Utilising Interviewer Observations on Housing Unit Characteristics in the Hungarian LFS Nonresponse Analysis: A Research Plan

Authors: Mohay, Linda; Mújdricza, Ferenc

Affiliation: Hungarian Central Statistical Office, Methodology Department

Contact: Ferenc Mújdricza, Ferenc.Mujdricza@ksh.hu

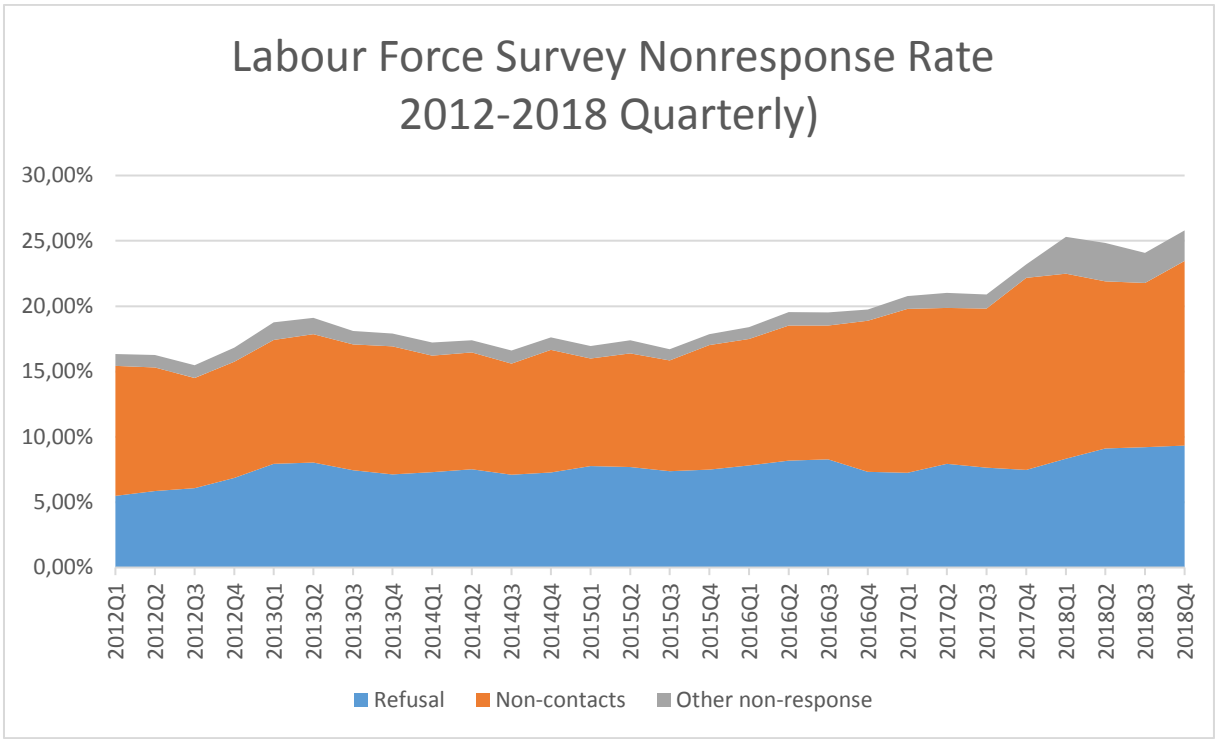
Introduction

Numerous papers report the relationship between interviewer observations of face-to-face surveys on housing unit/neighbourhood characteristics and survey participation as well as nonresponse (see for instance Groves & Couper 1998, Lynn 2003, or the rather exhaustive literature reviews in Casas-Cordero 2010, West 2011, and Casas-Cordero et al. 2013). The present paper is about a work-in-progress research on how interviewer observations on housing unit and/or neighbourhood characteristics could be useful for nonresponse analysis and, hopefully, for nonresponse adjustment. The research is still in its design phase, therefore the paper tells about the advances that have been made so far since the devising of the first version of the research plan (see the abstract submitted for the workshop). These advances include delving deeper in the available literature, analysis of the existing interviewer observation variables of the Hungarian LFS, and in the lights of the thus acquired knowledge, necessary changes in and refining of the preliminary research plan.

