Survey Research Methods (2008)
Vol.2 , No.2 , pp. 61-73
ISSN 1864-3361
http://www.surveymethods.org

A Note on Interviewer Performance Measures in Centralised CATI Surveys

Oliver Lipps
University of Lausanne

Interviewer performance with respect to convincing sample members to participate in surveys is an important dimension of survey quality. However, unlike in CAPI surveys where each sample case 'belongs' to one interviewer, there are hardly any good measures of interview performance for centralised CATI surveys, where even single contacts are assigned to interviewers at random. If more than one interviewer works one sample case, it is not clear how to attribute success or failure to the interviewers involved. In this article, we propose two correlated methods to measure interviewer *contact* performance in centralised CATI surveys. Their modelling must take complex multilevel clustering effects, which need not be hierarchical, into account. Results are consistent with findings from CAPI data modelling, and we find that when comparing effects with a direct ('naive') measure of interviewer contact results, interviewer random effects are largely underestimated using the naive measure.
Keywords: non-response, interviewer performance, contact level, multiple membership, cross-classified multilevel, random assignment

Introduction

In CAPI surveys, interviewers usually work all contacts on a sample member until the latter is either ready to complete the interview, refuses, or leaves the interviewer with a pending appointment. In the case of a CAPI survey, the assignments of the sample members contacts to interviewers can therefore be schematised as follows (see Figure 1).

Here, it is straightforward to measure interviewer performance in convincing sample members to participate in the survey, simply by calculating the mean number of *finally* participating sample cases worked by the interviewer. Methodologically, the only problem is a possible confusion of area and interviewer effects, because interviewers may obtain more or less 'difficult' areas.[1]

In centralised CATI surveys, separation of these effects is guaranteed by the randomised sample case interviewer - contact assignment (see Figure 2).

Here, although the problem of interviewer-area confusion is usually resolved (unless, inter alia, interviewers are used according to the dialect spoken in an area), it is not obvious how to measure interviewer performance. Most existing approaches focus on single interviewersample member contact results, where generally only cooperation rates based on first contacts are retained (e.g. Mayer and O'Brien 2001). The reason is "...to avoid contaminating the measure with the performance of a previous interviewer" (Durand 2005:763). This is in line with analysis by Groves and Couper (1998:256), who conclude that for the later contacts, the attributes of the prior contacts are the most important indicators of cooperation likelihood. However, if a final disposition is not achieved after the first contact, as in refusal conversion cases or if appointments are made, this approach is not applicable. Recent models therefore assign bonus or malus points to *transitions* (Durand 2005) achieved; i.e., they assess single contact results dependent on the previous contact result of the sample case. In addition, single contact results are directly assessed, with the result of the previous contact controlled for in regression models (Lipps 2007b).

These measures suffer from various problems:

- Arbitrariness of 'point' assignment according to call or contact achievement. Durand, for example, attributes one credit point for a completed interview from a previous appointment (2005:766). However, this procedure is not very convincing for appointments with a fixed date and time. No special interviewer performance is required to conduct a standardised interview at a fixed date. The achievement is rather to convince the sample member to fix a date and time for an interview. However no points are attributed for this achievement. Moreover, the degree of bindingness of appointments may vary widely. Lipps (2007b) shows that there is a substantial difference in the probability of finally completing a case, depending on whether a *vague* or a *fixed*[2] appointment has been made in the Swiss Household Panel Survey. In addition, the *order* of the contact on a sample case plays a role: after a first fixed appointment with the target person has been agreed, 88% of all household interviews are fi-

Contact information: Oliver Lipps, Swiss Foundation for Research in Social Sciences (FORS), Lausanne, Vidy, CH - 1015 Lausanne, oliver.lipps@fors.unil.ch

[1] In the British Household Panel Survey (BHPS) wave 2, an interpenetrated sample experiment has been performed on a subsample in order to be able to separate interviewer and area effects (O'Muircheartaigh and Campanelli 1999).

[2] Fixed means with a fixed date and time for the interview.

