Mark Brooks (Leibniz University Hannover), Rattiya Lippe (Leibniz University Hannover), Hermann Waibel (Leibniz University Hannover)

Comprehensive Data Quality Studies as a Component of Poverty Assessments

Understanding and measuring poverty is not possible without household data. Especially for dynamic and multi-dimensional poverty measures, high quality panel data on income and consumption are essential for realistic poverty assessments. Hence, reliable data are the basis for designing effective policies to sustainably reduce poverty.

Income measures are often plagued by non-sampling errors (e.g. missing information and measurement error) due to the sensitive nature of income data (e.g. Meyer et al., 2015). Therefore, a better understanding of the determinants of non-sampling errors is necessary. While data access, use, and collection has benefited from technological innovations, these do not automatically solve the problem of data quality. In household surveys in developing countries, tools, such as "Survey Solutions" by the World Bank (https://mysurvey.solutions), have significantly increased the effectiveness of data collection. Experimental evidence (Caeyers et al., 2012) showed that "Computer Assisted Personal Interviewing" can prevent errors, especially in collecting consumption data, which are prevalent in paper-based questionnaires. Although electronic questionnaires allow the implementation of automated plausibility checks, monitoring data quality remains a necessary complementary task. Interviewing respondents is a challenge in the context of developing countries, in particular regarding management and communication. Issues such as misinterpretation of questions, lack of interviewer consistency and challenging interview conditions cannot be eliminated solely by electronic questionnaires.

Previous data quality research has largely focused on individual factors, which can cause non-sampling errors. For example, Beegle et al. (2012) found that long reporting periods lead to significant data errors resulting from recall bias. Fisher et al. (2010) concluded that male household heads frequently underestimate their wife's income. Using household panel data from Thailand and Vietnam, Phung et al. (2015) showed that aside from interviewer and respondent characteristics, the interview environment significantly affects the frequency of non-sampling errors.

Research on data quality of household surveys in developing countries has generated important lessons, which if implemented can significantly improve data for poverty estimates. However, three important aspects of data quality have not received sufficient attention so far. Firstly, many studies rely on either cross-section data or experiments, which can limit the scope in which data quality is analyzed. Secondly, while quantifiable interviewer and respondent characteristics (e.g. age) are included in such studies, qualitative information such as subjective assessments of interviewer behavior (e.g. professionalism) are not yet considered. Thirdly, most

studies focus on individual aspects of data quality such as the effect of interviewer and/or respondent characteristics. With few exceptions (e.g. Phung et al., 2015) studies rarely account for the interview environment (e.g. duration/timing of interview) and survey environment (e.g. survey schedule).

This paper presents results of a comprehensive study on data quality in an ongoing long-term household panel project in Thailand and Vietnam (www.tvsep.de) that began collecting household and individual level data from 4,400 households in 2007. Four potential sources of errors are accounted for: interviewer and respondent characteristics, and the interview and survey environment. Additionally, we control for respondent fatigue measured by participation patterns throughout the span of the project.

The objective of this paper is to assess simultaneously the effect and relative importance of these factors on the frequency of non-sampling errors. We determine the effect of non-sampling errors on income measures, and suggest measures that can improve their quality. In our analysis, we first establish the frequency of four types of non-sampling errors: missing values; refused answers; extreme values (i.e. outliers); and responses that do not comply with survey plausibility rules (i.e. implausible values). The share of each error type in relation to the total number of items relevant to income measures in the survey instrument is used as the dependent variable.

We specify, by country, five separate OLS equations, which differ in their dependent variables. Each equation accounts for one of the four error types, whereas the fifth uses the sum of non-sampling errors.

Explanatory variables include quantitative interviewer and respondent characteristics. In addition, we include qualitative variables such as subjective personality traits of interviewers and performance assessments by their supervisors. Exam scores of interviewers are used as a proxy for their initial survey knowledge. Moreover, subjective assessments of each interview by interviewers and respondents, which capture the quality of interviewer-respondent interactions, and disturbances during the interview, are included. Finally, variables describing the survey environment, e.g. dummy variables capturing travel-/rest-days are included.

Results confirm findings of earlier research; for example, characteristics of both the interviewer and respondent (e.g. age, education) as well as congruent characteristics have a significant effect on non-sampling errors. Respondent ethnicity only affects data quality in Vietnam as indicated by previous research (Dang, 2012). Both interviewer and respondent personality traits have a significant effect on data quality. Furthermore, lower quality data stems from negative interviewer-respondent interactions and poor interview conditions, such as mismatched interview times when compared to respondent preferences. Interviews that took place after travel days produced data of lower quality.

By means of a comprehensive model, we provide novel insights on the determinants of non-sampling errors. Furthermore, we show that adjustments to the survey environment and higher

quality interview environments can lead to more reliable income data, which are necessary to improve the validity of poverty assessments and to develop policies to reduce extreme poverty.

References

Beegle, K., Carletto, C., & Himelein, K. (2012). Reliability of recall in agriculture. Journal of Development Economics, 98(1), 34-41.

Caeyers, B., Chalmers, N., & De Weerdt, J. (2012). Improving consumption measurement and other survey data through CAPI: Evidence from a randomized experiment. Journal of Development Economics, 98(1), 19-33.

Dang, H.-A. (2012). Vietnam: A Widening Poverty Gap for Ethnic Minorities. In G. Hall & H. Patrinos (Eds.). Indigenous Peoples, Poverty and Development (pp. 304-343). New York, NY: Cambridge University Press.

Fisher, M., Reimer, J. J., & Carr, E. R. (2010). Who should be interviewed in surveys of household income? World Development, 38(7), 966-973.

Meyer, B. D., Mok, W. K. C., & Sullivan, J. X. (2015). Household Surveys in Crisis. Journal of Economic Perspectives, 29(4), 199-226.

Phung, T. D., Hardeweg, B., Praneetvatakul, S., & Waibel, H. (2015). Non-sampling error and data quality: What can we learn from surveys to collect data for vulnerability measures? World Development, 71, 25-35.