Background and aims of the study

Increasing nonresponse rates of the Hungarian LFS (see *Table 1*) demand researching the sample composition of completions and nonresponse. The need for better nonresponse adjustment methods necessitate the improvement of the design and use of the variables that can be used for devising appropriate weighting procedures.

Table 1.



As Lynn (2003) argues, adjustment for unit-nonresponse bias is typically made by weighting procedures based on indirect measures from administrative (frame or population) data. The Hungarian Central Statistical Office (HCSO) cannot follow this custom, for it does not have access to administrative data sources adequate to the purpose that other NSIs use to adjust for nonresponse bias. However, it might not even be feasible. As Lynn (2003) goes on, such procedures would require strong correlation between the weighting classes and survey measures as well as response propensity – however, the fact is that these correlations are often not strong enough. Lynn assumes that data collected *directly*

from the non-respondents might serve as better bases for weighting. Budget-friendly collection of adequate data with the Key Items Form (KIF) instrument that Lynn (2003) proposed is only possible (to a certain extent) regarding refusals (where contact has been made with the non-respondent). According to West's (2011) summary of the literature, auxiliary variables to be used for reducing nonresponse bias have to be available for respondents as well as nonrespondents.

Interviewer observation data are – at least ideally – free from nonresponse issues because they do not depend on the cooperation of the sample unit that is out of the control of the researcher (see Groves and Couper 1998), only on the proper usage of interviewer observation variables by the interviewers that is, at least in theory, under the control of the researcher. Therefore, interviewer observations on housing unit and/or neighbourhood characteristics may be the only data truly adequate to adjustment for nonresponse bias – although they are not collected directly *from* the entire group of non-respondents, they provide us with direct information *on* them. Lynn's (2003) above-mentioned note on the importance of correlation between key survey variables pertains to interviewer observation data as well, determining the extent of their utility. The problem is, as West (2011), Casas-Cordero (2010), and Casas-Cordero et al. (2013) summarise in the light of the relevant literature, that these correlations are also rather low in survey practice, and the predictive power of interviewer observations on housing unit/neighbourhood characteristics for survey participation is not as high as necessary for successful unit nonresponse adjustment.[1] As West (2011) assumes and Casas-Cordero et al. (2013) interpret Kreuter et al.'s (2010) discussion on the matter, the generally low correlations and limited effectiveness of nonresponse adjustments are potentially caused by measurement error in these auxiliary variables, which assumption seems to be in conflict with the proven reliability and validity of the instruments (see for instance Andresen et al. 2006, Andresen et al. 2008, etc.). According to the logic of Casas-Cordero et al.'s (2013) paper, this conflict is due to the sheer fact that survey interviewer observation data differ from the data of studies focusing on collecting housing unit/neighbourhood characteristics data for the sole practice of testing the reliability and validity of an instrument. This logic points out the importance of budget- and resource-efficient training of survey interviewers while maintaining their standard and appropriate use of the observation instruments to rule out interviewer effect caused by the differences in their individual perception and judgment.

Although interviewer judgments are prone to errors hindering successful adjustment of survey estimates for nonresponse bias (West 2011, West & Kreuter 2015), these measurement errors can be reduced and the quality of interviewer observations can be enhanced. There are at least two possible directions. As Sinibaldi et al. (2013, 190) note, 'as with questionnaire items, observation questions should accurately capture the construct of interest and be understood consistently by interviewers.' Taking this hint, it seems obvious that the solution of measurement error issues cry for cognitive testing of these instruments among survey interviewers to enable their correction, as it is routinely done with other survey instruments at most NSIs. Given that we know of no such study to date, the present paper focuses in part on this woefully neglected but rather important possible direction of improving the quality of the collected auxiliary data. One can assume, on the one hand, that quantitative validity and reliability tests with satisfactory outcomes might hide issues in the question –response process that may be the reason for the high levels of interviewer effect despite the good results (Andresen et al. 2006, Andresen et al. 2008). The fact that qualitative testing of these instruments is lacking shows that researchers tend to 'blame' the data collectors (i.e. the interviewers) entirely for the measurement error, and are uncritical towards the instruments themselves, which