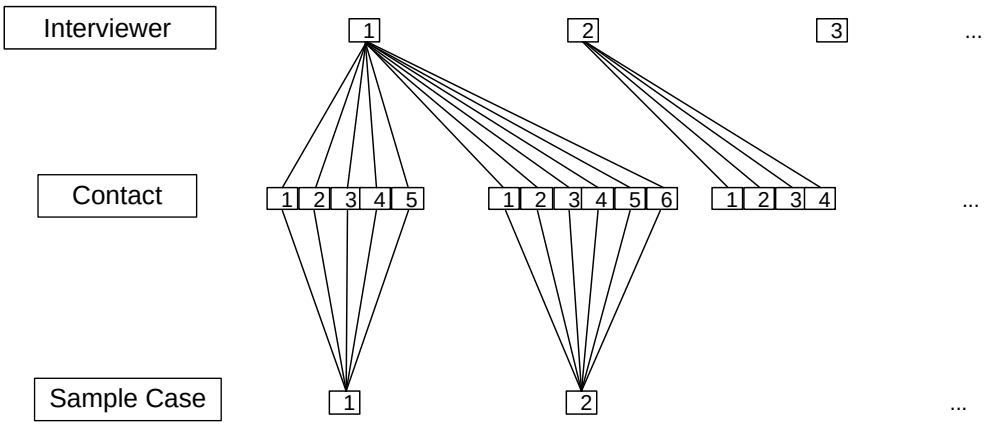

Figure 1. Interviewer-Sample Case Assignments via Contacts in CAPI Surveys

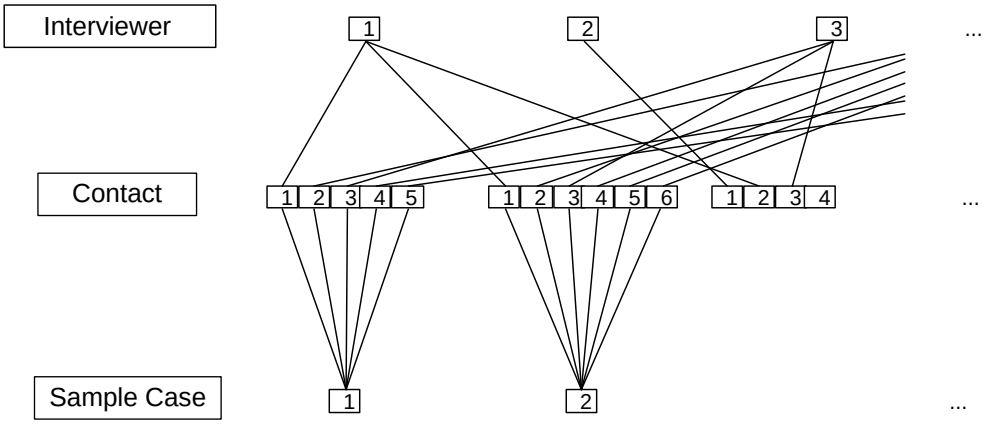

Figure 2. Interviewer-Sample Case Assignments via Contacts in Surveys with Random Assignment

nally completed. This probability of final participation decreases continuously with the number of fixed appointments with the target person: to 84% after the second, 82% after the third, 76% after the fourth, 67% after the fifth fixed appointment, etc. A similar picture emerges after vague appointments with the target person: after the first vague appointment, 71% of the households are finally completed, 66% after the second, 64% after the third, 57% after the fourth, 51% after the fifth, etc. This example shows the difficulty of assigning credit points after a certain contact result.

- Clustering effects of contacts for sample members with interviewers are not considered. For example, in Durand's analysis (2005), a multilevel model for growth (with a random effect of time) with interviewers as second level is used, in order to assess interviewer learning effects on performance over time. However, sample member clustering effects can be assumed to be much higher but are not taken into account. There is also no consideration made for the complex cross-clustering of contacts within interviewers and sample cases (see Figure 2).

Performance Measures for Centralised CATI Surveys

'Cooperation' performance

We define the first performance index by referring the contacts to the specific contact *target*, namely the survey participation of the sample case. Therefore, a straightforward way to measure interviewer performance is to define *final* participation of the sample member as binary performance index, and relate it to the assigned contacts. I.e., all contacts done on a participating case would be assigned a '1' if the sample member finally participates, '0' if not.

The first measure interprets each contact of the interviewers involved in working the sample case as one partial contribution to the participation of the sample member. We call this approach 'cooperation', because the target is to convince the sample case as a complete unit to participate, irrespective of the outcome of single contacts.

'Refusal Avoidance' performance

To make things more complicated, survey research theory might suggest a different measuring approach. Groves'

and Couper's concept of "maintaining interaction" with sample members (1998:243 ff.), which has already proved successful in CAPI surveys, might form the basis of such an approach. This concept is based on the strategy of avoiding termination of the interaction during initial contacts (p. 249), because "the odds of success are increased with the continuation of the conversation" (p. 250). If this is accepted as interviewer guidance to be strictly adhered to in terms of a single contact, it would mean that the interviewer tries to minimise the odds of a 'no' rather than to maximise the odds of a 'yes'. Thus, a high interview performance in a randomly assigned setting could also mean that the interviewer does *not obtain a refusal* from the sample member. This concept includes the ability of "stepping back" (e.g. Hox et al. 1998:174) as one possible interviewer tactic to adequately react to initially reluctant individuals.

