Can We Trust Shoestring Evaluations?

Martin Ravallion

Many more impact evaluations could be done, and at lower unit cost, if evaluators could avoid the need for baseline data using objective socio-economic surveys and rely instead on retrospective subjective questions on how outcomes have changed, asked post-intervention. But would the results be reliable? This paper tests a rapid-appraisal, "shoestring" method using subjective recall for welfare changes. The recall data were collected at the end of a full-scale evaluation of a large World Bank supported poor-area development program in China. Qualitative recalls on how living standards have changed are found to provide only weak and biased signals of the changes in consumption as measured from contemporaneous surveys. Importantly, the shoestring method was unable to correct for the selective placement of the program favoring poor villages. The results of this case study are not encouraging for future applications of the shoestring method, although similar tests are needed in other settings. JEL codes: C81, H43

There are a great many interventions that we would like to evaluate for which no (pre-intervention) baseline data are available. Think of all the projects for which no impact evaluation was ever planned.[1] In the absence of baseline data we cannot do the standard "double-difference" estimator—comparing outcome changes since the baseline between treated and untreated units.

But there is a potential way out: We can ask post-intervention questions of both the treatment and comparison groups on how much their welfare has improved since the intervention began. This would dramatically lower the costs of impact evaluations—an example of what Bamberger et al. (2004) call

Martin Ravallion holds the Edmond D. Villani Chair of Economics at Georgetown University. His email address is mr1185@georgetown.edu. This paper builds on a long-term evaluation developed by the author and Shoahua Chen at the World Bank, in collaboration with staff of the Rural Survey Organization of China's National Bureau of Statistics, which implemented the survey data collection. The author is also grateful to Solveig Buhl, then of Deutsche Gesellschaft für Technische Zusammenarbeit (GTZ), for comments on the recall module developed for this paper, and help in field testing and refining the module. Funding for the study was provided by the World Bank's Knowledge for Change Trust Fund. For their comments, the author is grateful to Kathleen Beegle, Gero Carletto, David McKenzie, Dominique van de Walle, and the Review's editors and anonymous referees.

1. For example, while there has been a substantial growth in impact evaluations of World Bank development projects, only 8.8 percent of its investment loans in 2009/10 had an impact evaluation. (In 1999/00 it was 2.4 percent.)

THE WORLD BANK ECONOMIC REVIEW, VOL. 28, NO. 3, pp. 413–431 doi:10.1093/wber/lht016
Advance Access Publication June 19, 2013

"shoestring evaluations." And it would open up many new opportunities for learning about policy effectiveness. It could be especially helpful in addressing a common problem in impact evaluations of development projects, namely that the time period is often constrained to fall short of the period in which the full impact is to be expected (King and Behrman, 2009; Ravallion, 2009). For example, in evaluating donor-financed operations it can be hard to ensure that the impact evaluation extends far enough beyond the disbursement period to credibly capture the impacts.[2] Yet for certain types of development projects—including infrastructure—longer-term impacts are expected.

There have been examples of the use of retrospective questions to create "instant longitudinal data" (Janson, 1990).[3] However, we know very little about the method's performance in impact evaluations, where the interest is in comparing results from two samples, one treated and one not. The limited references one finds to the idea in the literature on evaluation appear to be encouraging. In their book, *Real World Evaluation*, Bamberger et al. (2006) identify recall as one of the methods available for reconstructing baseline data ". . . to obtain estimates of major changes in the welfare conditions of the household" (98) and they provide examples.[4] They offer a cautiously positive assessment of this method:

> "Recall is a potentially valuable, although somewhat treacherous, method to retroactively estimate conditions prior to the start of the project and hence to reconstruct or strengthen the baseline data. Although the literature on the reliability of recall is quite limited, particularly in developing countries, available evidence suggests that although information from recall is frequently biased, the direction, and sometimes magnitude, of the bias is often predictable . . . so that usable estimates can often be obtained." (Bamberger et al., 2006, 98.)

How confident can we be about the potential for using retrospective recall of outcome changes as a proxy for the actual changes in an impact evaluation?[5] Some observers have argued that long-term recall of changes in the overall standard of living provides a usable signal. For example, Narayan, Pritchett and Kapoor (2008) use a 10 year recall period for changes in living standards in studying poverty dynamics in developing countries. Krishna (2004) and Krishna et al. (2006) use a 25 year recall period for essentially the same purpose. There have been very few tests of long recall, but in one of the few examples, Berney and Blane (1997) found evidence that 50 year (!) recall of relatively simple

2. This arises from the externalities in evaluation, given that main support for the evaluation typically comes from the manager of that specific project, while benefits accrue more broadly. See Ravallion (2009).

3. On the strengths and weaknesses of such designs see Featherman (1908), Janson (1990) and Solga (2001).

4. Also see Broegaard et al. (2011). Retrospective recall of baseline information has also been used in evaluative medical research. See, for example, Watson et al. (2007) and McPhail and Haines (2010).

5. One might argue that recall of changes in subjective welfare is of intrinsic interest, even if it does not accord well with reality. For example, this is argued by Narayan et al. (2008, p.8). That may well be, but here it is assumed that one is only interested in recall as a proxy for missing data on outcomes, for the purpose of an impact evaluation where there is an objective outcome measure appropriate to the specific project.

information (father's occupation, type of dwelling, number of rooms, water and sanitation facilities) was quite reliable in a small sample of British adults.

Yet the literature is also replete with warnings on how unreliable retrospective studies can be, given the limitations of human memory. "Telescoping" is thought to be common, whereby important events are remembered reasonably well but placed at the wrong time; errors of both omission and commission also occur.[6] Recall of precise quantities, such as food consumed, is unlikely to be reliable over more than a month or so. A degree of recall failure is to be expected although it should not be presumed that longer recall periods necessarily give less accurate answers. That depends on what one is asking about. Longer term recall of changes in overall standard of living may well be more reliable than for (say) quantities of food consumed.

