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.

In this specific crime study, one of our main concerns was that men would under report potentially sensitive behaviors due to social desirability bias, and that the therapy treatment might further increase social desirability related under reporting, leading us to observe a treatment effect that in reality is merely a treatment-correlated increase in underreporting. The validation, however, found no evidence that this was the case. Survey-based reporting of potentially sensitive behaviors was quite high: at endline, 22% of men reported stealing in the past two weeks, and 48% admitted to marijuana use. For the four sensitive behaviors, survey responses and validated measures were identical about 80% of the time. Men reported slightly fewer sensitive behaviors in the survey than the validation, but this underreporting was driven mainly by the underreporting of gambling. To the extent that there is survey underreporting of the sensitive behaviors, it is the opposite of what we anticipated: the group that received both cash and therapy was the least likely to under report sensitive behaviors.

Another prior was that expenditure data would be less prone to social desirability bias or measurement error correlated with treatment. In contrast, we found that across all treatment arms expenditures seem to be underreported in the survey relative to validation. This underreporting of expenditures was largest in the control group.

One benefit of the validation is that it upended our priors about the nature and direction of measurement error. Another is that it affects our conclusions in the larger study. Using outcome data from the survey, Blattman et al. (2015) found that cash led to short run income and expenditure gains, which dissipated after a year. They also found that therapy reduced anti-social behaviors, such as crime, immediately and dramatically, but that this change persisted only if the men received therapy and cash. The validation exercise largely bolsters this core result. Indeed, it implies that the treatment effects based on survey data could underestimate the true effects by up to 20%. The validation calls into question, however, the finding that cash led to short run income gains, given that underreporting of expenditures is greater in the control group.

There are several possible explanations for these patterns. Regarding the low level of underreporting of sensitive behaviors, the majority of the men in the study sample are part of a counterculture where drugs and crime are commonplace. Thus, they may be less likely to feel stigma around these behaviors than "normal" society members. Additionally, these men (and their entire counterculture) are already seen by "normal" society as pariahs, and thus, there may be little advantage to hiding such behaviors.

Regarding the higher level of sensitive behavior reporting in the treatment groups, especially therapy plus cash, it's conceivable that therapy makes the men more accustomed and willing to discuss these sensitive issues openly with a member of the project, or that the control group wants to appear better behaved and therefore more deserving of a future program.

Turning to expenditures, underreporting across all arms is consistent with simple recall bias in consumption surveys. Underreporting of gambling and expenditures, especially in the control group, is also consistent with control group members hoping to become eligible for a future program by appearing poorer or more deserving.

Altogether, these findings are crucial to the credibility of the study's experimental estimates, in this case bolstering the claim that the therapy reduced crime and other anti-social behaviors, and moderating the claim that the cash transfer increased incomes. Perhaps more broadly, the findings also challenge conventional notions of the direction of measurement error.

It would be a mistake, however, to cite this paper as evidence that systematic measurement error of sensitive behaviors in high-risk populations is low; that behavioral treatments foster trust and reduce measurement error; or that low-salience expenditures are especially vulnerable to experimenter or recall bias. These are all plausibly true, but before we can generalize more validation needs to be done in more places. An important takeaway message is that, despite several years working with this and similar populations, including extensive quantitative surveys and qualitative interviews, our priors about the most important sources of measurement error were wrong.

We include a detailed description of our procedures to make it easier for other researchers to adapt and use the method. In principle, we think it is applicable to a wide range of risky or stigmatized sexual, health, and economic activities. The cost, in our case, was roughly 3% of the total evaluation budget, a modest amount given that measurement error in self-reported data was the key causal identification concern in the evaluation.

One analog to our approach is in psychology, where virtually every self-reported survey measure of mental health has been validated using structured clinical interviews (e.g. Spitzer et al., 1999). Another is a recent surge of behavioral and other measures to validate survey data on violence, prejudice, and other troublesome outcomes. In addition to the list and endorsement experiments mentioned before, Scacco (2010) interviewed a random subsample of potential religious rioters behind a screen that shielded their identity, and Paluck and Green (2009) measure cooperation by the patterns of distribution of a group survey gift. Finally, business profits and consumption have also proven troublesome to measure and have been the subject of experimental measurement studies. de Mel et al. (2009) experimentally test alternative approaches to measuring microenterprise profits, and find (counterintuitively) that the least intensive methods yield the least biased estimates. Beegle et al. (2012) have experimentally tested various consumption measures against one another. These studies have proven important to the studies where income is the crucial outcome. Ours could prove as useful to interventions targeted at violence, crime, and other risky or stigmatized behaviors. One thing is certain: systematic measurement error is a large and largely unaddressed problem, calling for more such new tools and their refinement and replication.

2. Context and experimental design

In poor countries like Liberia, governments are especially fearful of urban young men and the possibility they will commit crimes, rioting, or election violence. We designed a study to test the economic and behavioral roots of crime and violence among high-risk men. Blattman et al. (2015) describe the study in full detail.

2.1. Full experimental sample

The study recruited 999 young adult men in five neighborhoods of Monrovia, a city of roughly 1.5 million. The study sought out "hard-core street" men—men in their 20s and 30s who live in extreme poverty and may be involved in violence, drugs, and crime. We recruited and implemented the study in three phases over two years, typically in different, distant neighborhoods (see Appendix A). Table 1 describes the study sample at baseline.

On average the men were age 25, had nearly eight years of schooling, earned about $68 in the past month working 46 hours per week (mainly in low-skill labor and illicit work), and had $34 saved. 38% were members of an armed group during the two civil wars that ravaged the country between 1989 and 2003. At baseline, 20% reported selling drugs, 44% reported daily marijuana use, 15% reported daily use of hard drugs, 53% reported stealing something in the past two weeks, and 24% reported that they were homeless in the last two weeks.

2.2. Intervention and experimental design

We designed, implemented, and evaluated two interventions—group cognitive behavior therapy and cash—in a factorial experimental design. We first randomly assigned half the sample to an offer of therapy. Therapy was completed within eight weeks.

Table 1
Description of the study sample (n = 999).

Baseline covariate	Mean	S.D.	Baseline covariate	Mean	S.D.
Age	25.4	(4.86)	Average weekly work hours in:		
Married/Living with partner	16%	(0.37)	Potentially illicit activities	13.6	(27.26)
# of women supported	0.5	(0.64)	Agricultural labor	0.4	(3.69)
# children under 15	2.2	(3.17)	Low-skill wage labor	19.4	(28.85)
Muslim	10%	(0.30)	Low-skill business	11.5	(23.98)
Years of schooling	7.72	(3.29)	High-skill work	1.5	(7.59)
Literacy score (0–2)	1.2	(0.90)	Ex-combatant	38%	(0.49)
Math score (0–5)	2.8	(1.57)	Currently sleeping on the street	24%	(0.43)
Health index (0–6)	4.9	(1.38)	Times went hungry last week	1.26	(1.36)
Disabled	8%	(0.26)	Sells drugs	20%	(0.40)
Monthly cash earnings (USD)	68.30	(84.49)	Drinks alcohol	75%	(0.43)
Durable assets index, z-score	0.00	(1.00)	Uses marijuana daily	44%	(0.50)
Savings stock (USD)	33.70	(67.41)	Uses hard drugs daily	15%	(0.35)
Able to get a loan of $300	11%	(0.31)	Stole in past two weeks	53%	(0.50)

Notes: Surveys were completed with all men, but there are a small number of missing baseline values per respondent. For purposes of regression analysis, these are imputed with the sample median to avoid losing the observation.

Following this, we held a second lottery for grants of $200 with the full sample.[8]

2.2.1. Treatment 1: cognitive behavior therapy and counseling

The therapy was designed and implemented by a local non-profit organization, Network for Empowerment and Progressive Initiatives (NEPI) Liberia. The 8-week program had two main goals. The first was "transformation," or the shift from the position (and self-identity) as an outcast living on the fringe of society to an economically- and socially-integrated member of mainstream society. The second goal was to shift men from present-oriented decision-making to future-oriented goals and behavior.

The approach and curriculum grew out of NEPI's experience, but were largely grounded in cognitive behavioral therapy (CBT) theory and practice. Group-based CBT approaches have been validated, typically in US populations, to reduce substance abuse, criminality, and aggression.

Participants met three times a week in groups of about 20, for four hours at a time, led by two facilitators. The only compensation provided for attendance was a bowl of rice and simple stew. On alternate days when the group did not meet, the facilitators visited the men at their homes or work areas to provide individual advising and encouragement. Many of the facilitators who ran the group intervention and individual counseling were themselves ex-combatants or reformed street youth.

The CBT element of the program manifested itself in the emphasis on small practical changes each session, which are reinforced through encouragement and praise. These included reducing substance use and abuse, improving body cleanliness, improving the cleanliness of the area in which they lived, and managing their anger without resorting to violence. Facilitators also formally encouraged participants to engage with society in planned and unaccustomed ways.