In the second measure we thus understand high interviewer contact performance as not obtaining a refusal. Accordingly, we call this approach "refusal avoidance" (Groves and Couper 1998, Mayer and O'Brien 2001).

Control: 'Optimisation' performance

Finally, we use a direct or 'naive' interviewer performance indicator as a control measure. The idea is that interviewers usually try to optimise single contact outcomes per se. As Sonnentag and Frese state: "... because teams are composed of individuals, team processes and team performance cannot be completely understood and improved without taking individual performance into account" (2002:17).

In order to approximate the performance of a single contact, we calculate the mean probability of sample case cooperation by contact result (see Lipps 2007b). The trivial values are contacts resulting in a completed interview (optimisation performance=1) and refusals (optimisation performance=0). Concerning the more interesting intermediate contact results, we distinguish between fixed appointments with the target person and an agreed contact date and time, fixed appointment with another person and an agreed contact date and time, and vague appointments. Because in this setting the interviewer tries to optimise his or her single contact outcome, we call this approach 'optimisation'. Because of the averaging over all sample cases, there is concern that the optimisation index - much more than the other two indices - contains only a small part of the true interviewer effects.

Before we model the three performance indices, the distribution of the optimisation index is depicted in table 1. We use data described later (SHP/SILC 2005/2006), separated by regular and refusal conversion fieldwork phases. For example, after a vague appointment with a target person has been agreed, 59 percent of these sample cases are finally completed. For the sake of completeness, we also depict the trivial cases, completion and refusal.

Table 2 contains the respective figures for the refusal conversion phase.

Because the optimisation index is measured as probabilities, we treat this variable as binomially distributed later in the models (Browne 2005).

Relationship of performance indices

Because the three performance indicators all intend to measure the same thing interviewer performance we expect them to be positively correlated. Even more than this, some contacts result in the same index value by definition: if an interviewer performs well on the cooperation index, s/he necessarily avoids a refusal. Similarly, a refusal implies a cooperation of 0. On the other hand, a refusal avoidance other than a cooperation may not necessarily mean a cooperation of 1, because another interviewer may still obtain a refusal by the sample member.

Table 3 lists the correlation matrix between the three performance indices, averaged over each interviewer.

As expected, we find positive correlations between the performance indices, with different degrees of correlation. Generally, high and highly significant correlations exist within either the regular or the refusal conversion fieldwork phase. For example, interviewers who perform well on the cooperation index during the regular phase, also do so on the refusal avoidance index during the same phase (R = .58). Correlations are not so high across different fieldwork phases: for example, the performance of interviewers on the refusal avoidance index during the regular phase has an insignificant (R = .05) correlation with performance on the cooperation index during the refusal conversion phase. Note however that the correlations across phases refer only to the subsample of interviewers who conduct refusal conversion contacts, with a supposedly higher performance also during the regular phase.[3]

Also interesting are the correlations across the two fieldwork phases on the same index: their magnitudes range from an insignificant R = .18 (cooperation) to a significant R = .28 (refusal avoidance). This means that interviewers tend to perform slightly better during the refusal conversion fieldwork phase if they already did so during the regular phase. This holds despite the positive interviewer selection for the refusal conversion phase mentioned above.

Modelling Interviewer Performance

In the modelling step, we are interested in the magnitude of fixed and random effects on interviewer and sample case level, using the three performance indices as dependent variables. Given that the indices all aim to describe the same thing, we would expect that the coefficients are similar.

Previous research has used CAPI data to analyse interviewer effects on sample member participation (Hox et al. 1991, Groves and Couper 1996, 1998, Japec 2005). This research has shown that considerable interviewer effects on survey cooperation exist, so we would expect that interviewer random effects would be significant for our measures. However, due to construction, it is probable that a

[3] The appointment of interviewers to conduct refusal conversion contacts is up to the survey agency. Exact selection methods are not known.