However, the issue here is not whether people can recall well how they lived 10 years ago (say), but rather whether such data are reliable for inferring the impacts of an intervention in the absence of baseline data. In past applications of long-term recall, the precise period of time is not so important. In an impact evaluation, telescoping could be a more serious concern since one wants to know welfare changes since the precise time the project started. And the date at which a program intervention occurred cannot be expected to generally coincide with some common and memorable event to aid recall.[7] The reliability of recall will depend on the time profile of benefits from the specific project. Given the scope for telescoping, it will clearly make a difference whether those benefits are evenly spread over time or concentrated in some sub-period.

There is another reason why we might be concerned about the reliability of this shoestring method. The questions asked are likely to be subjective-qualitative questions; indeed this is seen as recommended practice, on the presumption that recall of quantitative data is unreliable (Bamberger et al., 2006). Such methods can also reduce the cost of the evaluation; development outcomes such as consumption or income require relatively complex and costly surveys. However, the literature on subjective welfare ("well-being") points to concerns about this type of data, especially when used as a dependent variable, as here. If we could assume that the errors are white noise then they will not create bias, although they may make it harder to obtain precise estimates of impacts. However, there are reasons to expect systematic effects. The fact that these are typically subjective data on outcomes suggests that non-ignorable measurement errors and personality/mood effects on self-assessed welfare will be present, and there are good reasons to expect the errors in subjective data to be correlated with other explanatory variables (Bertrand and Mullainathan, 2001; Ravallion and Lokshin, 2001).

6. Janson (1990) reviews the evidence on recall errors. For useful overviews of these and other issues in survey design also see Fowler (1995) and Iarossi (2006).

7. For example, McIntosh et al. (2011) used fundamental events in the history of each sampled household to help improve recall for a situation in which take-up of the intervention varied (endogenously) across households. However, in most evaluations the intervention starts at the same date for all participating households.

There is also the possibility that the intervention alters the scales used by respondents to in subjective questions—such as what it means to be "poor" or "very satisfied" with life—thus biasing the results even with perfect recall. People will naturally interpret the scales used in a subjective question on welfare relative to their personal knowledge and experience, which might well be influenced by the intervention. There is evidence of systematic effects of respondent characteristics on how scales are interpreted in subjective questions (Beegle et al., 2012).

The upshot of these observations is that subjective responses on outcomes must be expected to contain statistically non-ignorable noise for the purposes of an impact evaluation.[8] If program placement was random and impacts common across all units then one would not be concerned, although heterogeneous impacts cloud the picture, even in an experiment.[9] In non-experimental evaluations, biases can be expected even under homogeneous impacts.

However, the literature on policy or program evaluation does not contain (to my knowledge) even a single example in which this type of baseline recall has been tested against conventional survey data collected at both the baseline and post-intervention.

This paper tries to help fill this gap in our knowledge. The paper reports on an experiment that was designed to test the idea of using retrospective data as a substitute for baseline data from a contemporaneous survey. After collecting baseline and post-intervention data for treatment and observationally comparable comparison units to allow estimation of a standard double-difference (DD), a series of recall questions were asked on how various dimensions of welfare had changed since the time the project was introduced. This allows what can be called the "shoestring double difference" (SDD) estimator. More precisely, the two estimators of mean impacts are:

$$DD \equiv E(\Delta Y_i | i \in T) - E(\Delta Y_i | i \in C) \tag{1}$$

$$SDD \equiv E(R_i | i \in T) - E(R_i | i \in C). \tag{2}$$

Here $\Delta Y_i = Y_{i1} - Y_{i0}$ is the measured change in consumption between the baseline (date 0) and post-intervention surveys (date 1) for the respondent in household i, where each respondent is assigned to either the treatment (the set T) or comparison group (C), and R_i denotes the subjective recall of the change in living standards over the same period.

8. This is well recognized in the literature on using subjective welfare data in economics, where the focus is on the regression function of subjective welfare on covariates rather than the actual values reported by respondents. For an overview of the literature see Ravallion (2012).

9. For example, suppose that the true impacts of a project are greater for poor people, for whom recall is less reliable. (There is supportive evidence for this conjecture in Das et al., 2011, who found health-status recall to be worse for poor people, using data for India.) Then we will obtain a biased estimate of the difference in impacts between poor and non-poor people even with randomized assignment of the intervention.

Two versions of the SDD estimator are studied here:

SDD1: This assumes that no baseline data are available. Only an ex-post survey can be done. Thus no adjustments are made for selection bias based on contemporaneously observed pre-intervention differences that might influence subsequent trajectories.

SDD2: This assumes that only the data on outcomes are missing. Thus standard corrections can be made for selection bias based on other observables at the baseline.

Note that the difference is in whether an allowance is made for selection on observables. If the recall of changes since the introduction of the project works well then both SDD1 and SDD2 will be able to address selection based on time-invariant unobserved factors.[10]

Importantly, the shoestring evaluations were "tacked onto" a full scale evaluation. This was for a large antipoverty program in poor areas of rural China, and the results are reported in Ravallion and Chen (2005) and Chen, Mu and Ravallion (2009). The benchmark evaluation was non-experimental, and so it could not eliminate time-varying selection bias based on unobserved factors, although its longitudinal design did allow it to address bias due to latent time-invariant factors. The evaluation did, however, put considerable effort into balancing a wide range of observed characteristics between the treatment and selected comparison villages, as documented in Chen et al. (2009). The present paper is thus able to compare SDD1 and SDD2 to the "actual" DD, as estimated from high-quality, comprehensive and contemporaneous baseline and follow-up surveys. The implications for the structure of recall errors are also examined.

The benchmark evaluation found that the average impact of the program on consumption and income was not significantly different from zero over the period as a whole, although there were significant income gains during the disbursement period; the bulk of these gains were saved and spread over a long period. So the present test of the shoestring evaluation method is done against a benchmark with little or no average impact.

