An Assessment of Data Quality of *The 5th National Survey on Family*: Evaluation of Survey Nonresponses and Data Matching with the Data Set of the 2013 *Comprehensive Survey of Living Conditions*

Masakazu YAMAUCHI, Keita SUGA and Jun KIKUCHI

This study has two objectives. First, we assess the data quality of *The 5th National Survey on Family* (NSF), which was conducted by the National Institute of Population and Social Security Research in 2013. Second, we examine the practical feasibility of combining two micro-data sets in a micro-level analysis. For the first objective, we analyze both unit and item survey nonresponses in the NSF. For the second, we perform a micro-level match-merging of the NSF to the 2013 *Comprehensive Survey of Living Conditions* (CSLC), conducted by the Ministry of Health, Labour and Welfare.

A unit nonresponse refers to the complete absence of a sampled unit's response to a survey. In the study, it was found that the overall rate of unit nonresponses in NSFs increased from 24.3% in the 2nd round (1998) to 36.0% in the 4th (2008), and the rate was 36.1% in the 5th NSF (2013). An examination of the data collection process revealed that the contact rate in the 5th NSF worsened from the 4th round, but the response rate in the 5th NSF improved.

Here, the contact rate refers to the proportion of sampled units who could be successfully contacted by field workers, and the response rate refers to the proportion of contacted persons who finally submitted completely filled questionnaires. The reason for unsuccessful contacts was that field workers could not deliver or collect questionnaires due to the lack of effective contacts, and we found that the rate of unsuccessful contacts in both the stages (delivery and collection) increased from the 4th round.

An item nonresponse refers to the absence of valid answers from the respondents. In the study, a higher number of item nonresponses was found in the 5th NSF compared to the 4th. Moreover, item nonresponse rates decreased in terms of both variables and cases in the 5th NSF micro-data set. However, we found that the data cleaning procedure in data processing, which identified inconsistent or illogical answers to determine whether a sensible correction was attainable, generally improved item response rates among both items and responded units.

For the second objective, we note that, in principle, the NSF micro-data can be merged on a one-to-one basis to cases in CSLC, because a common sampling frame is used for both. In practice, however, the exact matching cannot be realized. Indeed, we found that the proportions of matched cases were 93.3% of the 5th NSF units and 61.7% of the 2013 CSLC units.

On the basis of these results, we encourage those using the 5th NSF data to exercise caution and be aware of the limitations. Further, we point out desirable revisions to the survey design, especially in the construction of the questionnaire, for the next round of NSF, which is scheduled to be conducted in 2018.