Hawthorne effect in an experimental setting in the developing world and remove it. While other forms of unobtrusive objective monitoring exist—such as using administrative records when reliable (Angrist, Bettinger, and Kremer 2006) or tracking take up at a remote location via redeemed vouchers (Dupas 2009; Dupas 2014)—the recent explosion of small, inexpensive, and unobtrusive sensors expands researchers' ability to quantify and remove observation bias. A wide variety of emerging technologies can be utilized, a partial list includes: smart phones tracking locations through GPS, remote sensors that detect latrine usage (Clasen et al. 2012), sensors to remotely detect the use of water filters (Thomas et al. 2013), medical devices to monitor the hand hygiene of medical professionals (Boyce 2011), smart grid or other energy monitors (Darby 2010), and pedometers or other devices that monitor physical activity (Bravata et al. 2007). Adjusting for Hawthorne effects is important if the results of impact evaluations are intended to generalize beyond periods of intense in-person observation.

Bibliography

Angrist, Joshua, Eric Bettinger, and Michael Kremer. 2006. "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia." *American Economic Review* 96 (3): 847–62.

Bailis, Rob, Kirk R. Smith, and Rufus Edwards. 2007. *Kitchen Performance Test (KPT)*. Univerisity of California, Berkeley, CA.

Banerjee, Abhijit V., and Esther Duflo. 2009. "The Experimental Approach to Development Economics." *Annual Review of Economics* 1 (1): 151–78. doi:10.1146/annurev.economics.050708.143235.

Bento, Antonio, Daniel Kaffine, Kevin Roth, and Matthew Zaragoza-Watkins. 2014. "The Effects of Regulation in the Presence of Multiple Unpriced Externalities: Evidence from the Transportation Sector." *American Economic Journal: Economic Policy* 6 (3): 1–29. doi:10.1257/pol.6.3.1.

Bertrand, Marianne, and Sendhil Mullainathan. 2001. "Do People Mean What They Say? Implications for Subjective Survey Data." *American Economic Review* 91 (2): 67–72.

Boyce, John M. 2011. "Measuring Healthcare Worker Hand Hygiene Activity: Current Practices and Emerging Technologies." *Infection Control and Hospital Epidemiology* 32 (10): 1016–28. doi:10.1086/662015.

Bravata, Dena M, Crystal Smith-Spangler, Vandana Sundaram, Allison L Gienger, Nancy Lin, Robyn Lewis, Christopher D Stave, Ingram Olkin, and John R Sirad. 2007. "Using Pedometers to Increase Physical Activity: A Systematic Review." *Jama* 298 (19).

Buchanan, Kathryn, Riccardo Russo, and Ben Anderson. 2014. "Feeding Back about Eco-Feedback: How Do Consumers Use and Respond to Energy Monitors?" *Energy Policy* 73 (June): 138–46. doi:10.1016/j.enpol.2014.05.008.

Clasen, Thomas, Douglas Fabini, Sophie Boisson, Jay Taneja, Joshua Song, Elisabeth Aichinger, Anthony Bui, et al. 2012. "Making Sanitation Count: Developing and Testing a Device for Assessing Latrine Use in Low-Income Settings." *Environmental Science & Technology* 46 (6): 3295–3303. doi:10.1021/es2036702.

Darby, Sarah. 2010. "Smart Metering: What Potential for Householder Engagement?" *Building Research & Information* 38 (5): 442–57. doi:10.1080/09613218.2010.492660.

Das, Jishnu, Jeffrey Hammer, and Kenneth Leonard. 2008. "The Quality of Medical Advice in Low-Income Countries." *Journal of Economic Perspectives* 22 (2): 93–114.

Duflo, Esther, R Glennerster, and Michael Kremer. 2008. "Using Randomization in Development Economics Research: A Toolkit." Edited by T Paul Schultz and John A Strauss. *Handbook of Development Economics* 4 (07): 3895–3962. doi:10.1016/S1573-4471(07)04061-2.

Dupas, Pascaline. 2009. "What Matters (and What Does Not) in Households' Decision to Invest in Malaria Prevention?" *American Economic Review* 99 (2): 224–30. doi:10.1257/aer.99.2.224.

———. 2014. "Short-Run Subsidies and Long-Run Adoption of New Health Products: Evidence from a Field Experiment." *Econometrica* 82 (1): 197–228.

Ermes, Miikka, Juha Pärkkä, Jani Mäntyjärvi, and Ilkka Korhonen. 2008. "Detection of Daily Activities and Sports With Wearable Sensors in Controlled and Uncontrolled Conditions." *IEEE Transactions on Information Technology in Biomedicine* 12 (1): 20–26. doi:10.1109/TITB.2007.899496.

Ezzati, Majid, Homayoun Saleh, and Daniel M Kammen. 2000. "The Contributions of Emissions and Spatial Microenvironments to Exposure to Indoor Air Pollution from Biomass Combustion in Kenya." *Environmental Health Perspectives* 108 (9): 833–39.

Falk, Armin, and James J Heckman. 2009. "Lab Experiments Are a Major Source of Knowledge in the Social Sciences." *Science* 326 (5952): 535–38. doi:10.1126/science.1168244.

Flay, Brian R. 1986. "Efficacy and Effectiveness Trials (and Other Phases of Research) in the Development of Health Promotion Programs." *Preventive Medicine* 15 (5): 451–74.

Harrell, Stephen, Theresa Beltramo, David I. Levine, Garrick Blalock, and Andrew M. Simons. 2014. “What Is a Meal?: Comparing Methods of Auditing Carbon Offset Compliance for Fuel Efficient Cookstoves.” Dyson School of Applied Economics and Management Working Paper. Ithaca, NY.

