Kiel

# Working Papers

Kiel Institute for the World Economy

# Guidelines for the Use of Household Interview Duration Analysis in CAPI Survey Management

by Marcus Böhme and Tobias Stöhr

**No. 1779** | July 2012

Web: www.ifw-kiel.de

Kiel Institute for the World Economy, Hindenburgufer 66, 24105 Kiel, Germany

Kiel Working Paper No. 1779 | July 2012

# Guidelines for the Use of Household Interview Duration Analysis in CAPI Survey Management *

Marcus Böhme and Tobias Stöhr

Abstract:

This paper provides evidence based guidance for practical survey work, namely choosing interviewers and their workload. Analyzing a survey of 3568 households obtained through computer assisted personal interviews (CAPI) we find that interviewers learn considerably while the survey progresses. Time requirements for field work increase concavely with sample size which allows larger samples to be realized with a given budget than would be expected in planning such projects with simplistic cost estimates. We find a decrease of interview duration of almost 50 percent which translates into a significant increase of the average hourly wage the interviewers receive. These learning effects cease after around 20 interviews. Based on our results we recommend targeting interviewer training by age and technology-affinity of interviewers for CAPI surveys.

An updated version is forthcoming in *Field Methods*.

Keywords: household surveys, CAPI, learning curves, interview duration, survey planning

JEL classification: C81

**Böhme, Marcus**
Kiel Institute for the World Economy
24100 Kiel, Germany
Telephone: +49(0)431-8814-571
E-mail: marcus.boehme@ifw-kiel.de

**Stöhr, Tobias**
Kiel Institute for the World Economy
24100 Kiel, Germany
Telephone: +49(0)431-8814-367
E-mail: tobias.stoehr@ifw-kiel.de

* We would like to thank Gregory Martin (U.S. Census Bureau) for helpful advice on CSPro during the programming of CELB questionnaire.

# I. Introduction

A large share of empirical research in economics is based on individual or household survey data. For precise and reliable analyses high data quality is a necessary condition. However, data collection is often planned and conducted on an ad hoc basis. In particular, it is often unclear to which extent survey schedules that emerged from the drawing board are relevant in the "messy" reality of field work in developing countries. One aspect that is central to data quality is the number and selection of interviewers. Still, decisions in this regard are often driven by gut instinct or even exogenous if local partners are contracted to conduct interviews. By employing para-data we will demonstrate how survey duration analysis can improve survey planning and provide important insights into the dynamics of survey duration.

The duration of a questionnaire has two important implications. First, it will be important to plan the overall duration and budget of the survey. Researchers are frequently faced with the tasks of budgeting and planning primary data collection. This typically comes after outlining the questionnaire and will typically be subject to patchy information regarding details of implementation. Even after questionnaires have been thoroughly pretested, interviewers will get more used to questionnaires and software over time. Thus the average time needed for one interview will decrease during the first weeks of field work. Interviewers will typically make mistakes or be unsure how to react to idiosyncrasies that were not discussed during their training. With an increasing number of interviews conducted by each interviewer, the occurrence of such difficulties and therefore the duration of interviews can be expected to decrease.

Secondly, the duration of the questionnaire is important for the real wage of the interviewer and therefore might affect the behavior of the interviewer. If interviewers are paid on a per piece basis, a learning curve would suggest that their hourly wage would increase with survey progress until familiarity means that no further learning effect can be achieved. Thus, it would be in the interest of the researcher to set the number of interviewers employed during field work at a level that allows the learning curve to take its full effect for each interviewer. This suggests some minimum workload per interviewers which should not be undercut. Fixed costs per interviewer such as training and material (e.g. computers) mean that this is also relevant for the budgetary efficiency of the field work.

Although empirical evidence regarding differences in interview duration between CAPI and classical pen-and-paper interviewing (PAPI) is mixed and inconclusive (Baker et al. 1995, Fuchs et al. 2000) the use of CAPI holds various advantages. Through the use of automatic routing and validation algorithms during the interview key stroke errors, nonsensical data entry and missing data can be avoided and overall data quality can be improved (e.g. Caeyers et al. 2012). This automatic validation is also relevant for panel surveys were the availability of previous responses reduces the respondent burden (Jäckle 2008). However, one major concern regarding the implementation of CAPI surveys in developing countries is that low computer literacy may restrict the pool of potential interviewers and exclude many experienced interviewers. Using a recently conducted CAPI survey, we demonstrate the importance and size of these learning effects. Furthermore we provide evidence for the role of personal characteristics of the interviewers that need to be considered when introducing CAPI techniques. Our results are therefore especially relevant for surveys that employ CAPI for the first time.