The findings from this case study suggest that SDD methods are vulnerable to biases that confound identification. The shoestring method reproduces well the zero average impact finding of the benchmark evaluation. However, when one looks more closely at the structure of recall errors it becomes clear that the shoestring method has performed poorly. Respondents' perceptions of how their living standards have changed are found to provide a weak and biased signal of consumption changes measured from contemporaneous surveys. There are also signs of "false positives" stemming from the weak ability of retrospective recall of welfare changes to neutralize selection bias based on unobserved initial conditions influencing program placement.

10. On the distinction between selection bias based on observables and that based on unobserved factors see Heckman et al. (1998).

The following section describes the project and data. Section II presents the impact estimates, while section III explores the responses on recall further to help understand the results in section II. Section IV concludes.

I. Setting, Data, and Methods

The project being evaluated is the World Bank's Southwest China Poverty Reduction Project—the Southwest Program (SWP) for short. This comprised a package of multi-sectoral interventions targeted to poor villages using community-based participant and activity selection. The aim was to achieve a large and sustainable reduction in poverty. The project was implemented in selected poor villages in the designated poor counties of Guangxi, Guizhou, and Yunnan. The total investment per capita under the SWP was roughly equal to mean annual income per capita of the project villages.

Within the selected villages, virtually all households were expected to benefit from the infrastructure investments under SWP, such as improved rural roads, power lines and piped water supply. Widespread benefits were also expected from the improved public services, including upgrading village schools and health clinics, and training of teachers and village health-care workers. Those with school-aged children also received tuition subsidies as long as the children stayed in school. Over half of the households in SWP villages also received individual loans at a lower interest rate than for commercial sources of credit. The loans financed various activities including initiatives for raising farm yields, animal husbandry and tree planting. There was also a component for off-farm employment, including voluntary labor mobility to urban areas and support for village enterprises. The selection of project activities aimed to take account of local conditions and the expressed preferences of participants.

Chen, Mu and Ravallion (2009) report results from an intensive survey data collection effort over 1995–2005 spanning both treatment and comparison villages.[11] All surveys were implemented by the Rural Household Survey (RHS) team of the government's National Bureau of Statistics (NBS). The baseline survey covered 2,000 randomly-sampled households in 200 villages, with roughly half not participating in the SWP. A final post-intervention survey was done in 2004/05. Surveys were also done during the disbursement period up to its end, in 2000.

There are 112 SWP villages and 86 non-SWP villages in the sample. The SWP villages were a random sample from all project villages, while the non-SWP villages were a random sample from all other villages in the designated poor counties. Ten randomly sampled households were interviewed in each village.

The surveys included community, household, and individual questionnaires. The community schedule collected data on natural conditions, infrastructure and

11. The attrition rate was 12 percent over the full period. Chen et al. discuss tests for attrition bias and for bias in selecting replacement sample households.

access to services. The household survey collected data on (*inter alia*) incomes, consumptions and assets. The individual questionnaires covered gender, age, education, and occupation.

Relative to other household surveys, unusual effort went into obtaining accurate data on consumption and income. While the community, individual and project activity surveys used conventional one-time interviews each year, the household surveys were quite different. The surveys were closely modeled on NBS's Rural Household Survey (RHS) (which is described in detail in Chen and Ravallion, 1996). This is a good quality budget and income survey, notable in the care that goes into reducing both sampling and non-sampling errors. Similarly to the RHS, sampled households maintain a daily record on all transactions plus log books on production. Local interviewing assistants visited each household at two-three weekly intervals to monitor compliance and check questionable data entries or inconsistencies found at the local (county-level) NBS office. Other trained interviewers also visited at regular intervals to collect additional data. This intensive interviewing method is in marked contrast to most surveys in which the respondent is visited only once or twice.

The consumption aggregate was built up from very detailed data on cash spending on all commodities and imputed values of consumption from own household production, valued at local selling prices. Living expenditures exclude spending on production inputs (which are accounted for in net income from own-production activities).[12] The income aggregate includes cash income from all sources and imputed values for in-kind income. Income is measured net of all production costs, including interest on debt (including loans from the SWP). The out-migrating workers were not tracked, although the income aggregate includes remittances received from family members who migrated, including those supported by the SWP. Remittances are expected to be the main means by which the out-migration component reduced poverty in the short run.

For the 2004/05 follow-up survey, exactly the same survey instrument was used as for the prior surveys. However, toward the end of the period, a rapid-appraisal module was designed by the author and refined based on field testing. The Chinese and local language versions of the module were refined on the basis of field tests in poor villages in a number of locations.[13] For the purpose of the present paper, in 2005 the module was added to the final survey of treatment and comparison samples in the SWP evaluation to elicit perceptions of how welfare had changed over time since the project began. The module asked respondents to assess whether various aspects of their lives had improved over the preceding 10 years. (There was

12. Living expenditures exclude transfer payments, although these only account for a small share of total spending (3.7 percent over the whole sample in 1996).

13. In the development stage for the module, the first field testing was done over two days in two selected poor villages in Jiangxi, and then revised. The module was then fields in 12 villages in Jiangxi, and further refined. The Jiangxi work was supervised by Solveig Buhl (GTZ staff member assigned to the provincial poor-area program office). The module was further tested and refined by staff of the national and provincial statistics offices in each of the three study provinces.

no obvious common event 10 years ago that could be used to help anchor the recall period in respondents' minds.) The questions related to a long list of aspects of well-being and in each case the respondent was asked whether this item had improved or not over the last 10 years, on a 5-point scale, "much worse," "slightly worse," "no different," "slightly better" and "much better." Matching questions were asked about perceived current standards of living. The sample was restricted to adults who were at least 28 years of age at the time of the interview.

