Using the R-indicator to study attrition bias in a probability based Web panel

Paper to be presented at the International Workshop on Household Survey Nonresponse
30 August – 1 September 2017, Utrecht

Annette Scherpenzeel
SHARE – Survey of Health, Ageing and Retirement in Europe
Munich Center for the Economics of Aging (MEA)
scherpenzeel@mea.mpisoc.mpg.de

Thomas Klausch
Department for Epidemiology and Biostatistics
VU University Medical Center
T.Klausch@VUmc.nl

1. Introduction

The LISS panel is a large-scale online panel established in 2007 on the basis of a random probability sample from the Dutch population. This online panel has been constructed by selecting a random sample of households from the population register of The Netherlands. Selected households were approached by means of CAPI or CATI. Moreover, co-operative households without Internet access were provided with equipment giving them access to the Internet. The details of the sampling design, the recruitment procedure and the response rates of the LISS panel are described by Scherpenzeel and Das (2010).

As a result of recruitment nonresponse and attrition over time, the LISS panel might have some biases in its composition. In an earlier paper, we have estimated the R-indicator for the LISS panel initial sample, using population register auxiliary information, to study the representativeness of the panel for the target population. The research question we aimed to answer in that paper was how representative an online panel recruited on the basis of a probability sample is for the general Dutch population. In addition, we compared the LISS panel R-indicator with the R-indicator estimated for the Dutch Labour Force Survey, to see how large the relative selection bias in the online LISS panel was compared to a traditional non-Internet panel study.

In the present paper, we followed a different approach, estimating R-indicators across panel waves by means of the first wave variables. The R-indicator in this case does not indicate population representativeness but panel composition compared to the net sample in

the first wave. The advantage of this approach is that a large number of survey variables, ranging from socio-economic characteristics to general attitudes, can be included in the R-indicator. The research question we aimed to answer in this paper is how systematic the attrition in the LISS panel is with respect to the core research variables.

2. The LISS panel

The CentERdata research institute in Tilburg, the Netherlands, has set up the LISS panel (Longitudinal Internet Studies for the Social Sciences). It consists of about 8000 individuals that complete online questionnaires every month. LISS panel members complete online questionnaires every month, for which they get an incentive of €7.50 per half hour of interview time. They are invited each month by email and receive two reminders spread over the month when they do not complete the questionnaires. The analyses in this paper include only panel members who ware sampled in the original 2007 sample and participated in the first wave of the core questionnaire modules Health, Politics and Values, and Personality. In a later stage of the project, we will also add the panel members sampled in the refreshment samples of 2009 and 2011, but they are excluded in the present analyses.

4. The R-indicator

The R-indicator is developed by Statistics Netherlands (SN) as a measure of response bias (see Cobben, 2009; Bethlehem, 2010; Schouten, Cobben and Bethlehem, 2009 Schouten, Shlomo and Skinner, 2011), in the framework of the research project Representativity Indicators for Survey Quality (RISQ). The R-indicator ('R' for representativity) is based on estimated response probabilities. It represents the dissimilarity between the respondent and sample pool with respect to auxiliary variables that are available from other sources than the survey itself (Schouten, Cobben and Bethlehem, 2009).

In the first stage of the LISS R-indicator project, we estimated the R-indicator for the initial LISS panel sample using variables which were all derived from the population register (Scherpenzeel and Schouten, 2011). In order to obtain these register variables, the LISS panel response data were linked to the population registers by SN. The variables included were: Age (average age of core household), household composition, ethnical background (household definition), urbanization, average value of houses in postal area ("WOZ-waarde") and employment (at least one person employed in core household).

In the second stage, presented in this paper, we estimated the R-indicator on the basis of the first wave data collected with the 2007-2008 LISS panel core questionnaires. The variables included were: Socio-demographics (e.g. sex, age, household size, income), health indicators (e.g. BMI, mobility, smoking, drinking, number of health conditions), personality

traits (e.g. big five dimensions, need for cognition, self-esteem), survey attitudes and political traits (e.g. political interest, external and internal efficacy, political trust). The first wave core data were linked to a response data set which included a response score for each month of data collection, from November 2007 to January 2015, for each panel member of the 2007 net sample. The response score was 1 if a panel member completed at least one questionnaire that month, 0 if a panel member was selected for at least one questionnaire that month but did not complete any questionnaire, and -1 if a panel member was no longer selected for any questionnaire which in general indicated he or she had dropped out. In the analysis, the codes 0 (questionnaire nonresponse) and -1 (panel drop out) were collapsed, constituting a single nonresponse category for each data collection month. The R-indicator for each point in time constitutes the dissimilarity in the first wave core data of the total 2007 net sample and the first wave core data of the remaining 2007 sample members at that time point[1].