The next section will provide some context regarding the survey at hand. Section 3 estimates and discusses the learning curve and the influence of interviewer characteristics on it. Section 4 puts these results into perspective with survey cost and data quality. Section 5 concludes.

## II. Survey Methods and Data

The evidence presented is based on a large scale household survey that investigates the impact of migration on children and elderly left behind (CELB) in migrant families in Moldova. The sample was drawn from the Moldovan Labor Force Survey using stratified random sampling without replacement providing a nationally representative picture of migrant and non-migrant households with children or elderly. The face-to-face interviews were conducted from October 2011 to February 2012. The questionnaire consists of four modules, a household roster including demographic information for all household members, income and expenditure and other indicators of material wealth as well as a detailed migration section. Individual interviews were conducted with children aged 10-19 and caregivers for children of all ages. These covered topics such as health, education, behavior, and parenting practices. Elderly, defined as individuals above 59 years of age, were interviewed separately as well, covering topics such as health, mobility, and help arrangements. The final sample size was 3568 households with 12333 individual. The household non-contact and refusal rate were 7.9 and 10.6 percent respectively. In addition 3375 caregiver interviews, 1177 child interviews and 2170 elderly interviews were conducted with the main groups of interest of the survey.

To implement the survey a local survey company was contracted in a competitive tender. In order to decrease the data-entry cost and to allow close monitoring of the survey as well as almost instant feedback to interviewers CAPIs were implemented using inexpensive netbooks with long battery life[1]. The questionnaire was programmed using the public domain software package CSPro provided by the U.S. Census Bureau. The CAPI questionnaires contained a pre-programmed skip logic, i.e. automatic routing, and constructive error messages based on validation algorithms that helped to reduce the scope for error. The usage of CAPI also allowed us to include automatic time stamps for each section of the questionnaire. These recorded the exact system time at the beginning and end of each section as well as for the questionnaire as a whole and could not be edited by interviewers.

Interviewers had to complete the household roster before being able to interview individual household members for the caregiver and elderly modules of the questionnaire. We decided to set an incentive for high response rates from individual questionnaires by setting the hourly wage for these about 30 percent higher than for the household roster module. This helped to achieve very high response rates of 92.7 percent for caregivers, 89.7 percent for elderly, and 57.7 percent for children.

[1] Technical Details: 1.83 MHz, 1-2 GB DDR-3 RAM, On-board video and 160+ GB hard disk, 11 inch display diagonal.

A total of 51 interviewers took part in the field work. They were chosen by the survey company based on their experience and the ability to operate computers. Interviewers were on average 31 years old and came from all regions of the country. Due to the focus of our survey on children and elderly people, a majority of interviewers (67 percent) were female. The level of education among interviewers was high with an average of 14 years of schooling and almost half of the interviewers enrolled at university. On average our interviewers had participated in three previous surveys and there was no experience with CAPI. During our survey each interviewer completed an average 70 household modules and 94 individual interviews.

In training groups of 10 to 15, interviewers were introduced to questionnaires and trained to operate the software. We obtained basic demographic information of all interviewers from pairwise interviews they did with other interviewers during the training. Additionally, all interviewers were asked to fill out specific questionnaires regarding their survey experience, attitudes and expectations. These data were checked against records of the survey company to ensure their correctness.

During our survey's field work interviewers could always call their superiors at the main office if they had questions. These would then be clarified instantly either directly or, if required, with the authors who were on the ground during the first months of implementation. Typically, questions were clarified within minutes over the phone and interviewers could continue the household interviews on the spot.

Monitoring data quality and interviewer performance was possible almost instantly, since data were uploaded daily or bi-daily whenever possible. Hence we were able to give interviewers specific feedback quickly, such as additional explanation regarding particular questions when we suspected they had been misunderstood. The aforementioned time stamps and some quality measures allowed us to intervene where data quality seemed at risk. If data quality seemed low or interview duration dropped under critical values per section or question, an interviewer's supervisor was informed who discussed our concerns with interviewers. This typically led to improvements in performance that we could again monitor almost live but also two cases of interviewers being fired. With a classical paper-based survey this would have been all but impossible.

## III. Estimation & Results

The data collected by the interviewers can be written in an unbalanced panel structure. From the time stamps of each questionnaire we establish the exact order by which an interviewer conducted his interviews. Although interviews are not evenly spaced in time as several interviews are conducted per day, followed by several hours of rest or a weekend, we treat these alike by assuming that additional surveys, but not the time elapsed between them, matter for learning.