Facilitators also taught skills around planning and goal setting to help participants enhance their future-oriented attitudes, anticipate potential setbacks, and build skills for dealing with adversity. Finally, throughout the eight weeks, facilitators articulated a set of mainstream social norms and encouraged participants to adopt these norms.[9]

2.2.2. Treatment 2: unconditional cash grant

All men were eligible for a cash grant of $200. The cash was both a treatment and also a measurement tool (to see whether spending patterns were affected by the therapy).[10] The framing of the grant was minimalist—people were told that it was random, one-time, and unconditional.[11]

2.3. Survey data collection

The research team, a Liberia branch of the non-profit research organization Innovations for Poverty Action (IPA), presented themselves as independent evaluators.[12]

We attempted to collect survey data from each recruit five times: at baseline prior to the intervention; at "short-run" endline surveys roughly 2 and 5 weeks after the cash transfers; and at two "long-run" endline surveys 12 and 13 months after the cash grants.[13]

Because the sample was exceptionally mobile and difficult to track over time, we took special measures to minimize attrition. At baseline we were clear about our desire to stay in touch. We took photos and signature samples, and collected as many as ten different ways to contact each respondent. We documented contact information for each respondent, including all the places they said they sometimes stay, plus contact information for the network of people around them who have a more stable location. Respondents were often on the run from the police or other people, and so their contacts might be uncomfortable speaking to enumerators and revealing the respondent's location. Thus, after the baseline survey, we asked respondents to use the enumerator's phone to call their most stable contact and introduce the enumerator and study and give permission. At each endline, enumerators would

[8] None knew of the cash grant until after therapy was completed. Randomization was done through public draw in blocks of roughly 50. There is balance across treatment and control groups. 90% of all men assigned to the therapy attended at least six days of the therapy. Those who did not attend had slightly less schooling, slightly higher earnings and assets, and are less likely to use drugs or alcohol or steal. Thus it appears that the highest risk young men were the most likely to attend. See Appendix A for details.

[9] These include discouragement of crime, substance abuse, and interpersonal violence (encouraging instead the use of peaceful solutions to conflict). The program also encouraged good financial management, especially saving money, as an important aspect of future- and goal-oriented behavior.

[10] An international non-profit, Global Communities, conducted the cash distribution. These partners conducted all recruitment and program implementation to minimize the perceived connection between the research team and programs.

[11] Prior to the lottery, the group merely received a short lecture (15–30 min) on how to safeguard the funds once received. Of those assigned to the cash grant, 98% received it.

[12] They visually distinguished themselves from other organizations by wearing uniquely colored emerald green t-shirts and identification badges over the years of the study. The exception to this is the validators, who wore street clothes that helped them blend in with the study participants.

[13] The exception is the 100 men in the pilot phase, who had a single "short-run" survey 3 weeks after the grant, and a pair of "medium-run" surveys at 5 and 7 months in addition to the 12- and 13-month surveys. We ran pairs of short-run and long-run surveys because it allowed us to take two measures of relatively noisy outcomes with potentially low autocorrelation such as earnings, expenditures, criminal activity, drug use, and so forth. Taking multiple measurements at short intervals allows one to average out noise, increasing power (McKenzie, 2012). Each survey was roughly 90 minutes long, followed by roughly 90 min of interactive behavioral games and psychological tests. Liberian enumerators conducted face-to-face interviews in Liberian English using handheld electronic devices.

typically start with the phone numbers of the various contacts or respondent and try to arrange an appointment. Contacts received no financial incentive. If this strategy failed, the enumerators would begin visiting the various locations listed. A slight majority of respondents were found within a few hours. In other cases, all leads were cold and more extensive sleuthing and asking around the neighborhood was required. If someone had traveled or moved far away, enumerators either waited until they returned or traveled across the country to find them in person. On the upper tail, it could take three to four days of physical searching to find the hardest-to-locate people. Enumerators only stopped searching when all possible leads had been exhausted.

By making at least four attempts to track each man, we were able to track and survey around 93% of the target sample across all endline survey rounds. Attrition is not strongly correlated with baseline covariates or treatment assignment.[14]

3. Empirical strategy

To motivate the empirical tests, we outline a simple model of the effect of different forms of measurement error in outcomes in the context of an experimental intervention. We adapt the simple linear model from the Bound et al. (2001) review of measurement error for these illustrative purposes. We use the language of experiments throughout, but the same principles could be applied to observational causal inference using survey data. In this simple example, we suppose the true treatment effect specification is:

$$y^* = \alpha + \theta T + \varepsilon \tag{1}$$

where y^* is the true outcome and T is an indicator for assignment to treatment.[15] The observed survey outcome y^s, however, measures the true outcome with both systematic and random error:

$$y^s = \delta^s y^* + \gamma^s T + \mu \tag{2}$$

where we assume the random error μ is uncorrelated with y^*, T, and ε. Throughout this illustration, δ (which we take to be positive throughout) denotes systematic measurement error of the true outcome (such as underreporting due to social desirability bias) and γ indicates error associated with treatment only (as in the case of experimenter demand, for example).

Attempting to calculate treatment effects on y^* using only y^s, the researcher estimates the following potentially erroneous equation:

$$y^s = \hat{\alpha} + \hat{\theta} T + \hat{\varepsilon} \tag{3}$$

By substituting Eq. 1 into 2 and comparing to 3, we can see that the researcher estimates the treatment effect $\hat{\theta} = \delta^s \theta + \gamma^s$, and the bias from the true treatment effect θ is:

$$E\left(\hat{\theta} - \theta\right) = (\delta^s - 1)\theta + \gamma^s \tag{4}$$

There are three main cases to consider:

- $\delta^s = 1$ and $\gamma^s = 0$ is the special case of classical (random) measurement error involving μ only;
- $0 < \delta^s < 1$ and $\gamma^s = 0$ is the case where the survey measure systematically underreports the true outcome (but underreporting is uncorrelated with treatment status), in which case under reporting would bias the estimated treatment effect towards the null, and overreporting (if instead $\delta > 1$) away from the null; and
- $\gamma^s > 0$, which is the more worrisome case in which case we mistake measurement error (such as experimenter demand) for a treatment effect.

Now imagine we can collect validation data, y^v, for a random subsample:

$$y^v = \delta^v y^* + \gamma^v T + \eta \tag{5}$$

where η is uncorrelated with T, y^*, ε, and μ. We define the difference in the survey and validation measures as:

$$y^\Delta \equiv y^s - y^v = (\delta^s - \delta^v) y^* + (\gamma^s - \gamma^v) T + \mu - \eta \tag{6}$$

The key assumption in this paper is that the measurement error in the survey and validation data are in the same direction and that validation data correspond more closely to y^* than survey data. That is:

$$0 \le |\delta^v - 1| < |\delta^s - 1| \tag{7}$$

$$0 \le |\gamma^v| < |\gamma^s| \tag{8}$$

$$(\delta^v - 1) * (\delta^s - 1) \ge 0 \tag{9}$$

$$\gamma^s * \gamma^v \ge 0 \tag{10}$$

If assumptions 7 through 10 hold, then y^Δ is a proxy for overreporting (under reporting if negative). The approach described in this paper is only suitable for validation techniques that meet these assumptions. Note that in practice one cannot test them formally for clandestine or otherwise hidden behaviors, and the assumptions must be argued based on context and quality of the process. In much the same way, randomized response and list experiments rely on the assumption of less lying and no design effects, and instrumental variables estimates rely on the exclusion restriction.

If assumptions 7 and 8 hold, however, it means we can identify the direction and approximate magnitude of systematic survey error from the sample mean of y^Δ and assess whether the survey error is correlated with treatment by estimating the treatment regression:

$$y^\Delta = \alpha^\Delta + \theta^\Delta T + \zeta \tag{11}$$

where, since there is a treatment indicator in y^*, $\theta^\Delta = (\delta^s - \delta^v)\theta + \gamma^s - \gamma^v$.

As the validated measure approaches the true outcome measure, then θ^Δ approaches the value of the treatment effect bias described in Eq. 4. That is, as $\delta^v \to 1$ and $\gamma^v \to 0$ then $\theta^\Delta \to E(\hat{\theta} - \theta)$. The main focus of our analysis will be to calculate y^Δ in Eq. 6 and estimate θ^Δ from Eq. 11.

This formalization draws attention to several important caveats associated with any validation technique of this nature:

1. Identification of the bias $E(\hat{\theta} - \theta)$ hinges entirely on the credibility of the validation method and measure. The assumption of lower systematic measurement error is generally untestable and is a judgment call based on the nature and quality of the process.
2. Validation data cannot help us to separately identify the bias arising from general systematic error δ apart from treatment-specific error γ, except in the case where we are willing to make an a priori assumption about one of them, such as that $\gamma = 0$ (i.e. no "John Henry" effects or other forms of experimenter demand). In theory, the systematic and treatment-specific errors could run in opposite directions and cancel one another out. In that case, however, $y^\Delta \neq 0$.

[14] A majority changed locations between each round, many changing sleeping places every few weeks or nights. We generally made at least four attempts to locate each person, in all corners of the country, including prison (to be interviewed only when released). See Appendix A for formal analysis of attrition. The joint significance of all covariates and treatment assignment for survey attrition has a p-value of .53. Attrition is also roughly one percentage point lower in the treatment groups (not statistically significant).

[15] Bound et al. (2001) consider a continuous covariate X rather than indicator T. They also assume that other right-hand side variables are measured without error and have been partialled out. We ignore other covariates in this simple example, but the basic intuitions would hold with them present.

3. So long as the validation measures are imperfect and $0<|\gamma^v|$ or $0<|\delta^v-1|$, the estimates from Eq. 11 will tend to underestimate measurement error. The confidence interval on θ^{Δ} also increases with any noise in the validated measure, η.
4. Nonetheless, to the extent that the validation measures are credible, if we validate a random subset of the study sample we can adjust the distribution of y^* (conditional on T or other covariates) or estimate the "true" treatment effect θ using $\hat{\theta}-\theta^{\Delta}$.