The questions did not refer directly to the SWP since it was known from past survey experience in the SWP villages that no clear distinction was held by local residents between SWP and more general funding of local development through the county government. It should be recalled that all SWP villages (treatment and comparison group) were participants in the government's own long-standing poor-area development schemes. In practice this was essentially merged with SWP in treatment villages. Instead, the survey asked services to the village and responding household about services provided by the county government.

In measuring impacts for such a program one should allow for selection bias arising from how differences in initial characteristics influence subsequent trajectories; this is known to be important in poor-area development projects.[14] Chen et al. (2009) used propensity score (PS) weighting and trimming for this purpose. The method proposed by Hirano, Imbens and Ridder (2003) was used for PS weighting.[15] This allows for heterogeneity in the (observable) baseline characteristics that may be correlated with subsequent changes over time and so bias the DD results. The samples were also trimmed to assure sufficient overlap in propensity scores.[16] Of course, these adjustments for bias require the baseline data. Only the results without PS weighting and trimming would be feasible in a single survey round post-intervention.

In estimating the probits for whether a village was selected for SWP, covariates were chosen to reflect the selection criteria used by the project staff as well as the research team's priors on how other factors (such as remoteness and village ethnicity) may have influenced SWP placement. Chen et al. (2009) discuss the results in greater detail. They found that project villages tended to be in more hilly/mountainous areas, less well endowed with infrastructure, with lower mean income and consumption in the baseline. In most respects, the SWP villages tended to be poorer than other villages within the project counties.

Using the propensity scores based on the probit to re-weight the data, Chen et al. (2009) were able to obtain a close balancing of the characteristics of the two samples (including in the means of the initial outcome variables), particularly after trimming the samples, as is evident from Table 1 gives summary statistics on baseline village characteristics and household outcomes for both the

14. Jalan and Ravallion (1998) provide evidence using regional growth models for rural China.

15. For details see Chen et al., (2009).

16. For their "trimmed sample" Chen et al. (2009) chose the PS interval (0.1, 0.9), corresponding to the efficiency bounds recommended by Crump et al. (2006) for estimating average treatment effects with minimum variance.

TABLE 1. Balancing Tests for Baseline Village Characteristics and Household Outcomes Between Treatment and Comparison Villages

	Standardized means		Difference in standardized means					
			Un-weighted		PS weighted for total sample		PS-weighted for trimmed sample	
	SWP villages	Non-SWP villages	mean	s.e.	mean	s.e.	mean	s.e.
Baseline village characteristics								
Total population	0.009	-0.012	0.021	0.143	0.013	0.137	0.076	0.186
Electricity	−0.151	0.196	−0.347	0.141	−0.229	0.138	0.104	0.164
Phone	0.053	−0.069	0.122	0.143	0.109	0.141	0.155	0.168
Road	−0.061	0.079	−0.139	0.143	−0.094	0.141	0.211	0.164
Radio	0.044	−0.058	0.102	0.143	0.075	0.135	0.271	0.155
TV	−0.084	0.109	−0.193	0.142	−0.131	0.143	0.117	0.175
Nearest market < 5km	−0.036	0.047	−0.083	0.143	−0.068	0.148	0.078	0.187
Elementary school in village	−0.009	0.011	−0.02	0.143	−0.031	0.143	−0.075	0.182
Clinic in village	0.021	−0.028	0.049	0.143	0.051	0.141	0.043	0.170
Net income per capita	−0.162	0.211	−0.373	0.141	−0.241	0.142	0.073	0.164
Cultivated land per capita	0.134	−0.173	0.307	0.141	0.238	0.135	0.299	0.151
Baseline household outcomes								
Consumption per capita	−0.156	0.203	−0.36	0.141	−0.217	0.190	−0.069	0.181
Income per capita	−0.168	0.219	−0.39	0.141	−0.23	0.139	−0.182	0.181
Headcount poverty index for poverty line of:								
600 yuan (income)	0.175	−0.227	0.402	0.141	0.384	0.169	0.248	0.201
1000 yuan (income)	0.212	−0.277	0.489	0.139	0.41	0.202	0.165	0.201
600 yuan (consumption)	0.161	−0.21	0.371	0.141	0.455	0.186	0.259	0.198
1000 yuan (consumption)	0.171	−0.222	0.393	0.141	0.319	0.268	0.006	0.182

Notes: The standardized mean is (sub-group mean *minus* mean for full sample)/standard deviation for full sample. In total sample, there are 112 project villages' 86 comparison villages. In the trimmed sample, there are 71 project villages and 66 comparison villages. Household income, consumption and poverty measures are weighted by household size. Poverty lines are annual per capita in 1995 prices.

treatment and comparison villages, with and without trimming and matching. For compactness this is not a complete listing of all characteristics available in the data set. An Addendum is available with complete summary statistics for almost 70 characteristics; there is no sign of any significant difference in any attribute after matching and trimming. This appears to provide an observationally indistinguishable comparison for assessing impacts.

Note that these adjustments for selection bias are based on observable differences in the baseline. That still leaves any bias due to unobserved factors with time varying effects. Only selection bias due to (additive) time-invariant unobserved factors is removed using the time differencing component of the DD.

II. Impact Estimates

The findings from the full impact evaluation, as reported in Chen et al. (2009), are summarized in Table 2. The table gives DD estimates of the impacts of SWP on consumption and income for both the full sample and the sample trimmed for common support and using PS weighting. There is little or no impact of the SWP on consumption or income over the full period. This holds using a standard DD estimator as well as the PS weighted estimator on the trimmed sample.[17]

Chen et al. (2009) also provide impact estimates for 2000, at the end of the SWP's disbursement period. They found a significant impact on incomes during the disbursement period, but evidently this was all saved (as discussed further in Ravallion and Chen, 2005). Given this uneven spread of the impacts of SWP—concentrated in the earlier half of the study period—telescoping could well be a problem in using recall. The reliability of the SDD method will depend critically on the ability of respondents to recall the income gains over five years ago, and correctly identify those gains as being within the last 10 years.