[1] Although, according to West's (2011) literature review, there are some studies demonstrating stronger correlations with response propensity and/or key variables (e.g. Lynn 2003, Durrant et al. 2010, etc.), they did not aim for nonresponse adjustment.

might well be inadequate for practical use as auxiliary survey instruments despite the fact that they passed quantitative validity and reliability tests. On the other hand, as described above, the use of the instruments in such tests is far from being identical to their use in surveys, and therefore it is quite reasonable to assume that interviewers who are less specifically trained for the sole use of these instruments and are burdened with the much more important collection of the survey data may well interpret and/or handle the items differently than data collectors of the reliability/validity tests did. At any rate, it is safe to assume that the instruments may be at least partly inadequate in a survey environment, and might need smaller or bigger revisions. Although West (2011) refers to Pickering et al's (2003, cf. West 2011) study in order to support his hypothesis on interviewer judgment inaccuracy with regard to housing unit types, the results he refers to seem to be much better examples for the demonstration of inadequate, ill-considered instrument design. This is also in line with Sinibaldi et al's (2013) results and conclusion assuming that observations that demand subject-specific knowledge or are not easily observable might not be a good idea to collect. Although West (2011) does a fair attempt on showing the social psychological aspects of judgments based upon brief ('thin-slice') observations, it is still seems to be rather unreasonable expecting interviewers to assess correctly whether or not respondents own or rent the dwelling they live in without asking them – 46% accuracy rate in Pickering et al's (2003, cf. West 2011) study show the reality one can anticipate: they may have only made sheer guesses. Sinibaldi et al. (2013) found better accuracy rate on the assessment of Council or self-owned dwellings (87%), however, it was the worst in their study. The use of such instruments as interviewer observation variables is thus safe to deem as a possibly mistaken design decision. The obvious logical shortcomings aside, such a mistake would most probably have come to light had a proper qualitative (cognitive) testing of these instruments been done among interviewers.

The other, much better-trodden possible direction is a special training of the interviewers on the use of housing unit and neighbourhood characteristics variables. Provided the instruments are truly adequate, well-tailored trainings may effectively reduce idiosyncracy in interviewer observations and minimise measurement errors and interviewer effect by standardising the use of observation instruments as hypothesised in the literature (Casas-Cordero 2010, Casas-Cordero et al. 2013, West 2011, Sinibaldi et al. 2013, West & Kreuter 2015, 2018). Although Andresen et al. (2013) found no improvement in measurement quality after enhancing interviewer training (as well as field protocols), the second aim of the present study is to investigate the possible benefits of such a training.

Progress of the research to date

The Hungarian LFS has three existing interviewer observation variables on sample housing units, of which only one is designed to reflect neighbourhood characteristics to a very limited extent. This limited scope would not entail in itself that the data collected by these variables could not work as expected in a nonresponse adjustment procedure. What expectations can be raised towards a set of interviewer observation variables? There are at least three core requirements it should meet:

1. Flawless design: is it logically sound, are the item formulations free of any flaws, imprecisions, inadequacies, etc., do they represent appropriately the possible typical housing unit/neighbourhood characteristics specific to the population of the survey?
2. Clarity and ease of use: does it consist of clear, straightforward, easy to use items that interviewers interpret exactly the same as or as close as possible to the meaning intended by the designers?
3. Suitability for adjustment: is it suitable for assessing nonresponse bias, that is, do they correlate strongly enough with key survey variables and survey participation?

In theory, the first and second of these requirements might not be taken as separate because meeting the second could be taken as a design task. However, given the literature's apparent ignorance on it, the present paper treats it separately. Furthermore, as it follows from the considerations above, the third requirement is closely related to the way the instruments are being used by the interviewers, which depends greatly on the first and second requirements as well as on the quality of the training that could be taken as an external requirement (from the variable aspect). Although the third requirement is the most important from the point of view of nonresponse adjustment, it is partly a derivative of these factors. Therefore, in theory, it is more of an indicator of possible interpretation, design, training, actual field use, etc. problems. However, for successful nonresponse adjustment, the variables have to meet this requirement as well, through the data collected with them.