3.1. The special case of binary outcomes

We can further refine this empirical strategy based on the fact that our analysis in this paper will be confined to binary outcomes. Taking into account the binary nature of our dependent variable allows us to derive simple characterizations of under and over-reporting rates ($P(y^s=0|y^v=1)$ and $P(y^s=1|y^v=0)$, respectively), which we can estimate. For example, suppose we specify a model in which under and over-reporting rates differ by whether the validated measure is a 0 or 1:

$$y_i^s=\tilde{\beta}_0+\tilde{\beta}_1 T_i+\tilde{\beta}_2 y_i^v+\tilde{\beta}_3\left(y_i^v\times T_i\right)+\tilde{\mu}_i. \tag{12}$$

Then we can interpret:

- $\tilde{\beta}_0$ as the share of untreated subjects who do not do outcome y according to the validation measure, but report doing it in the survey (over-report);
- $1-\tilde{\beta}_0-\tilde{\beta}_2$ as the share of untreated subjects who engage in outcome y according to the validation measure, but do not report it in the survey (under report);
- $\tilde{\beta}_1$ as the increase in the over-reporting rate due to treatment;
- $\tilde{\beta}_1+\tilde{\beta}_3$ as the decrease in the under reporting rate due to treatment.

See section C.2 for the full derivation of 12. While this more complex specification yields several new estimands of interest, these additional model parameters come at a high cost of statistical power. In our case, with 240 observations in total, each parameter is estimated off of roughly 30 observations, putting us on a steep part of the power curve. While we report results of this approach, we choose to focus on the initial, more simplified, approach because of both these power concerns, and because we are most interested in correcting for the average bias in treatment effects using survey data, which we get from Eq. 11. Nonetheless, the more flexible approach is an option for instances where the returns to validation are especially high, or where the cost is low, such that the analysis is statistically powered. It is also possible to formulate a version of this approach for non-binary outcome variables, e.g. by specifying one or more threshold values above or below which the accuracy of reporting is of special interest.

4. Validation methodology

We selected six variables for validation, all with recall periods of two weeks. We chose outcomes with varying degrees of salience (or memorability) and potential social stigma and experimenter bias. The variables were:

1. *Stealing*. The survey asked how many times in the last two weeks the respondent stole someone's belongings or deceived or conned someone of money or goods.[16] Based on our fieldwork, we hypothesized that stealing would be the most salient and least socially desirable of all six measures.
2. *Gambling*. The survey asked how many times in the last two weeks the respondent gambled or bet on sports. Beforehand, we hypothesized gambling had a lower level of salience and sensitivity than stealing, but was still somewhat stigmatized.
3. *Marijuana use*. The survey asked how many times in the last two weeks the respondent smoked marijuana. Marijuana use is not socially acceptable across Liberian society overall, but is fairly prevalent in our target demographic. We initially hypothesized underreporting could arise not so much from social stigma, but from the discouragement of drug use in the therapy treatment.
4. *Homelessness*. The survey asked how many times in the last two weeks the respondent had to sleep outside, on the street, or in a market stall because he had no other place to sleep or stay. This is a salient variable where we hypothesized respondents might have under-reported from embarrassment or over-reported in order to appear needier (and eligible for more programs).
5. *Phone charging*. In the expenditure section of the survey, the survey asked how many times in the last two weeks the respondent charged his phone for money. This corresponds to taking one's phone to a kiosk with electricity where one pays a small fee to recharge the battery, a common and routine expense for many Liberians, without stigma and possibly not very memorable. 38% of our sample had a mobile phone at the endline, and 38% reported charging a phone in the last two weeks.
6. *Video club attendance*. In the expenditure section of the survey, the survey asked how many times in the last two weeks the respondent went to a video club. These clubs are private businesses where one can go to watch a movie, television show, or football match for a small fee. This is a popular and socially acceptable pastime, as most Liberians do not have electricity or home entertainment. Salience was unclear, but likely greater than phone charging.

These behaviors also exhibited diversity in program emphasis. Some, like stealing and marijuana use, were highly emphasized in the cognitive behavioral therapy, while others like video club and phone charging were not.

4.1. Validator staff

Eight local staff performed validations over the two years of data collection. We selected validators from the study's qualitative research staff. These people typically began as survey enumerators, but displayed such skill and rapport with the subjects that we hired and trained them to conduct a separate qualitative research component: longitudinal, formal, open-ended interviews with a different subsample of subjects. All conducted the qualitative validation when they were not working on the formal open-ended interviews.[17] Each validator received at least 10 days of training on the methods, including both classroom learning and extensive field training.[18] Like any qualitative study, we believe staff recruitment and training to have been among the most important tasks and also the largest start-up cost of this method.

4.2. Approach

Validators tried to determine whether each respondent had engaged in any of the measured behaviors, even once, in the two weeks preceding the respondent's survey date, as the survey asked about behaviors occurring during the two weeks prior to the survey. We found it optimal

[16] The survey also measured more serious forms of theft, such as armed robbery, but our qualitative validation focussed on non-violent theft.

[17] All but one were men, and all had a high school education. Two of the men completed roughly half the validations with the remainder doing roughly 10 to 20% each. To find these validators, we trained roughly two to three times the number of people needed from the pool of research staff, selecting only those with the most natural questioning and rapport-building skills for the validation exercise.

[18] Details of validator selection and training, team structure, tools and forms are in Appendix B.

for validators to visit each respondent four times, on four separate days, with each visit or "hangout session" lasting approximately three hours. The validator aimed to begin hanging out the day after subjects completed their quantitative surveys and to conduct all four visits in the 10 days following the respondent's endline survey date.

Validators deliberately avoided the feeling of a formal interview and would typically accompany respondents as they went about their business.[19] Validators sometimes took notes during visits, but only in isolated areas out of sight from the respondent.[20] The idea follows from basic principles of ethnography, which seeks to study subjects in their natural settings, similar to those the researcher hopes to generalize about (Wilson, 1977). The intent is to reduce the sense of being in an experimental situation, which ethnographers perceive as creating bias.

The main approach was to engage in casual conversation on a wide range of topics, including the six target topics/measures. The target topics were raised mainly through indirect questions while informally chatting. For example, validators typically started conversations with discussions of family. This was both customary among peers in Liberia and a sign of respect and interest in respondents' lives. It was also a stepping stone for discussing the target behaviors—either because the validator can discuss an issue in their family (someone engaging in one of the activities) or how the respondent's family feels about their current lifestyle and circumstances.

In general, validators found it helpful to tell respondents stories or scenarios about another person or themselves, related to the target measures, then steer the conversation to get information about how respondents had behaved in similar situations, eventually discussing the two weeks prior to the survey. Validators were careful to present these behaviors and incidents in a non-stigmatized light, for instance by discussing a friend who stole in order to get enough to eat, or how they themselves had periods of homelessness or used drugs and alcohol. Validators found that these personal stories (all of which were truthful) and genuineness were essential to building rapport and trust.

Validators might hold these conversations once or twice over the three hours, spending perhaps twenty or thirty minutes in conversation each time, to avoid unnaturally long or awkward conversations. The validator spent the remainder of the three hours in the general vicinity, observing respondents engaging in their daily activities. This could involve taking a rest in the shade or in a tea shop (as is common) or engaging others in conversation. Validators would also try to talk casually with the respondent's friends, relatives, or neighbors to learn about him (although we considered information from these second-hand sources as insufficient to support a conclusion about the respondents' behaviors, but merely as supporting information).

We found that building a rapport with participants in a short space of time was crucial. To develop trusting and open relationships, validators used techniques, including becoming close to respected local community and street leaders, eating meals together, sharing personal information about themselves, assisting subjects with daily activities, and mirroring participant's appearances and vernacular, as appropriate. In addition, validators tried to maintain neutrality and openness while discussing potentially sensitive topics. For instance, conveying—through stories or otherwise—that illicit behaviors were not perceived negatively, allowed respondents to feel comfortable sharing their involvement in such activities. Validators did not lie to or deceive respondents, however.

Overall, this approach—trust-building, spending time together over the course of several days, assuming the role of an "insider," attempting to obtain admission or discussion of the behavior, clandestine but fairly immediate note-taking, and (as discussed below) close examination of the evidence for each respondent with the investigators—was designed to counter the observer bias and selective recall that concern participant observation.[21] Developing a rapport with respondents, spending time to develop a relationship, and obtaining insider status are considered central to obtaining more honest and valid responses (Baruch, 1981; Bryman, 2003; Fox, 2004). We are not aware of any study, however, that has quantitatively tested this proposition.

4.3. Validation sampling and non-response

In each endline survey round we randomly selected study respondents to be validated, stratified by treatment group.[22] Table 2 describes the samples selected for validation in each survey round over the course of the study. In total, we randomly selected 7.3% of all surveys, 297 in total, for validation.

We found 240 (81%) of the 297.[23] This attrition is an identification concern, but there is little evidence of biased attrition. Excess validation attrition (those who were surveyed but not validated) was not robustly associated with baseline characteristics (see Appendix A).

