Journal of Development Economics 120 (2016) 99–112

Contents lists available at ScienceDirect

Journal of Development Economics

journal homepage: www.elsevier.com/locate/devec

Measuring the measurement error: A method to qualitatively validate survey data☆

Christopher Blattman [a,*], Julian Jamison [b], Tricia Koroknay-Palicz [c], Katherine Rodrigues [d], Margaret Sheridan [e]

[a] Columbia University SIPA, 420 W 118th St., New York, NY, United States
[b] Global Insights Initiative, The World Bank, 1818 H St NW, Washington, DC, United States
[c] The World Bank, 1818 H St NW, Washington, DC, United States
[d] International Rescue Committee, Research Department, 122 East 42nd St., New York, NY, United States
[e] University of North Carolina at Chapel Hill, Clinical Psychology, United States

ARTICLE INFO

Article history:
Received 25 November 2014
Received in revised form 11 January 2016
Accepted 15 January 2016
Available online 29 January 2016

Keywords:
Measurement error
Survey data
Validation
Field experiments
Liberia
Crime
Drugs
Risky behaviors

ABSTRACT

Empirical social science relies heavily on self-reported data, but subjects may misreport behaviors, especially sensitive ones such as crime or drug abuse. If a treatment influences survey misreporting, it biases causal estimates. We develop a validation technique that uses intensive qualitative work to assess survey misreporting and pilot it in a field experiment where subjects were assigned to receive cash, therapy, both, or neither. According to survey responses, both treatments reduced crime and other sensitive behaviors. Local researchers spent several days with a random subsample of subjects after surveys, building trust and obtaining verbal confirmation of four sensitive behaviors and two expenditures. In this instance, validation showed survey underreporting of most sensitive behaviors was low and uncorrelated with treatment, while expenditures were under reported in the survey across all arms, but especially in the control group. We use these data to develop measurement error bounds on treatment effects estimated from surveys.

1. Introduction

The trouble with many survey topics, whether it's abortion, drug use, crime, domestic violence, or support for terrorism, is that people may not tell the truth. This makes survey data on any sensitive topic suspect. Even without incentives to misreport, self-reported data are often inaccurate. Studies show people even misreport their gender and education.[1] When measuring subjects that can embarrass or endanger the respondent, we worry that people might misreport their attitudes or actions.[2]

When we are interested in the impact of a program or event, measurement error will also affect our ability to estimate unbiased causal effects. In dependent variables, random measurement error reduces precision but won't bias estimates.[3] Systematic reporting errors, however, generally bias causal estimates, especially when the measurement error is correlated with the treatment or exogenous event of interest. For instance, people who receive an anti-crime message or an addiction treatment might be more likely to respond that they are non-violent or drug free, both because it's socially desirable and because of perceived experimenter demand (where participants conform to the expectations of the people who ran the program).

☆ Acknowledgements: For comments we thank Neal Beck, Alex Coppock, Dan Corstange, Macartan Humphreys, Don Green, Cyrus Samii, Chris Udry, several anonymous referees, and participants at the NYU 2014 CESS conference. This study was funded by the National Science Foundation (SES-1317506), the World Bank's Learning on Gender and Conflict in Africa (LOGiCA) trust fund, the World Bank's Italian Children and Youth (CHYAO) trust fund, the Department of International Development, UK (DFID, GA-C1-RA2-114) via the Institute for the Study of Labor (IZA), a Vanguard Charitable Trust, the American People through the United States Agency for International Development (USAID, AID-OAA-A-12-00066) DCHA/CMM office, and the Robert Wood Johnson Health and Society Scholars Program at Harvard University (Cohort 5). The contents of this study are the sole responsibility of authors and do not necessarily reflect the views of their employers or any of these funding agencies or governments. Finally, for research assistance we thank Foday Bayoh Jr., Natalie Carlson, Camelia Dureng, Mathilde Emeriau, Yuequan Guo, Rufus Kapwolo, James Kollie, Rebecca Littman, Richard Peck, Patryk Perkowski, Colombine Peze-Heidsieck, Joe St. Clair, Joseph Sango Jr., Helen Smith, Abel Welwean, Prince Williams, and John Zayzay through Innovations for Poverty Action (IPA).