The housing unit/neighbourhood observation variables of the Hungarian LFS are very limited in their scope, and they also suffer from multiple design issues. The variables are the following:

A. Neighbourhood characteristics of the close residential area of the building
    1) traditional urban built environment
    2) microdistrict
    3) block of flats and villa district
    4) detached houses
    5) rural characteristics
    6) industrial area
    7) peripheral
    8) socio-economically deprived neighbourhood
    9) other, namely .....
B. What is the condition of the building in which the dwelling resided by the household is situated?
    1) excellent (newly built, renovated)
    2) satisfactory
    3) good
    4) degraded condition
C. In what type of building is the dwelling?
    1) detached house
    2) semi-detached or terraced house
    3) block of flats with less than 10 flats
    4) block of flats with 10 or more than 10 flats
    5) other dwelling unit, namely …

The major issues of the instrument are easy to identify. The question formulations are more or less appropriate, but the response categories are quite problematic. Variable A seems to be the most flawed. The preliminary examination uncovered that is seems to

- have undercoverage, clarity, or training issue: interviewers did not overuse the 'other' category, however, the analysis of their textual responses shows that the vast majority pertained to vacation homes. It indicates either category deficit, that is, the categories do not cover every possible case, or improper interpretation of the category in which vacation homes should have been categorised that could be a result of insufficient training or unclear, perplexing wording of the category;
- have undercoverage or overlap issue: certain special, but not infrequent cases may be difficult to categorise into the existing categories, e.g. the special Hungarian farmhouse neighbourhood

might belong to categories 5 or 7, but it does not fully belong to either – instead, it represents a separate neighbourhood characteristic that may require a new category;
- have overlapping or not properly discernible response categories: affected categories 4-5, 4-7, 1-4, etc., mainly because category 4 is linked to a certain type of dwelling, and the other affected categories are defined by environmental characteristics; another example of this issue is the overlapping of categories 1 and 3, for various types of block of flats built next to each other in a close line can be a distinct characteristic of traditional Hungarian urban built environments, particularly in bigger towns or cities (it does not, however, pertain to mansions of course, see the next issue);
- match completely different characteristics in the same category: as in category 3;
- be a hidden double-barreled question: category 8 represents a different phenomenon, that is, the social 'suitability' of the area that is rather obscure in this formulation. It might also be a euphemism for huts, lodges, or severely deteriorated lodgings. In this case, the double-barreled question issue does not hold, but then the formulation of the category is improper, unclear, and should be reformulated.

The variable and the characteristic it refers to should be subjected to further expert analysis. Based on the results of the analysis, its issues can be fixed and a proper formulation can be designed that will be ready for cognitive testing.

Variables B and *C* also suffer from issues, although from less serious issues than *Variable A*. For example, the fact that more than 60% of the responses on *Variable B* were response category 2 ('good') assumes that interviewers may not be able to interpret category 3 properly ('satisfactory') or they cannot differentiate it from 'good'. The overuse of category 2 might thus indicate that the scale is inappropriate in its range (in practice it is used as a scale with three response categories instead of four) or wording (overlapping meanings of categories 3 and 4). *Variable C* also has issues, for its sensibility seems to be very limited, and the assessment of the number of flats in a block of flats can be rather inaccurate in case the interviewer cannot even enter the building (noncontact), s/he has to make a judgment only by looking at its façade, sometimes from a significant distance, or partly or fully hidden behind a fence. The correction of these issues also require careful analysis and cognitive testing in order to deem them ready for subsequent quantitative validation/reliability tests or field use.