Table 3 summarizes the findings from asking in 2005 whether various aspects of well-being had improved over the previous 10 years. (This is a complete listing of results for all the recall questions asked.) The first main column of the table gives the SDD1 estimator, which is the single difference between SWP villages and non-SWP villages in the proportion of the population saying that the item in question had "obviously improved" or better.[18] Note that since the question already embodies the change over time, the single difference can be interpreted as a double-difference estimate of the impact on the underlying level of that variable.

The subjective assessments by SWP participants of whether their living standards had improved since the project began are not significantly different to those found for the non-SWP villages. For example, 36 percent of those in the SWP villages reported that their overall standard of living had "obviously

17. Chen et al. (2009) also study the heterogeneity in impacts and find that SWP could have had substantially higher overall welfare impacts if it had been targeted differently. They also study spillover effects of the program, given behavioral responses of local governments.

18. Sensitivity was tested to using both a lower and higher cut-off; in neither case was there any significant difference between SWP villages and the comparison villages.

TABLE 2. Impact of SWP on Household Consumption and Income

	Baseline mean in SWP villages	Gain in treatment project	Gain in comparison villages	Double difference	t-ratio
Full sample					
Income	989.45	401.316	360.644	40.673	0.537
Consumption	843.559	287.029	266.772	20.258	0.371
Trimmed sample with propensity score weighting					
Income	981.906	432.325	387.399	42.975	0.455
Consumption	841.729	345.947	287.687	58.535	0.786

Notes: Yuan per capita per year at 1995 prices. The time period is 1995/6 (baseline) to 2004/5. Standard errors are robust to heteroskedasticity and serial correlation of households within each village. Full sample comprises 112 project villages and 86 comparison villages. In the trimmed sample, there are 71 project villages and 66 comparison villages.

Source: Chen, Mu and Ravallion (2009).

improved" over the last 10 years. But this was also true of 36 percent of those in the non-SWP villages, implying zero impact of the project. Nor is there any significant impact on perceptions of the quality of the public spending in the village by the county government. (It will be recalled that the SWP spending was largely indistinguishable at local level from country spending more generally.)

Ostensibly these SDD1 results are consistent with the findings reported in Chen et al. (2009) indicating little or no long-term impact of the SWP on consumption (or income). However, a closer inspection leads one to question how much comfort one can get from this finding, from the point of view of assessing the scope for using SDD. There is in fact very little correlation between the perceived changes in standard of living and the changes in log consumption per person between 1996 and 2004/05; the correlation coefficient is 0.09 for SWP villages, which is only significant at the 8 percent level ($t = 1.78$); in the non-SWP villages, the correlation is even lower at 0.01. So the fact that SDD1 accords well with the DD estimator using actual consumptions is not because subjective welfare is revealing well the changes in consumption measured in the baseline and follow-up surveys. SDD1 would also show no impact if the recall data was a pure white noise error process.

When we turn to the SDD2 estimator—incorporating an allowance for selection bias on observables—we start to see signs of impact on overall living standards (Table 3). There is also a sign of impact on perceptions of living standards in the village as a whole. Possibly these signs of impacts are statistical flukes; with 30 outcome variables, one could easily get one or two significant effects by pure chance. However, looking at the entire column of differences between outcomes in treatment and comparison villages in Table 2 it is notable how much more positive they are (although often not significantly so) using SDD2. There appears to be something else going on here. The rest of this paper will try to figure out what it might be.

It might be conjectured that the signs of positive welfare impacts using SDD2 reflect some broader concept of "welfare" than captured by consumption. Or one

TABLE 3. Impacts on Self-Assessed Satisfaction With Life Compared to 10 Years Ago

	SDD1: Total sample			SDD2: Trimmed sample with propensity-score weighting		
	Mean in treatment villages	Difference (treatment-comparison)	t-ratio	Mean in treatment villages	Difference (treatment-comparison)	t-ratio
Overall standard of living of h'hold	0.357	0.001	0.018	0.343	0.108	1.635
Income	0.328	−0.005	−0.094	0.324	0.026	0.326
Food	0.377	0.017	0.334	0.356	0.073	1.050
Clothing	0.363	0.028	0.55	0.345	0.094	1.364
Housing	0.313	−0.045	−0.952	0.292	0.006	0.089
Electricity	0.464	0.029	0.515	0.426	0.083	1.020
Hygiene	0.184	−0.058	−1.457	0.186	0.005	0.089
Household appliances	0.275	−0.013	−0.298	0.244	0.010	0.152
Asset accumulation	0.173	−0.009	−0.228	0.151	−0.079	−0.974
Agriculture skill	0.101	−0.026	−0.935	0.087	0.009	0.285
Non-agricultural skill	0.152	−0.057	−1.412	0.146	0.016	0.341
Marketing of agriculture products	0.219	−0.028	−0.627	0.239	0.067	1.132
Credit availability	0.190	0.011	0.251	0.190	0.035	0.589
Affordability of primary/ mid. school	0.22	0.007	0.163	0.209	0.067	1.265
Health	0.302	−0.035	−0.71	0.285	0.092	1.550
School infrastructure	0.392	−0.053	−0.928	0.382	0.039	0.497
School quality	0.306	−0.024	−0.462	0.304	0.047	0.682
Health infrastructure	0.240	−0.059	−1.217	0.219	0.006	0.090
Road conditions	0.377	0.009	0.148	0.376	0.023	0.261
Transportation	0.426	−0.050	−0.846	0.411	−0.061	−0.686
Environment	0.132	−0.030	−0.875	0.129	0.005	0.114
Ecology	0.145	−0.072	−1.852	0.114	−0.024	−0.559
Safety	0.226	0.008	0.156	0.212	0.045	0.687
Knowledge of village affairs	0.170	−0.010	−0.227	0.161	0.048	0.992

Participation in decision-making	0.174	−0.017	−0.413	0.157	0.026	0.536
Democracy	0.232	0.015	0.321	0.216	0.075	1.353
Service to village by county govt.	0.200	0.017	0.369	0.157	0.009	0.133
Service to h'hold by county govt.	0.180	0.034	0.783	0.167	0.041	0.637
Overall village standard of living	0.345	0.034	0.626	0.325	0.116	1.761

Notes: Comparison (with 10 years ago) is based on a scale of 10, 1 being much worse off, and 10 being "totally improved". We redefine those outcomes as dummy variables, equal to 1 if the answer is "obviously improved (8)" or above, 0 if "improved (7)" or below. All the respondents were 28 years or older at the time of interview. Single double difference estimation is made on the total sample of 104 project villages and 79 comparison villages. Weighted double difference estimation is made on the trimmed sample of 66 project villages and 60 comparison villages. Robust standard errors allow for clustering within villages.