* Corresponding author.
E-mail addresses: chrisblattman@columbia.edu (C. Blattman), julison@gmail.com (J. Jamison), t.gonwa@gmail.com (T. Koroknay-Palicz), katherine.rodrigues@rescue.org (K. Rodrigues), sheridan.margaret@unc.edu (M. Sheridan).

[1] See Asher (1974); Bound et al. (2001).
[2] For instance, Karlan and Zinman (2008) find that large numbers of borrowers do not report high-interest consumer loans, potentially because they feel embarrassed.
[3] See Asher (1974); Hausman (2001). This statement applies primarily to linear models.

http://dx.doi.org/10.1016/j.jdeveco.2016.01.005
0304-3878/

Researchers have come up with a number of ways to limit bias in self-reported data. In developed countries, it is common to use administrative data. For example, studies of crime-reduction programs (such as the one we study in this paper) often prefer arrest and incarceration records to self-reported crime (e.g. Deming, 2011). Such data are seldom available in developing countries, however. Moreover, arrest data have serious systematic measurement error problems of their own.[4]

Others use survey experiments and indirect questioning. In list experiments, respondents report the number of items they agree with on a list, which randomly includes or excludes a sensitive item.[5] In endorsement experiments, respondents rate their support for actors expressing sensitive ideas (Bullock et al., 2011). These are valuable tools, albeit with limitations. They can be imprecise and require large samples, and they can be cumbersome when measuring an array of items. Survey experiments also rely on two key assumptions: that people do not lie when counting on a list or endorsing a person, and that the presence of sensitive items doesn't affect reporting of non-sensitive ones (Blair and Imai, 2012).

Finally, in some cases data are physically verifiable and researchers can use a little of what Freedman (1991) called "shoe leather" and simply verify behavior. For instance, in Mexico, the government sent administrators to audit self-reported asset data used to decide who was in or out of a cash transfer program and found underreporting of assets to increase eligibility (Martinelli and Parker, 2009).

This paper develops and field tests an alternative approach for testing the direction and degree of survey misreporting. It is intended to be useful when objective administrative data are not available, survey experiments are impractical, and direct physical verification is impossible. We pilot the approach on self-reported measures of crime, drug use, homelessness, gambling, and discretionary spending. In principle the method could be applied to other sensitive topics where objective assessments are difficult—intimate partner violence, prostitution, risky sex behaviors, participation in communal violence, voting behavior, sexual identity, stigmatized diseases, and so forth.

The approach is relatively simple. We use intense qualitative work—including in-depth participant observation, open-ended questioning, and efforts to build relationships and trust—to try to elicit more truthful answers from a random subsample of experimental subjects. We focus on a very small number of key behaviors, and over several days of trust-building and conversation, we try to elicit a direct admission or discussion of the behavior.

We then compare these qualitative findings to survey responses, and use the difference to estimate the direction, magnitude, and patterns of measurement error. It is effectively a shoe leather approach for difficult-to-verify, often covert behaviors. Like survey experiments, the method relies on the assumption that people are more truthful in this context than in a survey. The techniques we use—spending time with respondents, interacting in their natural environment, developing a rapport, and trying to attain "insider" status—are central techniques in qualitative and ethnographic research to obtain honest and valid responses (e.g. Wilson, 1977; Bryman, 2003).

This paper illustrates the approach, including when, where, and how it could be applied to other field experiments or other causal analysis using survey data. It also describes the patterns of reporting bias that we observe in this particular crime-reduction study, upending the priors we held about the nature and direction of measurement error in these circumstances.