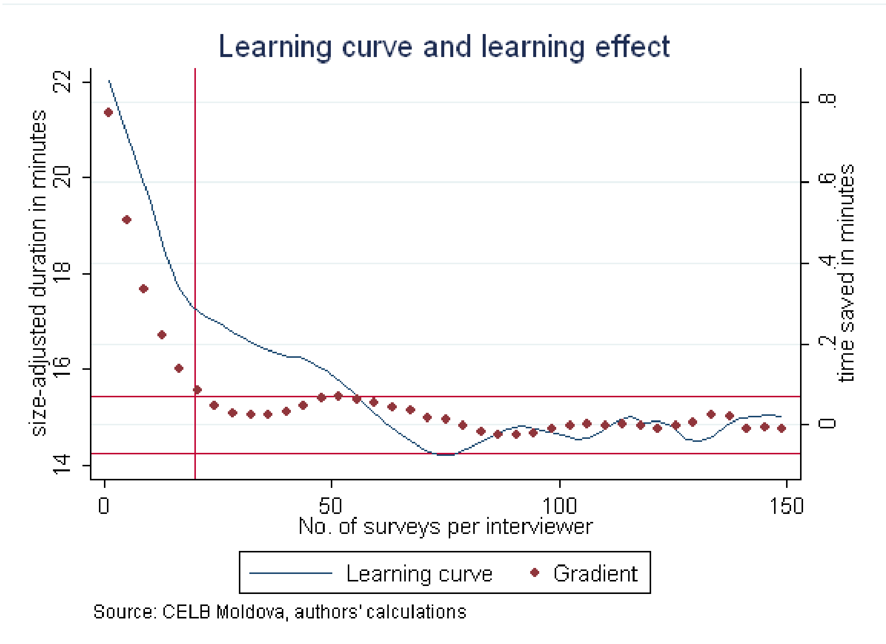
Figure 1 shows the estimated learning curve and its gradient based on a local polynomial estimation. Interview durations in the following are adjusted by dividing the duration of sections which had to be filled out for several individuals per household by the household size[2]. The left

[2] As there will be economies of scale, we use squared terms as additional controls in estimations later on.

hand y-axis reports the minutes used per interview while the right hand axis indicates the decrease in minutes per survey per additional interview. The fitted curve suggests that the statistically significant learning effect ceases after around 20 observations, which is marked by the vertical line. After that there is still some movement of the learning-estimate but the gradient stays within a 99 percent confidence band around 0 which is reflected by the horizontal lines. The relevance of the cumulative learning effect is apparent from the graph. The learning effect reflects that the median size-adjusted surveys take about 15 minutes in contrast to 25.5 minutes at the first interview.

This means that first field tests of a questionnaire overestimate interview duration vastly. Furthermore, even after several days of training that includes applying the questionnaire several times in mock interviews there are sizeable learning effects. The size of the learning effect also highlights how important close monitoring of the survey during the first weeks is. Without help and quick feedback interviewers could continue making mistakes and be left confused, especially in complex surveys like the one at hand.

Figure 1




To quantify the effects suggested by the local polynomial graph in Figure 1 we employed two different approaches. First we used a simple inverse function:

$$\tau_{it} = \alpha_i + \beta_{1i}\frac{1}{x_{it}} + \varepsilon_{it} \ , \qquad (1)$$

where $\tau_{it}$ is the size-adjusted interview duration, $\alpha_i$ captures the long-run minimum duration, $x_{it}$ reflects the chronological rank of the interview of interviewer i during the interview process and $\varepsilon_{it}$ is the error term. There is one major drawback of this functional form. It implies by assumption that 95 percent of the learning effect after the first interview will take place between the first and the 20th interview. To address this concern we use the well-known decay function, which has a less restrictive functional form then the inverse, as a robustness check:

$$\tau_{it} = \alpha_i + \theta_i * e^{-\beta_{1i}*x_{it}} + \varepsilon_{it} \qquad (2)$$

Estimation results for both approaches are given in the corresponding columns of Table 1. All estimations include variables that indicate the size of the household to control for possible biases in the size adjustment of duration and economies of scale. In column (1) the learning curve is estimated without interviewer specific covariates. The coefficient $\beta$ indicates a total learning potential of 15.88 minutes at the limit compared to the first survey. The constant indicates a fading out at 13.54 minutes per interview – a decrease of more than 50 percent in interview duration compared to the beginning. The less restrictive non-linear estimation of the decay function in column (2) indicates a learning potential of 11.06 minutes approaching a constant of 13.82 minutes. The full effect also arrives later, with only 86.6 percent realized after 20 interviews. The decay provides a better fit but has to be estimated numerically so that convergence depends on the starting values. For the sake of parsimony, we prefer the inverse function over the alternative in column (2).