4.3.1. Statistical power

In order to minimize the confidence intervals surrounding any treatment-measurement error correlation, we chose the sample size that maximized the number of interviews we felt qualified validators could manage logistically.[24] Post hoc calculations of statistical power confirm the estimates we made at the design stage. With a sample of 240, we can detect general over- or under reporting greater than 17% of the survey mean (14% of the "true" validated mean).[25] Because each treatment arm is a subsample, however, we cannot precisely measure the effect of treatment on misreporting—it is difficult to detect effects greater than 33% of the survey mean (28% of the validated

[19] On the first visit validators would obtain verbal consent. We designed the consent script to be informal, and explained that the goal of hanging out with the respondent was to talk about some of the same things they discussed in the survey. In addition to this verbal consent, the formal consent form that preceded the recent survey said that qualitative staff may come and visit them again to gather more information.

[20] e.g. in a toilet stall or teashop. If validators were unable to find a secluded area in which to take notes, they sometimes recorded information in their cell phones, pretending to send a text message.

[21] For general discussions of validity in qualitative methods, see LeCompte and Goetz (1982); Power (1989); Wilson (1977).

[22] For each pair of survey rounds, study participants were randomly divided into blocks (e.g. 1, 2, 3, 4), and block 1 study participants were surveyed before block 2, and block 2 before block 3, etc. Within each block we randomly selected validation subjects using a computer-generated uniform random variable. The selection was performed without replacement in a given pair of survey rounds (e.g. the short-term endline surveys in a given phase), but sampling was performed with replacement across survey rounds. Twenty subjects were validated in more than one round.

[23] We could not find 15 for even the endline survey. We could not validate a further 42 because they were difficult to find even immediately after the survey or (more commonly) because they lived a long distance away. In general, we surveyed respondents who had moved far out of Monrovia, but we were unlikely to validate them because of the time and expense and opportunity cost.

[24] In general, the validation sample was a balanced subsample of the full sample (see Appendix A for sampling and balance details). Power calculations, based on roughly the first 60 validator interviews, indicated that there was a modest degree of underreporting of all behaviors, sensitive and non-sensitive, but that the correlation between treatment status and measurement error was uncertain—across outcomes it varied in sign and magnitude, but was about zero on average. Thus the chief advantage of maximizing the sample conditional on time available was to shrink the confidence interval to build confidence in our method and the main outcomes of interest. Further validation was mainly limited by the number of validators we felt could be trained and supervised.

[25] We calculated this minimum detectible effect (MDE) using a two-sided hypothesis test with 80% power at a 0.05 significance level, using baseline and block controls when calculating the R-squared statistic. We calculated an MDE for both the 0–2 expenditures index and the 0–4 sensitive behaviors index. The expenditures index had a mean of .82 in the survey and an MDE of .13 for general over- and under-reporting and .29 for a treatment effect on misreporting. The sensitive behaviors index had a mean of 1.12 in the survey and an MDE of .2 for general over- and under-reporting and .36 for any treatment effect on misreporting. We estimate that doubling the sample size would have increased power by about a third.

Table 2
Validation sample, totals and attrition.

Phase	Survey round	Target # of surveys	# selected for validation	# validated	Reason for no validation data		% validated		
					Unfound at endline	Unfound for validation	All	Treatment	Control
1	3-week	100	0						
	5-month	100	24	18	2	4	75%	75%	75%
	7-month	100	24	12	1	11	50%	50%	50%
	12-month	100	10	6	3	1	60%	63%	50%
	13-month	100	10	8	2	0	80%	86%	67%
2	3-week	398	26	24	0	2	92%	94%	89%
	5-week	398	27	17	0	10	63%	68%	40%
	12-month	398	28	25	2	1	89%	86%	100%
	13-month	398	44	38	1	5	86%	85%	91%
3	3-week	501	0						
	5-week	501	0						
	12-month	501	35	31	2	2	89%	89%	88%
	13-month	501	69	61	5	3	88%	88%	88%
All		4096	297	240	18	39	81%	81%	80%

Notes: The proportion selected in each round was principally a function of logistical feasibility (e.g. number of available staff), and in some none were selected. As procedures became more familiar and staff more experienced, more could be done over time. The percentage validated in the treatment group includes any treatment (cash, CBT, or both).

mean). Thus we are principally interested in the sign and magnitude of the treatment effect on misreporting by treatment group.

4.4. Coding validated data

Validators were unaware of the respondent's survey responses or treatment status, and formed their own opinions (based on the evidence collected) about whether respondents engaged in the six activities during the time period captured by the quantitative survey. Every coding recommendation was then discussed with and vetted by one of the authors.

A core part of the validator training included logical reasoning, supporting reasoning with evidence, and writing this down in a clear and structured manner. After each visit, validators made written notes about the relevant data collected, including evidence to support their conclusions, on a standardized form. At the conclusion of the four visits, the validator coded six indicators, one for each behavior, where "1" meant he had relatively direct evidence the respondent engaged in the behavior during the recall period, and "0" otherwise.[26]

Validators recorded an average of 1.35 "major" pieces of evidence per respondent per behavior to support their coding. This was typically the most persuasive piece or pieces of evidence rather than all evidence collected.[27] Table 3 reports evidentiary methods by behavior. In general, the validators used some form of direct or indirect questioning to elicit a direct admission of the behavior or persuasive statements that respondents did not engage in the behavior. The validators only witnessed or found direct evidence of the behavior in a fifth of cases, or had third party verification in about 6% of cases. In any event, witnessing or third party verification were not sufficient evidence for a final coding. For instance, witnessing had to be followed by questions confirming that the respondent also engaged in the behavior in the two weeks prior to the survey. This accounts for most of the cases where there was more than one piece of evidence highlighted.

Table 3
Evidentiary methods reported by validators, by behavior.

Main evidence techniques	Potentially sensitive behaviors				Expenditures	
	Steal	Marijuana	Gamble	Homeless	Video	Phone
	(1)	(2)	(3)	(4)	(5)	(6)
Avg. pieces of evidence	1.1	1.3	1.1	1.7	1.0	1.2
Obs. (all)	240	240	239	240	239	240
Direct question	36%	35%	38%	5%	32%	1%
Indirect question	28%	46%	42%	62%	59%	92%
Story/Scenario	36%	6%	13%	12%	2%	1%
Witnessed/Found evidence	3%	31%	9%	62%	5%	18%
Third party account	3%	6%	4%	21%	0%	0%
Other/Unclear	3%	9%	6%	13%	6%	5%
Obs. (coded "did not engage" in behavior)	191	118	170	190	93	125
Direct question	38%	44%	39%	5%	34%	0%
Indirect question	26%	46%	44%	60%	58%	98%
Story/Scenario	37%	7%	15%	12%	3%	2%
Witnessed/Found evidence	2%	3%	1%	65%	2%	1%
Third party account	3%	10%	4%	24%	0%	1%
Other/Unclear	2%	1%	1%	14%	4%	0%
Obs. (coded "did engage" in behavior)	49	122	69	50	146	115
Direct question	29%	25%	36%	4%	30%	2%
Indirect question	33%	46%	38%	70%	60%	86%
Story/Scenario	33%	5%	9%	10%	1%	0%
Witnessed/Found evidence	10%	59%	28%	52%	7%	37%
Third party account	4%	2%	4%	8%	0%	0%
Other/Unclear	8%	17%	17%	6%	8%	10%

Notes: Direct questions imply that the validator asked the respondent directly about his engagement in the activity. Indirect questions imply that the validator brought up the subject in general conversation (Where do you live?). Stories and scenarios are a form of indirect questioning where the respondent is invited to comment. Witnessing or found evidence implies that the validator saw the respondent engaging in the activity in question or found physical evidence that the respondent recently engaged in the activity. Third party accounts imply that the validator asked the family and friends of the respondent whether or not he engaged in the activity. Other or unclear methods include a handful of cases of unprompted information offered by the respondent, and also cases where the behavior could be inferred from other knowledge. Mainly it implies that coding was inconclusive or incomplete but is likely a form of questioning.

[26] Over the course of the exercise, different measures offered different experiences and lessons. Because of its relative frequency and visibility, we suspect marijuana use was the easiest to directly observe. But validators found other behaviors straightforward to discuss in conversation. In the survey and (especially) the validation, phone battery charging led to the most confusion—in particular, did simply charging one's phone count, or did only paying to charge one's phone count? Paid charging was the focus of the survey question (it appeared in an expenditure survey module), but we were concerned that the validators would use a more expansive definition. We attempted to mitigate such differences through trainings and regular discussions on the coding.

Homelessness also proved somewhat challenging to measure and validate, as we discovered its definition is subjective. Circumstances arose that were somewhat ambiguous, such as having no home of one's own but regularly sleeping on a friend's floor or in an acquaintance's market stall. To account for the potential variability in perceptions of homelessness, validators were instructed to include as much information as possible about respondents' living situations in their summary reports. The authors then worked with validators to code a somewhat broad definition of homelessness that included any ambiguous circumstances. Prior to analysis, it was not clear whether survey respondents applied the same definition, and hence we err on the side of finding underreporting in the survey.

[27] We do not have complete paper records of all evidence collected, and so the 1.35 pieces of evidence is probably an understatement of the full amount of evidence.

In general, the patterns of evidence are fairly commonsensical. Witnessing is limited to observable behaviors such as marijuana, gambling, homelessness, and phone charging. Stories and scenarios where the respondent is invited to comment or discuss are especially common for the most sensitive subject, stealing. Indirect questioning is most common for everyday topics such as homelessness (e.g. "Is this your house?") and phone charging (e.g. "I need to charge my phone. Where do you usually charge yours?").