Source: Calculations for this paper from the primary data used in Chen et al. (2009).

might argue that welfare recall uses different implicit weights, possibly reflecting missing or imperfect markets. Looking at the SDD2 results in Table 3, there is some sign of an impact on "health" that may account for the implied gain in overall standard of living. Amongst the consumption goods, clothing shows the strongest positive impact using SDD2. However, there is less sign of impacts on any of the many other dimensions of welfare for which the recall questions were used. Nor is it clear why these effects would only emerge when one uses the SDD2 method.

As an additional test for differences between project and comparison villages in "non-income" factors in subjective welfare one can exploit the fact that the recall module included questions on perceived current living conditions (for the same items in Table 3). The relationship between the answers for overall standard of living and consumption per person in the 2004/05 survey data was examined to see if there are any signs that the relationship is different between SWP and non-SWP villages, as might arise from impacts of SWP on "non-income" dimensions of welfare captured in the subjective assessments. The test entailed regressing each subjective measure of the level of welfare on log consumption per capita in 2004/05, a dummy variable for SWP villages and the interaction effect between these two variables. There were no significant differences between SWP and non-SWP villages for all except one of the categories in Table 3. The one exception was for roads ("are you satisfied with village road conditions?"); households with higher consumption in the SWP villages tended to rate road quality higher, but there was no such gradient in non-SWP villages. One might take this to suggest that the SWP enhanced perceived road quality for better-off households, although one cannot dismiss the possibility that one will get at least one significant result in 30 tests purely by chance.

These observations are hardly conclusive, but they don't leave one confident that SDD2 has revealed some genuine impacts that were somehow missing from the DD and SDD1 estimates. As the next section will show, a further insight into the SDD estimates can be obtained by looking more closely at the relationship between the recall of changes in household living standards since the project began and the measured changes in consumption.

III. Relationship between the Impact Estimators

To better understand the relationship between the two estimators, the following regression models are postulated for the recall responses:

$$R_i = \alpha^T + \beta_1^T \Delta Y_i + \beta_0^T Y_{i0} + \gamma^T X_i + \varepsilon_i^T \quad \text{for } i \in T \tag{3}$$

$$R_i = \alpha^C + \beta_1^C \Delta Y_i + \beta_0^C Y_{i0} + \gamma^C X_i + \varepsilon_i^C \quad \text{for } i \in C. \tag{4}$$

Here X_i is a vector of controls and $\varepsilon_i^k (k = T, C)$ are error terms. Notice that all parameters can vary according to whether the treatment is received or not. So this specification allows for the possibility that the two groups have different

TABLE 4. Regressions For Retrospective Assessments Of The Change in the Overall Standard Of Living in the Last 10 Years

	(1) Treatment villages		(2) Comparison villages		Difference	
	coefficient	t-ratio	coefficient	t-ratio	(1)−(2)	t-ratio
Intercept	1.987	1.281	1.905	0.798	0.081	0.029
Change of log consumption between 1996 and 2004/05 (β_1)	0.365	2.522	0.315	1.811	0.05	0.222
Log consumption in 1996 (β_0)	0.321	1.668	0.731	3.025	−0.41	−1.333
Gender of respondent	0.217	0.936	0.400	1.811	−0.183	−0.574
Age of respondent	0.083	2.166	−0.001	−0.017	0.084	1.331
Age^2	−0.001	−1.685	0.000	−0.189	−0.001	−0.837
R^2	0.036		0.043			
Prob. H_0:$\beta_0 = \beta_1$	0.820		0.094			

Notes: The de pendent variable is whether the respondent reported that the household's standard of living had "obviously improved" or better over the last 10 years. Estimation on a balanced panel with 913 households in 100 project villages and 681 households in 75 non-project villages. Standard errors are robust to heteroskedasticity and serial correlation of households within each village.

Source: Calculations for this paper from the primary data used in Chen et al. (2009).

perceived changes in welfare at given (Y_{i1}, Y_{i0}, X_i). In estimating (3) and (4) it will be assumed that $E(\varepsilon_i^k | Y_{i1}, Y_{i0}, X_i, i \in k) = 0$ ($k = T,C$) (as required for OLS to be unbiased). This can be questioned. For example, there may well be omitted variables influencing recall on how living standards have changed and correlated with Y_{i1}, Y_{i0}, X_i.

The estimates of equations (3) and (4) are found in Table 4. The dependent variable takes the value 1 if the overall standard of living is deemed to have "obviously improved" or better, and zero otherwise. The full samples are used (without trimming for common support) and the X vector comprised gender, age and age squared.

Recall that the answers on retrospective recall of changes in overall living standards were essentially orthogonal to the contemporaneously measured changes in consumption. With the controls, significant partial correlations emerge, for both treatment and comparison villages (Table 4). There is also a significant (positive) effect of baseline consumption after controlling for the measured change in actual consumption. This is suggestive of a systematic economic effect on recall errors. Comparing two households with the same actual consumption gain, the poorer one is less likely to report that its standard of living has improved based on recall. There are also signs of gender and age effects. However, the R^2's are low; over 95 percent of the variance in recall of changes in the household's overall standard of living is left unexplained.