5. Results

5.1 Initial representativeness

In the 2011 paper (Scherpenzeel and Schouten, 2011) it was shown that the R-indicator in the recruitment stage and starting stage was somewhat lower in the LISS panel than in the Labour Force Survey (0.85 versus 0.91 in the contact stage; 0.79 versus 0.84 for the panel registration/first interview[2]). The final overall recruitment response rate in the LISS panel in 2007/2008 was 48%, defined as: the number of households in which at least one person online registered as panel member as proportion of the total number of households in the gross sample excluding unusable addresses[3]. The R-indicator calculated on the basis of the gross sample and the net registered sample of 2007/2008, using the register variables, was 0.67 (CI 0.66 - 0.69)[4].

5.2 Sample development over time

On average, about 12% of the individual panel members leave the LISS panel per year. The overall monthly response varies between 65% and 79%. Figure 1 shows the participation

[1] It is hence not based on the dissimilarity of the wave one data and the data collected in later waves of the core questionnaire.
[2] Scores based on population and samples between age 16 and 65.
[3] Rate based on samples older than 15 years, without upper age limit.
[4] Score based on population and samples older than 15 years, without upper age limit

rate and the R-indicator from November 2007 to January 2015. In January 2008, 73% of the persons in the original 2007 sample who had registered and started in the panel completed at least one questionnaire that month. In January 2015, this was only 33%. The graph excludes the refreshment samples that were recruited over the years.

Figure 1. Participation rate and R-indicators of the 2007 net sample of the LISS panel, from 2007 to 2015.

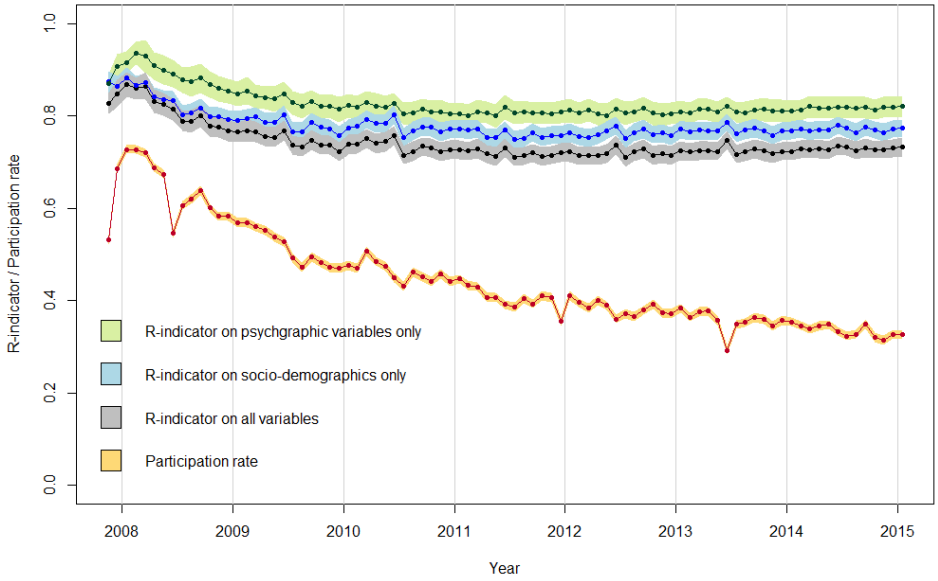
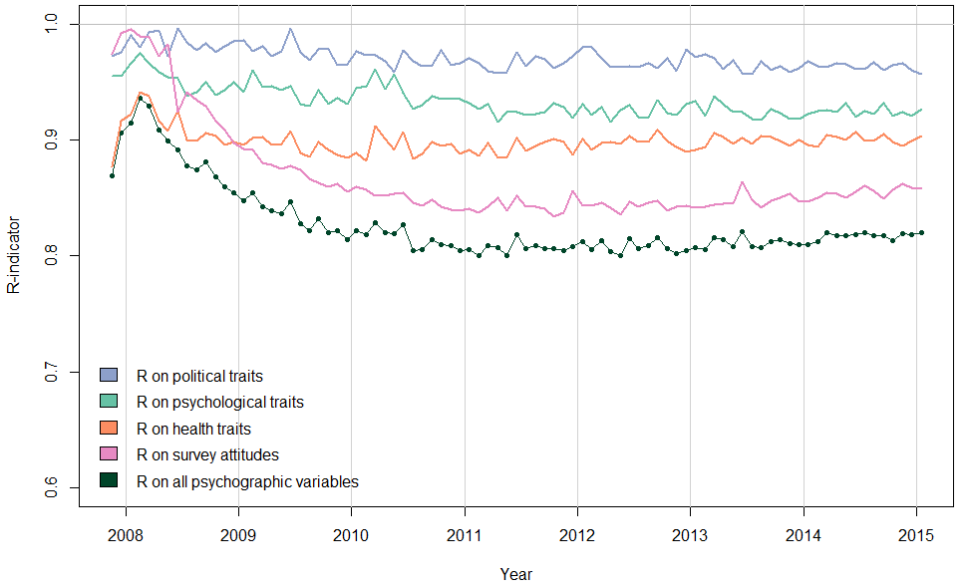