4.5. Limitations of the approach

While we think, based on our experiences, that this validation exercise gave enough time to gather detailed, accurate information and fostered trust and frankness, there are nonetheless limitations to this approach.

1. Potential disruption. The presence, and interactions and conversations with the validators may be intrusive and might disrupt respondent's daily activities, thereby altering the findings. To mitigate this risk, validators wore clothes that would blend in with their respondent's environment, and typically accompanied and assisted respondents in their activities as appropriate (e.g. helping a scrap metal collector scavenge).
2. Differences in recall periods. The validation occurred after the time period about which the survey questions had been asked, and validators or respondents could have made errors about the relevant window of time (e.g. homelessness could have been observed the week after the survey, and inferred to the time of the survey incorrectly). This is most likely a source of random measurement error.
3. Inconsistent questions. The survey and validation questions might have been interpreted differently, making it difficult to compare results. However, we used close consultations and reviews of the data, and focus groups with survey and validation staff, to maximize consistency.
4. Reverse Hawthorne effect. Training validators to look for certain behaviors could lead them to overreport those behaviors (akin to the problem of "when you have a hammer everything looks like a nail"). This reverse Hawthorne effect would probably be more of a risk if the validation method relied on passive observation. Rather, validation involved active discussion and (usually) a direct admission of the behavior. Also, one of the authors reviewed and discussed the evidence for every subject with the validator.
5. Increasing social desirability bias. In principle the participant observation method, by building rapport, could lead to a different source of measurement error by (for example) increasing social desirability bias. Our strong sense is that the opposite is true, that trust and rapport reduced the bias, but this is a subjective interpretation and not independently verifiable.
6. Consistency bias. In principle, respondents could recall their survey response and try to remain consistent despite trust-building. This could motivate randomizing the order of validation and survey in the future.
7. Non-blinded validators. The researcher is not immune from bias in qualitative research (LeCompte and Goetz, 1982; LeCompte, 1987). We are especially concerned with any bias correlated with treatment. While validators weren't given the subject's treatment status, it's possible and even likely that this could come up during the extended conversations. Thus there is a danger that the validators' biases will be correlated with treatment. The trust-building and preference for direct admission of the behavior was intended to mitigate this risk, but it still remains.

Most importantly, it seems unlikely that validators would commit most of these errors differentially across study arms. Misreporting correlated with treatment is still a risk under the consistency bias and non-blinded limitations, but the in-depth focus on a handful of questions, time invested, and trust-building is designed to counteract these biases as much as possible.

Finally, like any qualitative work, this is not an off-the-shelf tool. To select and refine the variables, recruit and train validators, and monitor quality of the data requires that researchers have some familiarity with the context and population and at least basic experience in qualitative data collection.

4.6. Replicability of the approach

There are three reasons to think that this method could be replicated in other developing country field experiments and observational analysis using surveys. First, the expertise needed to implement the method effectively exists in most countries. Indeed, it should be considerably simpler to implement outside of Liberia. After fourteen years of civil war, and with one of the lowest human development indices in the world, Liberia has very low local research capacity, even compared to other poor and post-conflict states.

Second, most social scientists are nearly as well prepared to design and implement the approach as they are a new survey instrument or measure. Like any measure or method, it takes local knowledge, care, and extensive pretesting to develop a credible approach, and can benefit from someone with expertise in the subject area. In our case, one of the field research managers had some background in qualitative work and quality assurance, which we believe improved the quality of training and selection of the validator staff.

Third, the cost of the data collection is not necessarily large relative to many field experiments or large-scale panel surveys. In this instance, the fixed cost of startup was primarily in the recruitment and training of the small number of validators—approximately 2 to 3 weeks of work. We estimate the marginal cost of validation was roughly $80 per respondent, mainly in wages and transport. By comparison, the marginal cost of surveying a respondent was roughly $70.[28]

While this method is considerably more expensive than survey experiments, it is more in line with the depth and cost of commonplace efforts to improve consumption measurement through the use of diaries or physical measurement.[29] For crucial measures in large program evaluations, or for statistics informing major policies, the cost is small relative to the intervention, larger study, or larger purpose. For instance, as a proportion of total expenditures on the study, this validation exercise cost under 3% of all research-related costs, and less than 1–2% of program plus research costs.

5. Results

For each of the six behaviors, we construct indicators for that behavior using survey data and the qualitative validation technique, pooling responses from all endline surveys. We also construct additive indices of the four potentially sensitive behaviors and of the two expenditures. Table 4 reports means in the full sample and each treatment arm, as well as the percentage of times the two measures are in agreement. Table 5 reports estimates of y^{Δ}, the difference between the survey and validation measures. Table 6 displays estimates of the correlation between treatment and measurement error. Finally, Table 7 reports treatment effects using the original survey data ($\hat{\theta}$); estimates of bias from measurement error (θ^{Δ}, the correlation between treatment and the survey-validated difference); and adjusted treatment effects that correct for this observed measurement error ($\hat{\theta} - \theta^{\Delta}$).

[28] Both figures were driven by the fact that it typically took one to two days of searching to find each respondent for surveying, plus the time to survey itself. Both surveying and validating in Liberia were expensive by the standards of household surveys, largely because of the cost of operating in a fragile, post-conflict state and the great difficulties in tracking such an unstable population.

[29] In one extreme example, in the India NSS consumption survey, enumerators physically measure the volume of all food consumption (N.S.S.O. Expert Group, 2003).

Table 4
Comparison of survey and qualitative validation means at endline.

	Potentially sensitive behaviors					Expenditures			All
	All (0–4)	Steal	Marijuana	Gamble	Homeless	All (0–2)	Video	Phone	(0–6)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
a. Full sample									
Survey mean	1.12	0.22	0.48	0.18	0.23	0.82	0.42	0.39	1.93
	(1.14)	(0.42)	(0.50)	(0.39)	(0.42)	(0.73)	(0.50)	(0.49)	(1.31)
Validation mean	1.21	0.20	0.51	0.29	0.21	1.09	0.61	0.48	2.30
	(1.18)	(0.40)	(0.50)	(0.45)	(0.41)	(0.74)	(0.49)	(0.50)	(1.21)
% in agreement		79%	85%	72%	82%		62%	82%	
b. Control group									
Survey mean	1.25	0.27	0.48	0.23	0.27	0.68	0.37	0.32	1.93
	(1.31)	(0.45)	(0.50)	(0.43)	(0.45)	(0.70)	(0.49)	(0.47)	(1.44)
Validation mean	1.30	0.23	0.49	0.34	0.23	1.18	0.65	0.54	2.48
	(1.23)	(0.42)	(0.50)	(0.48)	(0.42)	(0.70)	(0.48)	(0.50)	(1.21)
% in agreement		80%	88%	72%	77%		47%	75%	
c. Therapy only									
Survey mean	1.06	0.19	0.48	0.17	0.22	0.81	0.41	0.41	1.87
	(1.11)	(0.39)	(0.50)	(0.38)	(0.42)	(0.75)	(0.50)	(0.50)	(1.35)
Validation mean	1.09	0.17	0.48	0.24	0.20	0.98	0.54	0.44	2.07
	(1.14)	(0.38)	(0.50)	(0.43)	(0.41)	(0.76)	(0.50)	(0.50)	(1.24)
% in agreement		80%	89%	74%	80%		72%	81%	
d. Cash only									
Survey mean	1.03	0.21	0.49	0.13	0.21	0.77	0.37	0.40	1.81
	(1.16)	(0.41)	(0.50)	(0.34)	(0.41)	(0.71)	(0.49)	(0.49)	(1.35)
Validation mean	1.32	0.23	0.53	0.33	0.24	1.00	0.55	0.45	2.32
	(1.26)	(0.42)	(0.50)	(0.47)	(0.43)	(0.81)	(0.50)	(0.50)	(1.33)
% in agreement		76%	82%	74%	90%		56%	85%	
e. Therapy + cash									
Survey mean	1.13	0.22	0.48	0.21	0.22	0.98	0.54	0.44	2.11
	(0.98)	(0.42)	(0.50)	(0.41)	(0.42)	(0.73)	(0.50)	(0.50)	(1.11)
Validation mean	1.11	0.19	0.52	0.24	0.16	1.17	0.70	0.48	2.29
	(1.11)	(0.40)	(0.50)	(0.43)	(0.37)	(0.68)	(0.46)	(0.50)	(1.05)
% in agreement		81%	83%	68%	81%		71%	87%	
Observations	239	238	238	238	239	239	238	239	239

Notes: The table reports the means (standard deviations) of the survey and the qualitatively validated measures for the full sample and by treatment arm. "% in agreement" is the percentage of respondents for whom the survey indicator equals the qualitatively validated indicator.

5.1. Misreporting

5.1.1. Rates of behavior

Overall these are relatively common behaviors within our study sample. According to the survey data reported in Table 4, in the two weeks prior to the survey, 22% stole, 48% used marijuana, 18% gambled, 23% were homeless for at least a night, 42% attended a video club, and 39% paid to charge a mobile phone.

5.1.2. Correspondence in the survey and validator data

In general, the survey and validated data are identical about 80% of the time for sensitive measures and about 70% of the time for expenditures (Table 4). Correspondence is lowest for video club expenditures (62% overall), perhaps because attendance is intermittent and has low salience.

On average, the unadjusted validation means were higher than the survey means, suggesting slight underreporting on the survey. The average person reported 1.21 sensitive behaviors and 1.09 expenditures

Table 5
Survey over-reporting, estimated by the mean difference between survey and validation measures (y^{Δ}).