Leonard, Kenneth L. 2008. “Is Patient Satisfaction Sensitive to Changes in the Quality of Care? An Exploitation of the Hawthorne Effect.” *Journal of Health Economics* 27 (2): 444–59. doi:10.1016/j.jhealeco.2007.07.004.

Leonard, Kenneth L, and Melkiory C Masatu. 2006. “Outpatient Process Quality Evaluation and the Hawthorne Effect.” *Social Science & Medicine* 63 (9): 2330–40. doi:10.1016/j.socscimed.2006.06.003.

Leonard, Kenneth L., and Melkiory C. Masatu. 2010. “Using the Hawthorne Effect to Examine the Gap between a Doctor’s Best Possible Practice and Actual Performance.” *Journal of Development Economics* 93 (2): 226–34. doi:10.1016/j.jdeveco.2009.11.001.

Levine, David I., Theresa Beltramo, Garrick Blalock, and Carolyn Cotterman. 2013. “What Impedes Efficient Adoption of Products? Evidence from Randomized Variation in Sales Offers for Improved Cookstoves in Uganda.” CEGA Working Paper Series. Berkeley, CA. http://escholarship.org/uc/item/86v4x8nn#page-1.

Levitt, Steven D, and John A List. 2007. “What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?” *Journal of Economic Perspectives* 21 (2): 153–74.

Mueller, Florian, Frank Vetere, Martin R Gibbs, Stefan Agamanolis, and Jennifer Sheridan. 2010. “Jogging over a Distance: The Influence of Design in Parallel Exertion Games.” *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games*, 63–68.

Ruiz-Mercado, Ilse, Nick L Lam, Eduardo Canuz, Gilberto Davila, and Kirk R Smith. 2008. “Low-Cost Temperature Loggers as Stove Use Monitors (SUMs).” *Boiling Point* 55: 16–18.

Schwartz, Daniel, Baruch Fischhoff, Tamar Krishnamurti, and Fallaw Sowell. 2013. “The Hawthorne Effect and Energy Awareness.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (38): 15242–46. doi:10.1073/pnas.1301687110/.

Simons, Andrew M., Theresa Beltramo, Garrick Blalock, and David I. Levine. 2014. “Comparing Methods for Signal Analysis of Temperature Readings from Stove Use Monitors.” *Biomass and Bioenergy* 70: 476–88.

Smith, Eleanor. 2014. “Better Hygiene Through Humiliation.” *The Atlantic*, August 13.

Smith, Kirk R, Karabi Dutta, Chaya Chengappa, P.P.S. Gusain, Omar Masera, Victor Berrueta, Rufus Edwards, Rob Bailis, and Kyra Naumoff Shields. 2007. “Monitoring and Evaluation of Improved Biomass Cookstove Programs for Indoor Air Quality and Stove Performance: Conclusions from the Household Energy and Health Project.” *Energy for Sustainable Development* 11 (2): 5–18.

Smith-Sivertsen, Tone, Esperanza Díaz, Dan Pope, Rolv T Lie, Anaite Díaz, John McCracken, Per Bakke, Byron Arana, Kirk R Smith, and Nigel Bruce. 2009. “Effect of Reducing Indoor Air Pollution on Women’s Respiratory Symptoms and Lung Function: The RESPIRE Randomized Trial, Guatemala." *American Journal of Epidemiology* 170 (2): 211–20. doi:10.1093/aje/kwp100.

Thomas, Evan A, Christina K Barstow, Ghislaine Rosa, Fiona Majorin, and Thomas Clasen. 2013. "Use of Remotely Reporting Electronic Sensors for Assessing Use of Water Filters and Cookstoves in Rwanda." *Environmental Science & Technology* 47 (23): 13602–10. doi:10.1021/es403412x.

Zwane, Alix Peterson, Jonathan Zinman, Eric Van Dusen, William Pariente, Clair Null, Edward Miguel, Michael Kremer, et al. 2011. "Being Surveyed Can Change Later Behavior and Related Parameter Estimates." *Proceedings of the National Academy of Sciences of the United States of America* 108 (5): 1821–26. doi:10.1073/pnas.1000776108.

Table 1

Regressions testing for Hawthorne effect: estimates of effects of the presence of measurement team in primary three stone fire (TSF) usage and combined Envirofit usage, the coefficients represent the change in hours cooked per day compared to hours cooked per day in the measurement week

	Primary TSF		Combined Envirofit	
	(1)	(2)	(3)	(4)
Week prior to and after measurement week constrained to be equal	1.78***		-2.97***	
	(0.46)		(0.60)	
Week prior to measurement week		1.17**		-2.58***
		(0.54)		(0.69)
Week after measurement week		2.37***		-3.30***
		(0.63)		(0.64)
Household fixed effects	Yes	Yes	Yes	Yes
Observations	316	316	229	229
R-squared	0.82	0.82	0.79	0.79
Household clusters	118	118	89	89

Standard errors clustered at household level in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: The unit of analysis is a measurement "week" (approximately 72 hours) at a household. The specification in columns 1 and 3 imposes that the weeks prior to and after the measurement week are equal. The specification in columns 2 and 4 tests usage in the week prior to and after the measurement week separately. The coefficient estimates in column 2 are jointly significantly not equal to zero (p<0.01), but not statistically different from each other (p=0.10). The coefficient estimates in column 4 are jointly significantly not equal to zero (p<0.01), but not statistically different from each other (p=0.20).