Table 1: Estimation of Learning Curve in Face-to-Face CAPI interviews

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Inverse | Decay | Inverse | Inverse | Inverse | Inverse |
| α | 13.54*** | 13.82*** | 15.32*** | 14.85*** | 15.79*** | 14.63*** |
| | (0.54) | (0.538) | (1.12) | (1.63) | (1.69) | (1.06) |
| β | 15.89*** | -.094*** | 14.53*** | 16.44*** | 5.32 | 6.07 |
| | (1.33) | (0.011) | (1.26) | (1.60) | (7.32) | (3.89) |
| θ | | 11.06*** | | | | |
| | | (0.93) | | | | |
| Female | | | | -1.86*** | -2.00*** | -1.75*** |
| | | | | (0.38) | (0.42) | (0.38) |
| Age | | | | .073*** | .056*** | .044*** |
| | | | | (0.017) | (0.018) | (0.016) |
| Yeduc | | | | -.058 | -.079 | |
| | | | | (0.04) | (0.09) | |
| Experience | | | | -.056 | -.065 | |
| | | | | (0.04) | (0.04) | |
| Enrolled | | | | -2.83*** | -2.82*** | -2.67*** |
| | | | | (0.49) | (0.54) | (.47) |
| β · age | | | | | .310* | .336** |
| | | | | | (0.17) | (.13) |
| β · enrolled | | | | | -.032 | |
| | | | | | (4.10) | |
| β · female | | | | | 2.75 | |
| | | | | | (3.51) | |
| Interviewer fixed effects | No | No | Yes | No | No | No |
| % of learning realized after 50 interviews | 98 | 98 | 97.6 | 98 | 98 | 98 |
| Adj $R^2$ | 0.31 | 0.33 | 0.43 | 0.36 | 0.36 | 0.36 |
| F | 89.9 | - | 30.6 | 54.9 | 48.2 | 58.86 |
| N | 3384 | 3384 | 3384 | 2312 | 2312 | 2312 |

Notes: Author's calculation based on CELB Moldova 2012. Additional controls for household size, number of children, number of elderly, and number of migrants. Standard error in parentheses; *** p<0.01, ** p<0.05, * p<0.1, robust standard errors used in linear estimations.

If we expect learning curves to vary across interviewers, it is likely that individual characteristics determine the speed of learning as well as the exact position of the curve. Hence, we introduce interviewer fixed effects as a double check in (3) and add interviewer specific covariates in columns (4)-(6). Parameters such as the age or sex of interviewers shift the level of the learning curve up or down when significant while interactions with the number of interviews (β) provide changes to the vertical size of the learning curve and thus the learning effect itself. We find that female interviewers are faster by about 12 percent of the adjusted interview time but do not have a different speed of learning. Older interviewers need longer for interviews than younger interviewers. The coefficient translates into 3 percent per survey for an additional 10 years of age at the estimated long run survey duration.

Column (5) and (6) show that the learning curve is in fact driven by older interviewers. The general parameter $\beta$ is not significant any more when adding interaction terms with interviewer characteristics. However, a linear interaction of age and beta suggests that learning takes place the older an interviewer is. Given our subjective observation from the training, we interpret this as younger interviewers being already further towards the flat part of the curve when they start interviewing, either because they are more impatient or because of their relatively high technology affinity. Older interviewers thus start from a higher base but catch up later on.

We do not find a significant influence of education on the learning curve. However, interviewers who are enrolled at university have significantly lower interview durations throughout the learning curve. We suspect that this is caused by university students being the most computer literate in our interviewer population even compared to similarly aged non-students. There is no increase in the speed of learning after controlling for the effect of age. When taking into account these factors also the number of surveys conducted before ("experience") does not matter for interview duration any more.