	Potentially sensitive behaviors					Expenditures		
	All (0–4)	Steal	Marijuana	Gamble	Homeless	All (0–2)	Video	Phone
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Full sample	−0.10	0.02	−0.03	−0.11	0.02	−0.27	−0.19	−0.08
	0.17	*0.57*	*0.24*	***<0.01***	*0.45*	***<0.01***	***<0.01***	***<0.01***
Control group	−0.07	0.03	−0.02	−0.12	0.03	−0.50	−0.29	−0.22
	0.64	*0.57*	*0.71*	*0.09*	*0.60*	***<0.01***	***<0.01***	***<0.01***
Therapy only	−0.04	0.02	0.00	−0.07	0.02	−0.17	−0.13	−0.04
	0.80	*0.77*	*1.00*	*0.29*	*0.77*	*0.08*	*0.07*	*0.53*
Cash only	−0.29	−0.02	−0.05	−0.20	−0.03	−0.23	−0.18	−0.05
	0.04	*0.80*	*0.37*	***<0.01***	*0.42*	***0.03***	***0.03***	*0.32*
Therapy + cash	0.02	0.03	−0.05	−0.03	0.06	−0.19	−0.16	−0.03
	0.91	*0.57*	*0.37*	*0.66*	*0.25*	***0.01***	***0.02***	***0.48***
Observations	239	238	238	238	239	239	238	239

Notes: Columns 1 to 8 report the simple mean differences in the survey and validation measures for the full sample and for each treatment arm, along with p values for a t test of whether the mean is different from zero. We bold p values ≤ 0.05.

Table 6
Estimates of the correlation between treatment and measurement error.

(a) Constrained, with round-block fixed effects (Equation 11)

Covariate	Dependent variable (N = 239)							
	y^s-y^v, Sensitive behaviors					y^s-y^v, Expenditures		
	Stealing	Marijuana	Gambling	Homeless	All (0–4)	Video Club	Phone Charging	All (0–2)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
β_o (Constant)	−0.029	0.062	−0.109	0.093	0.015	−0.326	−0.194	−0.517
	[.087]	[.061]	[.093]	[.076]	[.177]	[.118]***	[.095]**	[.158]***
β_1								
Therapy	−0.019	0.015	0.025	−0.025	−0.004	0.170	0.174	0.339
	[.084]	[.057]	[.097]	[.091]	[.199]	[.102]*	[.085]**	[.132]**
Cash	−0.038	−0.042	−0.085	−0.077	−0.237	0.109	0.165	0.269
	[.088]	[.067]	[.090]	[.079]	[.195]	[.111]	[.078]**	[.134]**
Both	−0.006	−0.024	0.077	0.031	0.079	0.127	0.181	0.304
	[.080]	[.062]	[.095]	[.089]	[.183]	[.103]	[.075]**	[.115]***

(b) Unconstrained, with round-block fixed effects (Equation 12)

Covariate	Dependent variable (N = 239)							
	y^s, Sensitive behaviors					y^s, Expenditures		
	Stealing	Marijuana	Gambling	Homeless	All (0–4)	Video club	Phone charging	All (0–2)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\tilde{\beta}_0$ (Constant)	0.301	0.098	0.231	0.283	0.976	0.367	0.047	0.048
	[.140]**	[.092]	[.118]*	[.129]**	[.287]***	[.130]***	[.093]	[.208]
$\tilde{\beta}_1$								
Therapy	−0.022	0.010	−0.009	−0.036	0.154	−0.190	0.100	−0.011
	[.070]	[.077]	[.074]	[.079]	[.228]	[.124]	[.072]	[.207]
Cash	0.003	0.025	−0.079	−0.138	−0.069	−0.072	0.040	0.089
	[.068]	[.081]	[.064]	[.070]**	[.220]	[.139]	[.068]	[.219]
Both	−0.013	0.025	0.064	−0.025	0.271	−0.113	0.041	−0.150
	[.064]	[.081]	[.076]	[.083]	[.241]	[.138]	[.064]	[.209]
$\tilde{\beta}_2(y^v)$	0.496	0.735	0.315	0.405	0.677	0.038	0.504	0.328
	[.158]***	[.096]***	[.108]***	[.158]**	[.108]***	[.123]	[.096]***	[.129]**
$\tilde{\beta}_3$								
Therapy×y^v	−0.166	−0.014	−0.131	−0.020	−0.210	0.434	0.079	0.222
	[.234]	[.125]	[.176]	[.220]	[.147]	[.169]**	[.143]	[.169]
Cash×y^v	−0.232	−0.114	−0.063	0.286	−0.134	0.133	0.196	0.064
	[.208]	[.134]	[.148]	[.202]	[.144]	[.174]	[.137]	[.169]
Both×y^v	−0.064	−0.085	−0.241	0.066	−0.230	0.386	0.234	0.379
	[.214]	[.129]	[.173]	[.232]	[.137]*	[.171]**	[.130]*	[.168]**

Notes: The table reports the degree and direction of bias in our treatment effects. In panel (a), we assume that our measurement error does not vary by whether or not the individual engages in the behavior, which allows for a simple way to use β_1 to adjust our ITT estimates. In panel (b), we relax this assumption and let the measurement error vary by behavior and treatment arm at the cost of reduced statistical power.
*** p < 0.01. ** p < 0.05. * p < 0.1.

in validation, and 1.12 sensitive behaviors and 0.82 expenditures in the survey.

Note that this is an average, however, and the qualitative validation finds cases of over and under reporting in the survey relative to the validation. There are 328 instances where the measures are not equal: 208 cases of survey under reporting (the survey indicator is a zero and validation indicator is a one), and 120 cases of over-reporting (see Appendix C.2). Most of these differences arose from direct or indirect questioning of the participant. Some form of external evidence, such as direct observation or third party confirmation, additionally supported 21% of the underreports found through validation (especially marijuana use, homelessness, gambling and phone charging) and 14% of over-reports (mainly homelessness).[30]

5.1.3. Underreporting of sensitive behaviors, particularly gambling

Table 5 reports our proxy of survey over-reporting: the simple survey-validation differences, with *p* values from a *t* test of the difference from zero. Negative values indicate survey under reporting, assuming the validator measure is more accurate of course. As noted above, we have the statistical power to detect differences greater than about 17% of the survey mean.

Overall, gambling seems to be slightly underreported in every treatment arm, and highly underreported by men in the control and cash only groups. For instance, 33% of the cash only group admitted to gambling during validation, compared to 13% during the survey. Some of this underreporting could be due to ambiguous behaviors being coded as gambling in validation interviews but not in the survey. But the fact that underreporting is smaller in the therapy arms suggests that the underreporting is not simply an artifact of different definitions, but rather something else, such as a strategic response to treatment status.

If we look at stealing, marijuana use, and homelessness, however, none of the survey-validation differences are statistically significant. There is possibly some slight underreporting of drug use and slight over-reporting of stealing, but the magnitudes are generally small in the sense that they are less than 10% of the survey means reported in Table 4. The sample size is small, however, and hence many of these differences are not precisely estimated.

5.1.4. Underreporting of expenditures

We see much stronger evidence of underreporting of expenditures in the survey. The difference for the combined expenditures is −0.27 in the

[30] Direct observation was more likely to support findings of underreporting than of over-reporting, but this is mechanical since it's not possible to observe a non-behavior. This does not apply to homelessness, however, since the absence of homelessness is having a home, and this is observable.

Table 7
Estimates of treatment effects ($\hat{\theta}$) and treatment effect bias (θ^{Δ}) by outcome and treatment.

(a) Round and block fixed effects (N = 3765/239)

Treatment arm	Outcome index	ATE using survey data ($\hat{\theta}$)		ATE on measurement error (θ^{Δ})		Adjusted ATE ($\hat{\theta} - \theta^{\Delta}$)	
		β	S.E.	β	S.E.	β	S.E.
		(1)	(2)	(3)	(4)	(5)	(6)
Cash only	Sensitive	−0.057	[.095]	−0.178	[.190]	0.121	[.195]
	Nonsensitive	0.080	[.052]	0.285	[.130]**	−0.205	[.143]
Therapy only	Sensitive	−0.186	[.092]**	0.004	[.199]	−0.190	[.198]
	Nonsensitive	0.000	[.050]	0.335	[.134]**	−0.334	[.154]**
Therapy and cash	Sensitive	−0.398	[.090]***	0.118	[.182]	−0.516	[.196]***
	Nonsensitive	0.076	[.050]	0.314	[.116]***	−0.239	[.134]*

(b) Baseline controls and survey round/block fixed effects (N = 3765/239)

Treatment arm	Outcome index	ATE using survey data ($\hat{\boldsymbol{\theta}}$)		ATE on measurement error ($\boldsymbol{\theta}^{\Delta}$)		Adjusted ATE ($\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\Delta}$)	
		$\boldsymbol{\beta}$	S.E.	$\boldsymbol{\beta}$	S.E.	$\boldsymbol{\beta}$	S.E.
		(1)	(2)	(3)	(4)	(5)	(6)
Cash only	Sensitive	−0.078	[.071]	−0.313	[.177]*	0.234	[.268]
	Nonsensitive	0.081	[.045]*	0.147	[.136]	−0.066	[.195]
Therapy only	Sensitive	−0.193	[.071]***	−0.215	[.180]	0.022	[.264]
	Nonsensitive	−0.009	[.046]	0.268	[.144]*	−0.277	[.213]
Therapy and cash	Sensitive	−0.402	[.069]***	0.052	[.183]	−0.454	[.270]*
	Nonsensitive	0.073	[.045]	0.267	[.122]**	−0.195	[.197]

Notes: The survey-based average treatment effect (ATE) estimates, θ, pool all survey rounds and regress each outcome on treatment indicators. Standard errors are robust and clustered by individual. Estimates of the bias from treatment, θ^{Δ}, come from a regression of the difference in the survey and validation measures on an indicator for treatment arms. Standard errors are robust and clustered by block. The difference, $\hat{\theta} - \theta^{\Delta}$, is an estimate of the true treatment effect after adjusting for observed bias. It is calculated as the linear difference of the estimates and the standard error is calculated via bootstrapping (we performed 1000 draws from the sample, with replacement; we calculated $\hat{\theta}$, θ^{Δ} and $\hat{\theta} - \theta^{\Delta}$ for each draw; and we generated the standard error on $\hat{\theta} - \theta^{\Delta}$ using the distribution from these draws).