To sum up, the presently used variables should be revised. However, despite its flaws, the underlying idea of this set of variables is unique even in international context. That is, not only the type/condition of the dwelling or building in which the dwelling is situated is to be observed as it can be usually found in the commonly used housing unit/neighbourhood observation instruments, but also the close built environment of the building, too. For instance, there are many examples for buildings distinctly unfitting to their built environments, such as a single detached house surrounded by blocks of flats, or a block of flats in an industrial area (apartments for workers were frequently built close to factories in the early 1900's). The necessary revision of the Hungarian variables will have to be made based on the suggestions and results of the literature. Of the housing unit/neighbourhood observation instruments, four seem to be the most significant. However, further examination of the other options is still ahead of us. At the present state of the study, it is clear that Krause's (1998) six-item scale would be unsuitable for the purposes of nonresponse adjustment because it cannot be filled out properly in case of noncontact. Only the four items that do not require observations on the inside of the respondents home could be useful, but they are not too sensitive. Andresen et al's (2006) African American Health Five-Item Neighborhood Assessment Scale (AAH 5-item NAS) based on Krause's scale seem to have a better approach, however, there are items which depend on rather subjective assessments (amount of noise and air quality), which entails that their potential for standardised use

seem to be quite limited, and it contains no item on the type and condition of the building in which the respondent's dwelling is situated. Furthermore, only the overall condition of the residential area of the respondent's dwelling is asked, the type and particular characteristics are not. Andresen et al's (2008) African American Health Seven-Item Neighborhood Assessment Scale (AAH 7-item NAS) is more sensitive and the users (i.e. interviewers) are much better provided with hints on when and how to use the response categories, therefore it is less prone to subjective assessments. However, it still lacks the items on the building in which the respondent's dwelling is situated, and on the type and characteristics of the residential area. The problem with these scales is that their purpose is to measure social disorder, and social disorder only. Significantly longer scales, such as the 18-item NAS scale (see Andresen et al. 2013), or the even more extensive and complex Neighborhood Observation Protocol used in the study of Casas-Cordero (2010) would be unfit for field survey use. Although the latter has the variables on the buildings and residential area that the five- and seven-item AAH NAS scales lack, there are a lot of possible logical doubts that can be raised against it on the level of probable subjectivity in the assessment, completeness and exclusivity of the response categories, scale granularity and so forth, besides the length issue. As Andresen et al. (2013) note, shorter scales have much better potential utility. The next step of our work is thus finding the best succinct measures, creating an adaptation by adjusting them to the specific Hungarian conditions, and complementing them, if necessary, by properly designed variables on the buildings/residential area/close built environment.

Next steps

The original plan on a simple split-sample experiment on the effect of interviewer training has already significantly changed. The first step is to design a new instrument as outlined above in detail, because using the existing variables for nonresponse adjustment would not be feasible due to the severe design flaws.

The subsequent steps will be the following:

2. Cognitive testing of the instrument among survey interviewers, and implementing the necessary corrections on the variables.
3. Devising a brief but efficient training protocol. The extensive trainings on the use of the housing unit/neighbourhood observation variables as described for instance in Andresen et al. (2013) are compelling, but they are not realistic in the HCSO and perhaps in other regular NSI interviewer training settings. Therefore, we plan to rely heavily on novel education solutions. Namely, where applicable, several typical Google Street View situations will be presented to the interviewers instead of using only pictures to enable them making better sense of the response categories. Two versions of the same training material will be made, one for traditional 'classroom' training, and the other for e-learning at home.
4. The interviewers will be divided into two groups. The control group will get a traditional 'classroom' training with tutors, the treatment group will only get the e-learning material.
5. The two groups will have to complete the same test, which consists of assessment of Google Street View situations using the variables. The results of the two groups will be compared in order to decide whether the cost-efficient e-learning material yield the same results as the traditional 'classroom' training, which is also more burdensome for the interviewers.

6. Correlation analysis of the new housing unit/neighbourhood observation variables with key survey variables and nonresponse data using the data to be collected in the next wave of the Hungarian LFS.

In case the after-training test-results are promising, we expect strong correlations and low levels of measurement error that enables the future devising of nonresponse adjustment procedures.

