www.elsevier.com/locate/worlddev

http://dx.doi.org/10.1016/j.worlddev.2013.11.008

World Development Vol. 71, pp. 25–35, 2015
© 2013 Elsevier Ltd. All rights reserved.
0305-750X/$ - see front matter

Non-Sampling Error and Data Quality: What Can We Learn from Surveys to Collect Data for Vulnerability Measurements?

T.D. PHUNG
Mekong Development Research Institute, Hanoi, Viet Nam

B. HARDEWEG
Leibniz University Hannover, Germany

S. PRANEETVATAKUL
Kasetsart University, Bangkok, Thailand

and

H. WAIBEL
Leibniz University Hannover, Germany

Summary. — This paper investigates the causes for non-response and measurement errors in household panel surveys designed for assessing vulnerability to poverty in Thailand and Vietnam. Using data from surveys conducted in 2007 and 2008 we show that interview environment, timing, interviewer, and some respondent characteristics significantly affect non-sampling errors. Investigating interviewer bias for household consumption we find no significant effect of interviewer variables, which suggests validity of the data collected. The paper maps out possibilities to reduce non-sampling errors such as observing suitable interview duration and timing and matching interviewer characteristics with those of respondents.

Key words — non-sampling error, household survey, vulnerability to poverty, Thailand, Vietnam

1. INTRODUCTION

Sampling errors usually can be controlled by choosing an appropriate sampling design, methodology, and sample size (e.g., Groves, 1989). In planning surveys for empirical research in development economics much is known about sampling designs that assures the representation of different groups in the sample and increases the probability of including smaller subgroups relevant to the purpose of the research. Less research however has been carried out on how to better manage non-sampling errors. Although they are considered as a problem in the conduct of surveys especially in developing countries, little is known about their causes and consequences. We argue that there is a need to better understand the role of non-sampling errors and that innovative ways to control them must be found.

There is a lack of empirical evidence on the existence and magnitude of non-sampling errors in surveys in developing countries. There are at least two reasons for this. One of them is that there is no well-established procedure on how to deal with such errors in data analysis. For example, outliers can be eliminated and missing data replaced with average values but both are rules of thumb and are not based on proven theory. The second reason is that the information required to undertake causal analysis of non-sampling errors is often not available, namely details on enumerator characteristics and interview conditions.

Non-sampling errors are important because they can lead to wrong conclusions drawn from the analysis of data. For example, over- or underestimation of crop yields can lead to wrong data on farm income and a high number of missing values can reduce the possibilities to establish significant relationships. While non-sampling errors are important for all empirical research data quality is of particular relevance in field research that deals with economic well-being, poverty, and vulnerability. In this area of research until recently often data were used which had been collected for other purposes (Hardeweg, Klasen, & Waibel, 2012) and where the circumstances of data collection were unknown. Therefore it has been argued that in order to advance research on vulnerability of poor people in developing countries special surveys are needed. To meet the theoretical requirements of advanced vulnerability concepts the instruments of such surveys need to include information on shocks and risks which is often quite sensitive to ask. Hence there is a need to assess how results of specifically designed vulnerability surveys could be influenced by non-sampling errors. In this paper we analyze causes and consequences of non-sampling errors in a household panel survey conducted in the context of a project that deals with vulnerability to poverty in rural areas of Thailand and Vietnam.

We examine whether variables that explain non-sampling errors are correlated with the consumption data of the surveyed household. We find that for two types of non-sampling errors namely missing values and “violation of plausibility rules” interviewer and household characteristics can be a source of error. Based on our results we suggest that observing survey management rules can reduce such errors. The results of our study offer some recommendations how to reduce the effects of these factors on data quality. Thus, lessons can be learned for further panel waves of our survey but also for the increasingly popular household panel surveys in developing countries.

In the next section we review some of the literatures that can provide insights of the factors hypothesized to affect non-sampling errors. In Section 3 a description of the organization and implementation of vulnerability surveys conducted in Thailand and Vietnam in 2007 and 2008 is provided. This is followed by a methodology section that provides the rationale for the empirical model developed to identify the effect of different variables to explain non-sampling errors. In Section 5 results are presented and discussed and in the last section we draw conclusions and submit recommendations for improving the organization and management of household surveys.

2. LITERATURE REVIEW

In the literature, generally three types of non-sampling errors have been defined (e.g., Banda, 2008; Groves, 1989): (i) coverage error, (ii) non-response errors, and (iii) measurement errors. A coverage error occurs when the sampling frame, i.e., the baseline data from which the sample is taken, does not sufficiently cover the target population. It arises during the sampling design phase and is a result of insufficient information about the chance of a sampling unit to be included in the sample (Dillman, 2007). The coverage error includes both, under-coverage, namely the failure to include important sampling units, and over-coverage, which means that untargeted respondents are included in the sample.