*** $p < 0.01$.
** $p < 0.05$.
* $p < 0.1$.

full sample (Table 5, Column 6). This difference is large—about a third of the survey mean reported in Table 4. Expenditure underreporting is largest for the video club measure, but both expenditures appear to be underreported. Interestingly, the mean differences appear to be smaller and less statistically significant if the men received one of the treatments. We return to these differences across treatment arms below.

5.2. Correlation between treatment and measurement error

In order to identify if measurement error is correlated with treatment, we estimate Eqs. 11 and 12 in Table 6. For sensitive behaviors, almost none of the coefficients on treatment indicators or interactions are statistically significant. For the index of four sensitive measures (Panel (a), Column 5), the coefficient on treatment, β_1, is actually greater than zero for therapy plus cash, implying that the impacts of therapy plus cash are, if anything, larger than the survey data imply. That said, the confidence intervals are relatively large, so we cannot rule out overstatement of treatment effects entirely. Nonetheless, there is almost no evidence of the bias we feared.

The results for our two expenditure measures suggest that all treatment arms are associated with a roughly 0.3 increase in the total number of instances (out of 2) in our proxy for measurement error (Panel (a), Column 8). There is underreporting across all arms, but it is greatest in the control group. As we see below, this implies that any expenditure gains we observe from the interventions may be the result of misreporting.

Before looking at these adjusted treatment effects, we consider the results of the more flexible regression in Panel (b). It does not materially change our conclusions. The effect on sensitive behaviors, in particular, is fairly homogeneous. Treated men who we think did not engage in the sensitive behaviors tend to report them ($\tilde{\beta}_1^{Both} > 0$) more than untreated men. Treated men who did engage in the sensitive behaviors tend to under report to a lesser extent than in the control group.[31] Put differently, treated men reported a slightly higher rate of sensitive behaviors regardless of their measurement in the validation exercise. Of course, this more flexible test has much lower statistical power given our number of observations, and so we interpret it with caution.

5.3. Adjusting treatment effects

5.3.1. Treatment effects using survey data

Blattman et al. (2015) report full treatment effect estimates, short-term and long-term, based on the survey data. These results indicate that cash (alone or in combination with therapy) led to an increase in consumption in the month after the grant, including a fall in homelessness, in part because the men spent the grant directly, but also because they invested in petty business and increased their earnings. After a year, however, these earnings and consumption gains had disappeared, likely because adverse economic shocks eliminated the men's additional cash, savings, and investments.

Therapy, meanwhile, led to self-reported falls in anti-social behaviors ranging from 30% to 50%, especially in interpersonal aggression, drug dealing, and theft. After a month, the falls were similar in both the therapy only and therapy plus cash groups. After a year, however, the fall was only sustained in the therapy plus cash group. The paper hypothesizes that therapy plus cash had a more sustained effect on anti-social behaviors because the cash grant positively reinforced the behavior change and enabled the men to practice their new skills and

[31] Also, note that, on average, $\tilde{\beta}_0 > 0$, $\tilde{\beta}_2 < 1$, and $\tilde{\beta}_0 + \tilde{\beta}_2 < 1$ for sensitive measures (Column 5). This is consistent with what we observe in Table 4: slight survey underreporting of sensitive behaviors, and 20–30% non-correspondence between survey and validated measures.

carry on with their new identity for longer. This large, sustained fall in self-reported anti-social behaviors in the therapy plus cash group is the central finding of the study.

We see the same patterns reflected in our pooled (over time) treatment effects on the sensitive and non-sensitive summary indexes, still focused on the survey data, which are displayed in Table 7. Cash weakly increased expenditures (our nonsensitive index) but had little effect on our sensitive behaviors index. The increased expenditures are driven mainly by short term impacts. Meanwhile, therapy decreased sensitive behaviors such as stealing and gambling. With therapy plus cash, the effects are largest and more statistically significant, in large part because they are sustained in the long run.

5.3.2. Adjusted treatment effects

Table 7 also reports the effect of each treatment on survey over-reporting, θ^{Δ}. These estimates effectively take the simple survey-validation differences in Panel (a) of Table 6 and estimate the difference across treatment arms. We present two cases: adjusting for survey round and randomization block fixed effects (Panel a), and adjusting for baseline covariates as well (Panel b). We use these to calculate an adjusted treatment effect, $\hat{\theta} - \theta^{\Delta}$, for our sensititve and nonsensitive indexes.[32]

First, the results imply that the true treatment effect of therapy plus cash on sensitive behaviors is no smaller than what we estimate with self-reported survey data. Indeed, the fall in these sensitive behaviors may even be greater than the survey reports suggest (Columns 5 and 6). This holds true for each of the individual sensitive behaviors, which are shown in Appendix C.3. Despite the large standard errors introduced by the small validation sample, the adjusted treatment effect on an index of all sensitive behaviors is larger in absolute value and significant at the 1% level (when we use round and block fixed effects, in Panel a). When baseline controls are added in Panel b, the coefficient is similar but the standard errors have increased and it is statistically significant at the 10% level only. This is partly because our validation sample is small, and the number of baseline controls is large, reducing the degrees of freedom.[33]

In contrast, adjusting expenditures changes the sign of the treatment effect we estimated using survey data, and hence affects our conclusions about the intervention. Based on the survey data, we estimated that the cash grant led to a short term increase in expenditures. But the slight underreporting of expenditures, especially the excess under reporting by the control group, may have exaggerated the effects of cash on expenditures and incomes (judging by these two expenditures at least). The adjusted (pooled) treatment effect on expenditures is negative for all treatments and both specifications, generally with nontrivial magnitudes but only statistically significant in a couple of instances (Columns 5 and 6).[34]

6. Discussion and conclusions

Perhaps the most important lesson from this exercise is that structured, in-depth, and representative qualitative work revealed patterns of measurement error that were quite different from our priors, despite extensive experience with the study group. There is little data on measurement error, however, and so (like many) our priors were unavoidably rooted more in what seemed like common sense and in common causal identification concerns rather than an informed understanding.

We worried, for instance, that high-risk young men might have special reasons to conceal their behavior—such as suspicion of outsiders, or a desire to receive program benefits in the future. Given that we were focused on measuring the treatment effects of a therapy program that discouraged various anti-social and unhealthy behaviors, we were also concerned that the treated would underreport such behaviors out of experimenter demand or social desirability bias induced by the therapy. Our multi-method approach revealed that the nature of measurement error was quite different.

For this specific field experiment, two findings stand out. First, the qualitative validation suggests that the underreporting in sensitive behaviors was modest, not statistically significant, concentrated in the control and cash only groups, and limited to one behavior in particular (gambling). Meanwhile, expenditures seemed to be broadly underreported in the survey, most of all in the control group.

Based on qualitative interviews, our impression is that these "sensitive" behaviors, while not acceptable within Liberian society as a whole, are not so stigmatized that most men in our sample feel ashamed to report them, perhaps because these men belong to a counterculture in which these activities are common. Moreover, the risk of punishment was minuscule.[35] Hence underreporting tended to be modest overall.

An exception was gambling. Gambling, unlike the other sensitive behaviors, is not a defining characteristic of the counterculture to which study participants belong. Furthermore, after receiving a cash handout, it's possible that men were reluctant to admit they'd gambled some of it away. The same could be true of the control group, to a lesser degree, who may have hoped for cash in future.[36] Alternatively, the therapy treatment could have increased the familiarity, trust or reciprocity between the subjects and implementers, and so men who received therapy were less likely to underreport.

The second major finding is that the expenditure-related activities were systematically underreported across all arms, and especially in the control group. The effect of treatment on measurement error is large and statistically significant in all arms. This finding is extremely important given that expenditure and consumption surveys are the principal means of measuring material well-being and poverty in most developing countries. We see two main possible explanations:

1. *Strategic behavior*. Since there was underreporting across all treatment arms, every study participant may have had an incentive to exaggerate their neediness in the hopes of future programs. This echoes our gambling result.
   Why more so in the control group? It's possible that the fewer treatments a man received, the more strategically he behaved on the survey, trying to appear poorer to encourage eligibility for future treatment. Those who received therapy, for example, might be interested in the cash. Phone charging and going to a video club are considered discretionary spending, and if a respondent wanted to signal destitution, he might underreport spending on these items.
   We view this explanation as plausible, although there are caveats. First, the control group did not over-report homelessness to the same degree, which is an obvious indicator of need (although perhaps observable enough that it was perceived as harder to falsify on a survey). Second, while drug use is technically an expenditure, this was not underreported to signal poverty. (One reason may be that the drug users were mainly heavy marijuana users, indeed so heavy that this was somewhat obvious, thus potentially making it

[32] Recall that $\theta^{\Delta} \rightarrow E(\hat{\theta} - \theta)$ as the validation measure approaches the "truth". Appendix C.3 contains the results for each component of these indexes.