## IV. Consequences for Survey Cost and Data Quality

Since interviewers are paid per item, i.e. completed questionnaire, the learning curve implies that the wage of interviewers is dynamic. In the beginning, the hourly wage per household interview was equivalent to about $2.05 excluding travel and other setup cost for the household interview[3]. On the flat part of the learning curve it had risen to about $3.59. Similar learning curves from different bases occur for the individual questionnaires. Of course, the fixed component of an interviewer's work is considerable, so these hourly wages do not translate completely into real wages. Let us give an example of how this dynamic affects budgeting. Assuming a fixed time mark-up of as high as 50 percent per interview, the hourly wages would still increase by 42.7 percent for the household interview and by an even larger margin for individual questionnaires. If we furthermore assume interviewers agreed to take part based on a good estimate of their overall hourly wage including the mark-up[4] and would accept any equal or higher offer, we can infer a counterfactual cost scenario.

[3] The following figures are rough approximations and exclude figures such as overhead to keep private information of the survey company from competitors.
[4] About $2.08 which translates into $356 per month and is reasonably close to local wages.

With a fixed cost of $6.000 for the whole survey consisting of $100 per interviewer for materials, training, management, and other expenses as well as $2000 overhead, our baseline translates into 40 interviewers finishing 4000 surveys at a cost of $36.000[5]. If one had set the initial per piece rate at the competitive hourly pay without considering the learning effect, one would have paid an overall $40.773[6]. Adding more interviewers would also make the survey more expensive. Even when the competitive wage is known in advance adding 20 interviewers would then not only increase the fixed cost by $2.000, but also require the initial per piece rate to yield at least the estimated hourly wage over the decreased workload. This would give an estimated cost of $38.265 or $42.774 without adjusting for the learning effect.

Our exemplary budget calculations show what a sizeable role the learning effect can play for survey costs. From the estimation we derive some more advice for practical work. First, depending on the age and technology-affinity of interviewers, training should be targeted. We find that potential learning effects were already achieved during training for younger, better educated interviewers, who took part in the training more enthusiastically. Second, during training one should observe in detail who is faster than the average and who is slower. If quality is given, people who are faster should be allowed to continue with a higher pace where possible, e.g. when filling out questionnaires in mock interviews with partners. Slower interviewers should get close supervision and extra training in tasks they struggle with.

Still, working at a slower pace should not be taken as an argument to employ only youngsters for interviewing. Rather older, less computer literate interviewers should receive additional training and close monitoring during the first days of field work. This way they can improve their performance quickly and overcome difficulties. Interviews visited by the authors gave the impression that due to their additional experience in interviewing but also better ability to relate to respondents (as the age gap is lower). This advantage means that these interviewers are likely to produce some of the best data. These are important considerations when introducing CAPI techniques in developing countries.

These learning effects cannot be discussed independently from other factors influencing the data quality. The trade-off between the number of interviewers and possible interviewer effects in the form of intra-class correlations needs to be considered. When for some reason field work has to be finished very quickly, there is no other solution than sending out more interviewers. This however would mean less efficient resource use. The statistical quality literature adds another facet (Groves, 2004, p.364): low workloads per interviewer are seen positively because individual interviewer effects (e.g. framing or ability to probe) have little influence on aggregates. Measuring them, however, is not always easy. For example, when the area covered by a survey is too large for interviews to be randomized perfectly across interviewers, spatial effects can be confounded with interviewer effects.

Intra-class correlations reflect the relation between the responses of one interviewer in relation to the correlation of all surveys. If the whole survey would be conducted by just one interviewer, the ICC would be maximal as the interviewer effect introduced by the only interviewer affected all interviews. If one interview per interviewer was conducted, the interviewer specific effect would be minimal but there would be no learning effect at all. Figure 2

[5] One household interview and 1.95 individual interviews expected per household.
[6] Calculations are available on request from the authors.

shows the development of the ICC over the survey horizon. Finishing additional surveys interviewers move along the x-axis to the right. Although theoretically the ICC must increase as more surveys are finished per interviewer, the ICC remains flat. This was possible by continuously informing the survey company about interviewer performance so they could shift workload from low to high quality interviewers.

Figure 2:

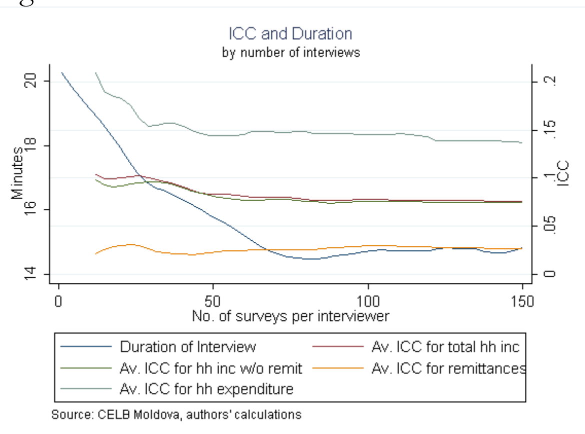
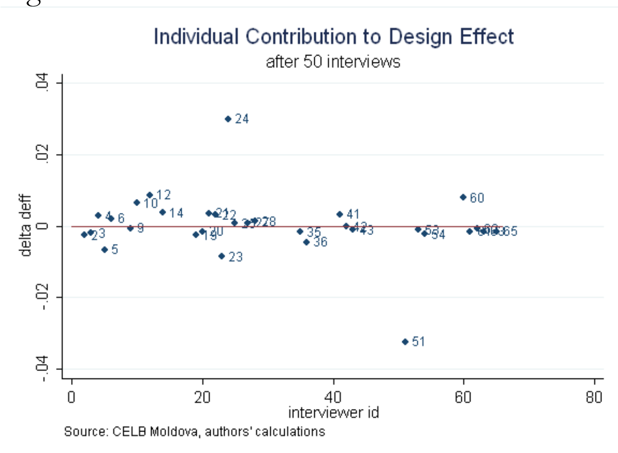



Figure 3:




The ICC can also be used to spot problematic performance during the implementation of the survey. Figure 3 shows the individual contribution of each interviewer to the overall design effect[7] at 50 conducted interviews. The indicator calculates the change in the design effect through a particular interviewer following Groves (2004) as

$$\Delta \mathrm{deff}_i = \frac{1}{n}\{(1 + \overline{icc} \cdot (m-1)) - (1 + \mathrm{icc}_{-i} \cdot (m-1)\} = \Delta \mathrm{icc}_i \cdot \frac{(m-1)}{n} \quad ,$$

where $n$ is the number of relevant interviewers, $m$ is the workload and $\Delta\mathrm{icc}$ is the ICC minus the ICC when excluding one interviewer at a time. We interpret this as the individual contribution to the design effect of a particular interviewer. High values suggest that an interviewer has a stronger than average correlation between values (e.g. interviewer 24 in figure 3). Thus the interviews conducted by this person decrease overall data quality as captured by the design effect. A low value suggests that there is particularly low correlation between responses in different interviews. Outliers such as interviewer 51 in figure 3 could suggest problems as well. In both cases, the researcher should look for explanations before confronting the interviewer. If she works in very similar areas, such as only in particular neighborhoods, whereas others interviewers had multisite assignments, the high ICC contribution are not necessarily attributable to interviewer performance. If there is no such explanation, one should monitor the data delivered by this interviewer in detail and intervene if necessary.

[7] A standard indicator reflecting the difference in variances introduced by the sampling structure and by differing variances (e.g. due the similarity of respondents from the same area or through interviewer effects) within subgroups of the sample. It reflects the loss of information from sampling within clusters compared to simple random sampling that is more expensive in face-to-face interviews.

## V. Conclusion

This paper has shown how survey work underlies large learning effects that should be taken into account when planning the number of interviewers and their individual workload during the setup of large household surveys. We have provided evidence for a nonlinear learning curve that suggests that the workload per interviewer should not be too low. We have argued that it is the right choice of interviewers and close monitoring during the survey which matter for data quality. Our results suggest that one should not shy away from training less computer literate interviewers CAPI techniques if they are good interviewers. Although for external validity additional studies are required, our analysis provides evidence to make more informed choices when facing a trade-off between interviewers' prior experience and their ability to catch up.

## VI. References

Baker, R. P., Bradburn, N. M. and Johnson, R. A. (1995): Computer-assisted Personal Interviewing: An Experimental Evaluation of Data Quality and Cost. Journal of Official Statistics, 11(4):413-431.

Caeyers, B., Chalmers, N. and 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

Fuchs, M., Couper, M. P. and Hansen, S. E. (2000): Technology Effects: Do CAPI or PAPI Interviews Take Longer? Journal of Official Statistics, 16(3): 273-286.

Groves. R. M. (2004): Survey Errors and Survey Costs, John Wiley & Sons, Hoboken, New Jersey.

Groves,R. M. and Heeringa, S. G. (2006): Responsive design for household surveys: tools for actively controlling survey errors and costs, Journal of the Royal Statistical Society, 169(3): 439-457.

Jäckle, A. (2008), 'Dependent Interviewing: Effects on Respondent Burden and Efficiency of Data Collection', Journal of Official Statistics, Vol. 24, No. 3, pp. 411-430.