Figure 1 presents three different R-indicator graphs: One including all demographic and core variables, one including only the standard socio-demographic variables, and one including only the core questionnaire variables in the domains health, personality, political values and survey attitudes (titled "psychographic variables" in figure 1). All three R-indicators were calculated by taking the wave 1 net sample of 2007/2008 as reference. The steadily decreasing response over years is not reflected in the three R-indicator patterns: The largest change in sample composition in comparison with the wave 1 sample is observed in the first two years. From the highest level 0.86 found in January 2008, the overall R-indicator decreases to a stable level of about 0.75 between 2010 and 2015, with small peaks in all June months. This pattern is observed for all variables, but the socio-demographic variables are more affected by systematic attrition than the core questionnaire variables.

Figure 2 shows separate R-indicators for the core questionnaire domains health, personality, political values and survey attitudes. The largest selection effect of sample attrition is observed in the health variables and survey attitudes. Personality measures and political values are much less affected by systematic attrition.

Figure 2. R-indicators of the 2007 net sample of the LISS panel, for different core questionnaire domains.



6. Discussion

This paper is a follow-up on an initial R-indicator estimation done for the LISS panel in 2011. The R-indicator calculated at that time was the representativeness indicator as it was originally intended: the score of 0.67 (CI 0.66-0.99) which was obtained at that moment signified that the first wave net sample which started participating in the LISS panel in 2007 represented the population moderately well on a set of standard demographic variables from the population register. For the follow up, we used a new approach implying a different meaning for the R-indicator: Instead of basing the R-indicator on the target population or gross sample, we take the first wave net sample of the LISS panel as the reference to which we compare the panel sample composition at later years. The reference sample has hence already been affected by initial recruitment nonresponse bias, as the originally estimated R-indicator of 0.67 showed. Consequently, the R-indicator based on this reference cannot be considered a true indicator of representativeness in the usual sense. Instead, it is meant to show whether the attrition in the LISS panel is systematic and affects the core research variables. An alternative approach would be to estimate the usual R-indicator across the panel years, using the already linked population register variables. This would show the development of the population representativeness of the panel over years and the effect of attrition on that representativeness. However, such an analysis would be restricted to the usual demographic variables and in our view not significantly contribute to the existing knowledge about attrition effects. By using the first wave panel sample as the reference, we

are able to include a large number of core research variables, such as health, personality, values and attitudes. Our results show that some research variables are sensitive to systematic attrition whereas others seem unaffected. The main discussion points are: 1) the utility of our approach in comparison to the alternative and 2) the consequences of the differential effect of attrition on different research variables. In a later stage of the project, we will use the same approach as in the present paper to estimate the change in panel R-indicator as a result of adding the LISS panel refreshment samples of 2009 and 2011.

References

Bethlehem, J. (2010). New developments in survey data collection methodology for official statistics. Discussion paper 10010, The Hague/ Heerlen: Statistics Netherlands. Available at: http://www.cbs.nl/NR/rdonlyres/BE555091-E0E7-47A6-84BC-6BE980646997/0/201010x10pub.pdf

Cobben, F. (2009). Nonresponse in sample surveys, methods for analysis and adjustment. Doctoral dissertation, The Hague/ Heerlen: Statistics Netherlands. Available at: http://www.cbs.nl/nl-NL/menu/methoden/onderzoek-methoden/onderzoeksrapporten/proefschriften/2009-x11-cobben-pub.htm.

Scherpenzeel, A. and Marcel Das (2010). True Longitudinal and Probability-Based Internet Panels: Evidence from the Netherlands. In: Das, M., P. Ester, and L. Kaczmirek (Eds.), Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies. Boca Raton: Taylor & Francis.

Scherpenzeel, A. and Schouten, B. (2011). LISS panel R-indicator: representativity in different stages of recruitment and participation of an Internet panel. Paper presented at the 22nd International Workshop on Household Survey Nonresponse, 5-7 September 2011, Bilbao, Spain.

Schouten, B., Cobben, F., Bethlehem, J. (2009), Indicators for the representativeness of survey response, Survey Methodology, 35 (1), 101 – 113.

Schouten, B., Shlomo, N., Skinner, C. (2011), Indicators for monitoring and improving survey response, Journal of Official Statistics, 27 (2), 231 – 253.