Non-response errors refer to the failure to obtain the intended information from respondents. This can be due to inaccessibility of the respondent as well as her refusal or inability to respond. It may also result from the way questions are being asked and to whom they are being asked (Bardasi, Beegle, Dillon, & Serneels, 2012). For example, if some labor activities are on a contractual basis, detailed questions on daily wages cannot be answered by the respondent. If the household head is the respondent she may not be able to provide accurate information of a migrant household member's labor activities.

There are two types of non-responses, namely unit non-response and item non-response. While unit non-response refers to the cases where a certain sample unit is missing as a whole, item non-response refers to the case where the information of a sample unit is only partially collected. For example, the respondent can give information on the yield of a cropping activity but she cannot remember details of the cost of production.

The third type of non-sampling error is measurement error. It occurs when the data obtained are likely to be incorrect. An example is that a respondent provides information on wage employment for unskilled labor and submits a wage value three times the wage level for skilled labor.

In this paper we concentrate on item non-response and measurement errors. We leave out the coverage error because it is more related to the prior information one can obtain in survey preparation and therefore strictly speaking it is not a non-sampling error. Research to measure the impact of data entry and questionnaire design on non-sampling error has mostly focused on a few factors (e.g., Glewwe & Dang, 2008). In this paper we are able to cover a wider range of factors including interviewer and respondent characteristics for a rather large data set in two countries.

There are at least five possible ways how non-sampling errors can occur. First, social psychologists (e.g., Kahn & Cannell, 1957) consider a survey interview as a structured social interaction and therefore, the demographic and socioeconomic characteristics of the interviewer can influence the behavior of respondents. Second, interviewers may deviate from the established standard procedure. For example, an enumerator can reword questions, may omit some (sensitive) questions, or make wrong recordings. Marquis and Cannell (1969) showed that the major reasons that contribute to errors in recorded data are the failure to read a question exactly as printed, incorrect compliance with skip patterns, and reading a question too fast. Third, even if the interviewer follows the guidelines and reads out questions exactly as written in the questionnaire, intonation or emphasis for certain words can vary, possibly prompting altering answers of respondents. Hyman (1954) argued that interviewers have a prior distribution of expected answers to the questions which influences the way they conduct the interview, e.g., by changing intonation and voice levels. Fourth, interviewers may assist the respondents in finding answers to difficult questions, e.g., by using different probing techniques for events which are difficult to remember. Fifth, a non-sampling error can also arise from proxy reporting, i.e., if one household member (usually the household head) is charged to provide information for all persons in the household.

Sources of non-sampling errors include imprecise definitions, faulty methods of enumeration, inappropriate survey instruments, using ambiguous questionnaires, definitions or instructions, lack of trained and experienced enumerators, inadequate supervision as well as inadequate scrutiny of the basic data, and errors in data processing operations such as coding, keying, verification, and tabulation (Banda, 2008).

Most of the recent studies on non-sampling errors analyze the relationship between interviewer characteristics and non-responses. Among the factors identified in previous studies are gender, age, experience of interviewers, interviewer–respondent interaction, organization of data entry, and incentives provided to respondents.

For gender, two studies (Fowler & Mangione, 1990; Lessler & Kalsbeek, 1992) found that female interviewers are more successful than male interviewers in obtaining information from respondents. For interviewer age results are mixed. While Lievesley (1986) found that middle-aged interviewers achieved higher response rates than young or old interviewers, Morton-Williams (1993) did not support this finding and Singer, Frankel, & Glassman, 1983 found higher response rates by older interviewers. Durbin and Stuart (1951), Lievesley (1986) and Couper and Groves (1992) found that interviewers' survey experience is positively correlated with the response rate. Campanelli, Sturgis, & Purdon, 1997 analyzed response rates in longitudinal surveys and found that interviewer continuity is important during earlier survey waves but less so in later phases. Interviewer-respondent interaction was found to affect data quality. Fisher, Reimer, and Carr (2010) found that when income composition has a strong gender focus, interviewing the household head alone did not produce statistically reliable results for poverty analysis. The study also showed that when men were asked about their wives' incomes, considerable inconsistencies occurred. Regarding data entry organization Glewwe and Dang (2008) found that entering data in the field within one or two days of completing the interview instead of doing it several weeks later improved the quality of data. They