Of course, what matters for the impact evaluation is the difference between the models for the treatment and comparison groups. If $\beta_1^T = \beta_1^C$ then (3) and (4) imply a linear relationship between SDD and DD. And this is supported by the

data. One cannot reject the null hypothesis that $\beta_1^T = \beta_1^C = \beta_1$. Using equations (1)–(4) we then see obtain SDD as the following linear function of DD:

$$SDD = \alpha^T - \alpha^C + \beta_1 DD + \beta_0^T E(Y_{i0}|i \in T) - \beta_0^C E(Y_{i0}|i \in C) + \gamma^T E(X_i|i \in T) - \gamma^C E(X_i|i \in C). \quad (5)$$

We can now identify three distinct reasons why SDD is a poor proxy for DD. First, other factors influence SDD besides DD (equation 5). Their weight depends on how similar the treatment and comparison groups are in terms of the means of initial consumption, other covariates and in the model parameters. Table 4 suggests that the (positive) effect of baseline consumption on the perceived change in living standards (after controlling for the measured change in actual consumption) is stronger for the comparison group, which also had higher mean consumption given the selection process. (This suggests that, without matching or trimming, $\beta_0^T E(Y_{i0}|i \in T) - \beta_0^C E(Y_{i0}|i \in C) < 0$.) Thus SDD1 is unlikely to perform well, since one would not have the baseline data on covariates of outcomes needed for matching. Essentially, the selection bias adds noise in the relationship between DD and SDD. But even SDD2 may perform poorly as an indicator of DD if initial outcomes are poorly balanced between treatment and comparison units.

Second, even for the classic case of a randomly assigned program with common impact—for which we could justify setting $\beta_0^T E(Y_{i0}|i \in T) = \beta_0^C E(Y_{i0}|i \in C)$ and $\gamma^T E(X_i|i \in T) = \gamma^C E(X_i|i \in C)$) in equation (5)—the coefficient on DD (β_1) of around 0.3 implies that a very large impact on consumption would be needed to switch the recall variable from zero to unity. Indeed, consumption would need to increase about 30 fold ($e^{1/0.3}$ is about 30)! Clearly SDD is a blunt indicator for DD.

Third, a further observation on Table 4 is that the coefficients on the change in log consumption and 1996 log consumption are very similar; indeed, one cannot reject the null hypothesis that they are the same ($\beta_1^k = \beta_0^k$ for $k = T,C$), although the restriction performs less well for the comparison group. Under this null, it is current consumption that is really driving perceptions of past welfare gains. This can be interpreted as "telescoping," although for the treatment group the recall of changes in the standard of living appears to put too little weight on baseline consumption, while for the comparison group the weight is too high. (Again this probably reflects the selection into the program.) Thus, equation (5) becomes:

$$SDD = \alpha^T - \alpha^C + \beta_1 [E(Y_{i1}|i \in T) - E(Y_{i1}|i \in C)] + \gamma^T E(X_i|i \in T) - \gamma^C E(X_i|i \in C). \quad (6)$$

This is the lethal blow to SDD: it ceases to have any value as an indicator of DD if there is any selection bias, generating observable baseline differences in consumption. And there must be a strong presumption that such differences exist.

In the light of these findings, let us return to the results in Table 3. Given that $\beta_1^k = \beta_0^k$, using recall of welfare changes since the baseline essentially amounts to ignoring the baseline differences. So (roughly speaking) one is regressing (subjectively-assessed) final welfare outcomes (plus the noise in subjective responses) on treatment status. The error term in the SDD1 estimator will contain the selection bias based on both observed and unobserved factors, whether time varying or not.

What then is SDD2 giving us? Adjusting the SDD estimate using PS weighting and trimming aims to balance the treatment and comparison groups in terms of baseline covariates. This provides some protection against selection bias. But the heavier contamination due to unobserved factors in the recall data may well be working in the opposite direction to the selection bias based on observables. The impact found using SDD2 could then be picking up some latent factor in subjective welfare that also helped facilitate village participation in the SWP. In this case study, it is safe to assume that SDD2 has largely removed the effects of the readily observable targeting criteria used to assign villages to the SWP, as we saw in Table 1. However, there are clearly unobserved factors, such as the influence of local political operatives. And these could well be correlated with subjective welfare levels in the village. So the ability of the DD estimator to eliminate the unobserved factors in selection is key to credibly estimating the impact. And (by the same logic) the evident inability of the SDD to do so makes it vulnerable to bias.

By this interpretation, given the structure of the errors in recall, on eliminating selection bias based on observables, SDD2 is revealing the remaining selection bias based on unobservables (including time-invariant factors) that is found in the recall responses on welfare changes.

IV. Conclusions

Given that it is rare to evaluate development projects by repeated observation over a long period, this case study has provided an opportunity to study a less costly method, based on respondent recall using subjective-qualitative questions. Success for this method would open up many low-cost opportunities for learning about development effectiveness.

Neither the "expensive" nor "shoestring" double-difference estimates suggest that the poor-area development program studied here had a significant long-term impact on living standards in poor areas of rural China. But their agreement is not because the retrospective qualitative assessments provided good proxies for the changes in consumption derived from high-quality contemporaneous surveys. Indeed, the analysis suggests that long-term recall of the household's overall standard of living contains only a weak and biased signal of changes in consumption. Controlling for the actual change in consumption, the recalled improvement in living standards tends to be higher for initially richer households. There are clear signs of telescoping in the recall responses, but the bulk of the

benefits occurred in the earlier half of the recall period, which is given too little weight by respondents in treatment villages. Recall is clearly also affected by many idiosyncratic factors not accountable to consumption.