[33] Meanwhile, the underreporting of gambling (displayed in Appendix C.3) does not have a statistically significant association with treatment. However, those who received cash alone underreported gambling to the surveyors more often than control group members, and so the measurement error in gambling is probably a combination of a general desirability bias as well as one correlated with treatments. A larger sample size would be needed to separate these more precisely.

[34] We see a similar pattern with another expenditure-related item, homelessness, in Appendix C.3—the survey-reported decline in homelessness tends to lose a lot of its significance with adjustment.

[35] The Liberian police are largely incapable of investigating and prosecuting all but the most grave crimes. Thus, these behaviors are not endangering to most of our sample and their peers, and they discuss them freely.

[36] There was no future program (this was communicated repeatedly), and the original field experiment actually used countervailing criteria for recruiting subjects, but these features of our program were unusual compared to standard NGO practice.

less prone to falsification.) Third, in principle those who received one of the earlier treatments also had incentives to behave strategically in the hopes of future programs. Treated men almost universally lobbied for additional assistance.

2. *Salience and recall bias*. Expenditures could be more subject to recall error, because they are less regular and possibly less salient than drug use or crime. There is ample evidence that consumption and expenditure data are underreported, and that underreporting increases with the period of recall, the lower the reported consumption per standardized unit of time, and the less salient the purchase (Beegle et al., 2012; Deaton and Grosh, 1997; Gibson, 2006). Neither video clubs nor mobile phone charging were particularly salient. People may also make cognitive errors when aggregating over a construct such as "the last two weeks." Finally, the expenditures survey module was long and much more subject to fatigue, compounding underreporting.
   Recall bias is plausible, but we are also looking for explanations that would correlate with treatment. There are a few possibilities. Treatment could have increased attention and mindfulness. The therapy was explicitly designed to reduce impulsive behavior and to increase planning. There is some evidence that impulsivity in fact decreased (Blattman et al., 2015). The cash transfer could have had a similar effect for different reasons. Studies have also shown that recall bias in consumption data increases with poverty (Beegle et al., 2012). This is consistent with evidence that cognition is taxed by poverty and scarcity (Mani et al., 2013). Presumably hunger would affect survey fatigue and mindfulness. The cash grant (and short run decrease in poverty) could have had a similar effect on the margin. Finally, receiving either treatment could have produced enough reciprocity that the treated group exercised more care in recalling less salient data. We regard these explanations with caution, but cannot reject them.

Both explanations are plausible but come with caveats, and so we refrain from a firm conclusion about the sources of measurement error. Given the importance of expenditure surveys in research it bears replication and further research.

In retrospect, we also see that, had the measurement error run in the opposite direction, it would have been difficult to distinguish the large effect of therapy and cash on crime from systematic measurement error given our sample size. The effect of treatment on our proxy for survey over-reporting would have been underpowered. We estimate that doubling the size of the validation sample would have increased power by about a third. The marginal cost of validation per respondent was roughly equal to that of running a survey. Thus we estimate that we could have doubled the number of validations by either increasing the evaluation budget by 3%, or reducing the total sample size by 3%. Given how much the credibility of these types of studies rests on self-reported data, this strikes us as a reasonable investment.

Overall, these results reinforce a fundamental principle of survey methodology: the importance of validating measurements with multiple instruments. To some extent our findings are unexpected (even puzzling), and the explanations are somewhat speculative. With more validation studies, of all kinds, we may start to see systematic patterns. We regard our multi-method approach as largely complementary to list (item count), random response, and endorsement experiments. It is useful to have more methods available.

Like other methods, ours requires a priori assumptions—in this case, that in-depth observation is less prone to bias and does not introduce major new biases. Our method is also more costly to implement, though not necessarily relative to the average cost of large surveys or modest impact evaluations. The stakes are high enough in many panel studies, field experiments, and other impact evaluations that validating a handful of key outcomes seems important for the individual project. Qualitative validation performs well enough, and yields sufficiently important results, that our approach deserves more systematic use and examination, ideally alongside these other methods.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jdeveco.2016.01.005.

References

Asher, H.B., 1974. Some consequences of measurement error in survey data. Am. J. Polit. Sci. 18 (2), 469–485.
Baruch, G., 1981. Moral tales: parents' stories of encounters with the health professions. Sociol. Health Illn. 3 (3), 275–295.
Beegle, K., De Weerdt, J., Friedman, J., Gibson, J., 2012. Methods of household consumption measurement through surveys: experimental results from Tanzania. J. Dev. Econ. 98 (1), 3–18.
Blair, G., Imai, K., 2012. Statistical analysis of list experiments. Polit. Anal. 20 (1), 47–77.
Blattman, C., Jamison, J., Sheridan, M., 2015. Reducing crime and violence: experimental evidence on adult noncognitive investments in Liberia. Working Paper.
Bound, J., Brown, C., Mathiowetz, N., 2001. Measurement error in survey data. Handb. Econ. 2, 3705–3843.
Bryman, A., 2003. Quantity and Quality in Social Research. Routledge, New York.
Bullock, W., Imai, K., Shapiro, J.N., 2011. Statistical analysis of endorsement experiments: measuring support for militant groups in Pakistan. Polit. Anal. 19 (4), 363–384.
de Mel, S., McKenzie, D.J., Woodruff, C., 2009. Measuring microenterprise profits: must we ask how the sausage is made? J. Dev. Econ. 88 (1), 19–31.
Deaton, A., Grosh, M., 1997. Consumption. Designing Household Survey Questionnaires for Developing Countries: Lessons from Ten Years of LSMS Experience.
Deming, D.J., 2011. Better schools, less crime? Q. J. Econ. 126 (4), 2063–2115.
Fox, R.C., 2004. Observations and reflections of a perpetual fieldworker. Ann. Am. Acad. Pol. Soc. Sci. 595 (1), 309–326.
Freedman, D.A., 1991. Statistical models and shoe leather. Sociol. Methodol. 21 (2), 291–313.
Gibson, J., 2006. Statistical tools and estimation methods for poverty measures based on cross-sectional household surveys. Handbook on Poverty Statistics.
Hausman, J., 2001. Mismeasured variables in econometric analysis: problems from the right and problems from the left. J. Econ. Perspect. 15 (4), 57–67 October.
Jamison, J.C., Karlan, D., Raffler, P., 2013. Mixed-method evaluation of a passive mHealth sexual information texting service in Uganda. Inf. Technol. Int. Dev. 9 (3).
Karlan, D., Zinman, J., 2008. Lying about borrowing. J. Eur. Econ. Assoc. 6 (2–3), 510–521.
Karlan, D.S., Zinman, J., 2012. List randomization for sensitive behavior: an application for measuring use of loan proceeds. J. Dev. Econ. 98 (1), 71–75.
LeCompte, M.D., 1987. Bias in the biography: bias and subjectivity in ethnographic research. Anthropol. Educ. Q. 18 (1), 43–52.
LeCompte, M.D., Goetz, J.P., 1982. Problems of reliability and validity in ethnographic research. Rev. Educ. Res. 52 (1), 31–60.
Mani, A., Mullainathan, S., Shafir, E., Zhao, J., 2013. Poverty impedes cognitive function. Science 341 (6149), 976–980.
Martinelli, C., Parker, S.W., 2009. Deception and misreporting in a social program. J. Eur. Econ. Assoc. 7 (4), 886–908.
McKenzie, D., 2012. Beyond baseline and follow-up: the case for more T in experiments. J. Dev. Econ. 99 (2), 210–221.
N.S.S.O. Expert Group, 2003. Suitability of different reference periods for measuring household consumption. Results in pilot survey. Econ. Polit. Wkly. 38 (4), 25–31.
Paluck, E.L., Green, D.P., 2009. Deference, dissent, and dispute resolution: an experimental intervention using mass media to change norms and behavior in Rwanda. Am. Polit. Sci. Rev. 103 (4), 622–644.
Power, R., 1989. Participant observation and its place in the study of illicit drug abuse. Br. J. Addict. 84 (1), 43–52.
Raghavarao, D., Federer, W.T., 1979. Block total response as an alternative to the randomized response method in surveys. J. R. Stat. Soc. Ser. B Methodol. 40–45.
Ram, P.K., Halder, A.K., Granger, S.P., Jones, T., Hall, P., Hitchcock, D., Wright, R., Nygren, B., Islam, M.S., Molyneaux, J.W., 2010. Is structured observation a valid technique to measure handwashing behavior? Use of acceleration sensors embedded in soap to assess reactivity to structured observation. Am.J.Trop. Med. Hyg. 83 (5), 1070–1076.
Scacco, A., 2010. Who Riots? Explaining Individual Participation in Ethnic Violence (*Dissertation*) New York University.
Spitzer, R.L., Kroenke, K., Williams, J.B.W., 1999. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. J. Am. Med. Assoc. 282 (18), 1737–1744.
Wilson, S., 1977. The use of ethnographic techniques in educational research. Rev. Educ. Res. 47 (1), 245–265.