The study recruited a thousand destitute young men in the slums of Liberia's capital, Monrovia, with an emphasis on men involved in petty crime or drugs. The formal evaluation by Blattman et al. (2015) randomized two interventions designed to reduce crime and violence: an 8-week program of group cognitive behavior therapy (CBT) to discourage impulsive, angry, and criminal behaviors; and an unconditional cash transfer of $200.

Obviously, we should be wary of self-reported survey measures of illegal or immoral behavior, especially from a population suspicious of authority, some of whom make their living illicitly. We should be doubly concerned when one of the treatments (therapy) tried to persuade people away from "bad" behaviors, potentially triggering additional social desirability bias or the perception of experimenter demand among the treated. We can imagine any informational or behavioral intervention would raise similar concerns. List experiments were one option, but we found them difficult to implement with a largely uneducated, illiterate population that was selected in part for impulsive behavior.[6] Thus we developed this alternative.

Of more than 4000 endline surveys conducted over the study, we randomly selected roughly 7.3% and attempted to validate survey responses on just six behaviors. Within days of the survey, one member of a small team of Liberian qualitative research staff ("validators") would visit the respondent four times over ten days, each day spending several hours as a participant observer or in active conversation with the man, his peers, and community members. Validators sought a direct admission of the behavior after building trust and familiarity. In effect the method is a very intensive, relationship-based form of survey auditing, which cost (per person) roughly as much as a regular survey to implement.

Validators and the authors then coded an indicator for whether or not the respondent had engaged in each behavior in the two weeks prior to the survey (i.e. during the timeframe about which survey questions on recent behavior were asked). Beforehand, we deemed four behaviors "potentially sensitive": marijuana use, thievery, gambling, and homelessness. Two others were common, non-sensitive behaviors that could be subject to recall bias or other forms of error: paying to watch movies in a video club, and paying to charge their mobile phone at a kiosk. We call these the "expenditure" measures.

This qualitative approach is not free from error: validators could still miss behaviors, make faulty inferences, or let suspicions of treatment status influence their judgment (among other things).[7] These limits of participant observation are well-known (Power, 1989). But these errors, we argue, are less likely to bias treatment effect estimates than the experimenter demand and social desirability bias we worried would cause underreporting in the survey. It comes down to the following proposition: that we can reduce the appearance of experimenter demand (plus other biases correlated with treatment) through four days building rapport and trust, and a focus on only six facts, in the context of what feels to the study participant like everyday conversation rather than a formal survey in which a stranger asks about the same six behaviors in a 300-question, 90-minute questionnaire.

This is the key assumption underlying the technique. It parallels the "no liars" and "no design effects" assumptions in list experiments. As in list experiments, the assumptions cannot be tested directly. But if we accept them, then by comparing survey data to the data collected by validators, we can assess the presence and degree of measurement error in the survey data, and its correlation with treatment assignment.

[4] Arrests underreport true criminal behavior, and they require strong assumptions: that arrests are responses to crimes rather than statistical or other discrimination; and that the treatment doesn't affect the likelihood of being arrested for a crime, by changing the location and observability of the crime for example.

[5] e.g. Raghavarao and Federer (1979). For recent applications see Blair and Imai (2012); Jamison et al. (2013); Karlan and Zinman (2012).

[6] For instance, a list experiment read aloud would require many ideas to be held in mind, and we were concerned that answers would be correlated with cognitive abilities.

[7] For instance, as with the survey, conversations between validators and participants may have been influenced by social desirability bias or experimenter demand. Additionally, had the validation exercise relied on observation as the primary source of evidence and the presence of an observer prompted good behavior, we would have underestimated sensitive behaviors in the validation. People have been shown to increase hand-washing behavior, for example, when directly observed, suggesting a Hawthorne effect of observation (Ram et al., 2010). This kind of desirability bias could be greater in a treatment arm, and validators might not eliminate it. Even validators could be biased if they can glean a subject's treatment status. Thus we cannot eliminate all measurement error correlated with treatment status through our approach.