Furthermore, there is an indication that the shoestring method can be deceptive. By not being able to effectively address the problem of selection bias based on the unobserved factors that determined which villages got selected for the program, the recall method becomes vulnerable to spurious impact signals. In this particular case, the recall method suggests positive impacts after controlling for observed differences between treatment and comparison villages at the baseline. The paper has argued that a plausible interpretation of this finding is that the selection bias based on observables is working in the opposite direction to that based on unobserved factors. Thus, only reducing the former bias (by balancing the distribution of observables between treated and comparison units) makes matters worse.

So (alas) this case study does not offer much encouragement on the reliability of this "shoestring approach." Of course, this is just one study and further tests are needed. Recall may be better in some other applications (such as when the intervention coincided with some important event.) Thankfully, the marginal cost of doing such tests in the context of a full-scale evaluation is not high. Hopefully this first study will prompt further tests.

References

Bamberger, M. 2009. "Strengthening the Evaluation of Programme Effectiveness Through Reconstructing Baseline Data." *Journal Of Development Effectiveness* 1(1): 37–59.

Bamberger, M., J. Rugh, M. Church, and L. Fort. 2004. "Shoestring Evaluation: Designing Impact Evaluations Under Budget, Time and Data Constraints." *American Journal Of Evaluation* 25: 5–37.

Bamberger, M., J. Rugh, and L. Mabry. 2006. *Real World Evaluation*. London: Sage Publications.

Beegle, K., K. Himelein, and M. Ravallion. 2012. "Frame-Of-Reference Bias on Subjective Welfare Regressions." *Journal Of Economic Behavior And Organization* 81: 556–70.

Berney, L. R., and D. B. Blane. 1997. "Collecting Retrospective Data: Accuracy of Recall after 50 Years Judged Against Historical Records." *Social Science And Medicine* 45(1): 1519–25.

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

Broegaard, E., T. Freeman, and C. Schwensen. 2011. "Experience from a Phased Mixed-Methods Approach to Impact Evaluation of Danida Support to Rural Transport Infrastructure in Nicaragua." *Journal Of Development Effectiveness* 3(1): 9–27.

Chen, S., R. Mu, and M. Ravallion. 2009. "Are There Lasting Impacts of Aid to Poor Areas?" *Journal Of Public Economics* 93: 512–28.

Chen, S., and M. Ravallion. 1996. "Data In Transition: Assessing Rural Living Standards in Southern China." *China Economic Review* 7: 23–56.

Crump, R., J. Hotz, G. Imbens, and O. Mitnik. 2006. "Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand." Technical Paper 330. National Bureau of Economic Research, Cambridge, MA.

Das, J., J. Hammer, and C. Sanchez-Paramo. 2011. "The Impact of Recall Periods on Reported Morbidity and Health Seeking Behavior." Policy Research Working Paper 5778. World Bank, Washington, DC.

Featherman, D.L. 1980. "Retrospective Longitudinal Research: Methodological Considerations." *Journal Of Economics And Business* 32(2): 152–69.

Fowler, F.J. 1995. *Improving Survey Questions: Design and Evaluation. Applied Social Research Methods Series*. Vol. 38. London: Sage Publications.

Heckman, J., H. Ichimura, J. Smith, and P. Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66: 1017–99.

Hirano, K., G. Imbens, and G. Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using The Estimated Propensity Score." *Econometrica* 71(4): 1161–89.

Iarossi, G. 2006. *The Power Of Survey Design*. Washington, DC: World Bank.

Jalan, J., and M. Ravallion. 1998. "Are There Dynamic Gains from a Poor-Area Development Program?" *Journal Of Public Economics* 67: 65–85.

Janson, C-G. 1990. "Retrospective Data, Undesirable Behavior and the Longitudinal Perspective," In D. Magnusson and L. Bergman, eds., *Data Quality In Longitudinal Research*. Cambridge, UK: Cambridge University Press.

King, E., and J. Behrman. 2009. "Timing and Duration of Exposure in Evaluations of Social Programs." *World Bank Research Observer* 24(1): 55–82.

Krishna, A. 2004. "Escaping Poverty and Becoming Poor: Who Gains, Who Loses, and Why?" *World Development* 32(1): 121–36.

Krishna, A., D. Lumonya, M. Markiewicz, F. Mugumya, A. Kafuko, and J. Wegoye 2006. "Escaping Poverty and Becoming Poor in 36 Villages of Central and Western Uganda." *Journal Of Development Studies* 42(2): 346–70.

McIntosh, C., G. Villaran, and B. Wydick. 2011. "Microfinance and Home Improvement: Using Retrospective Panel Data to Measure Program Effects on Fundamental Events." *World Development* 39(6): 922–37.

Mcphail, S., and T. Haines. 2010. "Response Shift, Recall Bias and Their Effect on Measuring Change in Health-Related Quality of Life Amongst Older Hospital Patients." *Health And Quality Of Life Outcomes* 8: 65.

Narayan, D., L. Pritchett, and S. Kapoor. 2008. *Moving Out of Poverty: Success from the Bottom up*. Washington, DC: Palgrave Macmillan and The World Bank.

Ravallion, M. 2009. "Evaluation in The Practice of Development." *World Bank Research Observer* 24(1): 29–54.

———. 2012. "Poor, or Just Feeling Poor? On Using Subjective Data in Measuring Poverty." Policy Research Working Paper 5968. World Bank, Washington, DC.

Ravallion, M., and S. Chen. 2005. "Hidden Impact: Household Saving in Response to a Poor-Area Development Project." *Journal Of Public Economics* 89: 2183–204.

Ravallion, M., and M. Lokshin. 2001. "Identifying Welfare Effects from Subjective Questions." *Economica* 68: 335–357.

Solga, H. 2001. "Longitudinal Surveys and the Study of Occupational Mobility: Panel and Retrospective Design in Comparison." *Quality And Quantity* 35(3): 291–309.

Watson, W.L., J Ozanne-Smith, and J Richardson. 2007. "Retrospective Baseline Measurement of Self-Reported Health Status and Health-Related Quality of Life Versus Population Norms in the Evaluation of Post-Injury Losses." *Injury Prevention* 13: 45–50.