Summary

The abstract of the early-phase research plan submitted for the workshop has changed significantly with the progress of the work. In the light of the acquired knowledge, the existing variables need revision. The adaptation of a ready-made instruments from the literature does not seem to be the right path due to the issues with each possible options. Therefore, new variables have to be elaborated, and then they will have to be subjected to cognitive testing in order to decrease measurement errors, a frequently reported problem with such instruments. A brief and realistic training protocol will then be devised, which will rely on selected Google Street View situations for demonstrating the full range of the rating levels. It will be made in two versions: as traditional 'classroom' training and as a less burdensome and more cost-efficient e-learning at home. The efficiency of the two types of training will be tested on two separate groups of interviewers. The new variables will then be put in field use in the next wave of the Hungarian LFS. In case the auxiliary variables of new housing unit/neighbourhood observation will be found to be strongly correlated with both response behaviour and key survey variables, they will be useful in the analysis of nonresponse and future application of effective weighting procedures. According to the results of Walsh et al. (2013) – provided the new variables perform well on the field –, it can also be anticipated that they will be predictors of level of effort (number of contact attempts and interim refusals) and they may enable the future upgrading of the Hungarian LFS data collection procedures to apply adaptive design.

Questions to the attendants:

- Does any of the countries collect interviewer observations in LFS?
- Which observations are found to be working for nonresponse analysis of the LFS?
- Does any of the countries use interviewer observations to estimate the level of sufficient fieldwork effort?
- Do interviewers get special training on collecting these observations?

References

Andresen, E. M., Malmstrom, T. K., Miller, D. K. & Wolinsky, F. D. (2006): Reliability and Validity of Observer Ratings of Neighborhoods. *Journal of Aging and Health* 18(1): 28-36.

Andresen, E. M., Malmstrom, T. K., Schootman, M., Wolinsky, F. D., Miller, J. P. and Miller, D. K. (2013): Observer Ratings of Neighborhoods: Comparison of Two Methods. *BMC Public Health* 13(1024).

Andresen, E. M., Malmstrom, T. K., Wolinsky, F. D., Schootman, M., Miller, J. P. and Miller, D. K. (2008): Rating Neighborhoods for Older Adult Health: Results from the African American Health Study. *BMC Public Health* 8(35)

Casas-Cordero, V. (2010): *Neighborhood Characteristics and Participation in Household Surveys.* PhD. dissertation, Faculty of the Graduate School of the University of Maryland, College Park.

Casas-Cordero, V., Kreuter, F., Wang, Y. & Babey, S. (2013): Assessing the Measurement Error Properties of Interviewer Observations of Neighbourhood Characteristics. *Journal of the Royal Statistical Society* A 176: 227-249.

Durrant, G. B., Groves, R. M., Staetsky, L. & Steele, F. (2010): Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys. *Public Opinion Quarterly* 74(1): 1-36.

Groves, R. M. & Couper, M. P. (1998): *Nonresponse in Household Interview Surveys.* New York: John Wiley & Sons.

Krause, N. (1998): Neighborhood Deterioration, Religious coping, and Changes in Health During Late Life. *The Gerontologist* 38(6): 653–664.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R. M. & Raghunathan, T. E. (2010): Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-response: Examples from Multiple Surveys. *Journal of the Royal Statistical Society* A 173: 389-407.

Lynn, P. (2003): PEDAKSI: Methodology for Collecting Data About Survey Non-Respondents. *Quality and Quantity* 37: 239-261.

Pickering, K., Thomas, R. & Lynn, P. (2003): *Testing the Shadow Sample Approach for the English House Condition Survey*. Report. London: National Centre for Social Research.

Sinibaldi, J., Durrant, G. B. & Kreuter, F. (2013): Evaluating the Measurement Error of Interviewer Observed Paradata. *Public Opinion Quarterly* 77(Special Issue): 173-193.

Walsh, R., Dahlhamer, J. & Bates, N. (2013): *Assessing Interviewer Observations in the NHIS.* Conference paper presented at the JSM 2013 Survey Research Methods Section.

West, B. T. (2011): *The Error Properties of Interviewer Observations and their Implications for Nonresponse Adjustment of Survey Estimates.* PhD. dissertation, University of Michigan.

West, B. T. & Kreuter, F. (2015): A Practical Technique for Improving the Accuracy of Interviewer Observations of Respondent Characteristics. *Field Methods* 27(2): 144-162.