Table 1: Sample Case Cooperation Probability by Contact Result. Data: SHP/SILC 2005/2006 Sample. Regular Fieldwork Phase

(N=39,207 Contacts, 2005/2006 SHP/SILC) Contact Result	N contacts	% contacts	Sample Case Mean Cooperation Probability
Completed Interview	10,200	26	1
Refusal	2,686	7	0
Fixed Appointment with target person	6,375	16	.86
Vague Appointment with target person	16,762	43	.59
Fixed Appointment made with another person	3.171	8	.65
All Contacts	39,194	100	.70

Table 2: Sample Case Cooperation Probability by Contact Result. Data: SHP/SILC 2005/2006 Sample. Refusal Conversion Phase

(N=39,207 Contacts, 2005/2006 SHP/SILC) Contact Result	N contacts	% contacts	Sample Case Mean Cooperation Probability
Completed Interview	888	15	1
Refusal	2,142	36	0
Fixed Appointment with target person	562	9	.70
Vague Appointment with target person	1,929	33	.37
Fixed Appointment made by another person	398	7	.37
All Contacts	5,919	100	.36

large portion of true interviewer variance on sample case participation is not captured by our measures. This is likely to hold especially for the optimisation index. As to fixed interviewer effects, it is usually hard to identify significant variables (Groves and Couper 1998, Pickery et al. 2001, Japec 2005, Lipps 2007b). If significant at all, main effects of interviewers are likely to be weak (Groves and Couper 1998). The most important effects of interviewers on cooperation seem to be training and experience (Snijkers et al. 1999, Hox and de Leeuw 2002). Groves and Couper state that "most of the acculturation process of producing effective interviewers occurs during training on the job" (1998:195).[4] Although the turnover in CATI is relatively high, even relatively short experience should have an impact. This can be expected because "performance initially increases with increasing time spent in a specific job and later reaches a plateau" (Sonnentag and Frese 2002). Therefore we use interviewer experience measuring covariates and survey related indicators in order to model the three modelling approaches.

For each of the three modelling variables, we build three subsequent models: first an intercept only model, which allows for calculating the variance portions on the level of the sample cases and the interviewers. In addition this model yields a baseline deviance statistic, which can be used to assess the model improvement by including fixed effects. In a second step, we include sample characteristics variables explaining the part of the total variance due to panel and sample cohort membership effects, which serve as controls. In the third step, we include fieldwork and interviewer experience characteristics, along with outcome characteristics of the previous contact for the optimisation model. It is the portion of the interviewer variance reduction between the second and the third step in the different models, and the coefficients of the covariates entering the third step, which we are especially interested in. The interviewer experience variables include whether the interviewer is already in his/her second panel year, and the number of contacts s/he already worked during the fieldwork period. We control the difficulty of accessing sample members measured by the number of the contact on the sample case (optimisation) and the total number of contacts on a sample case until final disposition (all indicators), the working shift at which the contact takes place, and the elapsed number of days in the fieldwork period. In addition, we are interested in the question of whether it is advisable to have the same interviewer conduct subsequent contacts. Rendtel et al. (2004) report highly positive response effects from interviewer continuity between waves for the European Community Household Panel. However, Campanelli and O'Muircheartaigh (1999) did not find such effects in a subsample of the BHPS.

In the cooperation index models we use the sample member's cooperation behaviour outcome as a constant dichotomous variable over all contacts on this sample member within one fieldwork phase.[5] In the parlance of multilevel modelling, we have a non-hierarchical multiple membership setting (e.g. Fielding and Goldstein 2006): each lowest level unit (sample case) is a member of possibly more than one higher level unit (interviewer). The (single) outcome on one sample member thus has contributions from possibly more than one interviewer. Interviewer related effects can be conceptualised as weighted contributions of the interviewers

[4] Japec (2005), however, reports findings that do not show a positive relationship between interviewer experience and response rates.

[5] Equal to 1 if the sample case finally cooperates, otherwise 0.

Table 3: Correlation of Performance Indices. Data: SHP/SILC 2005/2006 sample

Correlation Coefficient (N Interviewers) Significance Level	Regular Fieldwork Phase (N=202)			Refusal Conversion Phase (N=69)		
Modelling Approach/Fieldwork Phase	Coop	RA	Opt	Coop	RA	Opt
Cooperation (Coop): regular phase	1					
Refusal Avoidance (RA): regular phase	.58	1				
	.000					
Optimisation (Opt): regular phase	.74	.76	1			
	.000	.000				
Cooperation (Coop): refusal conv. phase	.18	.05	.16	1		
	.136	.691	.195			
Refusal Avoidance (RA): refusal conv. phase	.14	.28	.21	.73	1	
	.269	.020	.085	.000		
Optimisation (Opt): refusal conv. phase	.22	.21	.23	.89	.86	1
	.073	.085	.060	.000	.000	

working on that sample case. We set the weights according to the effort necessary to work the case and the suspected effect of the interviewer on the case: the n^{th} contact on a sample case is given a weight of 1/n. We thus take the increased difficulty of sample cases requiring more contacts to be finalised into account. To estimate the fixed and random coefficients of the multiple membership models, we use the Markov Chain Monte Carlo (MCMC) estimation technique (Browne 2005), which is implemented in the MLWin Software.[6] If, as in the second or third modelling approaches, single contact results are to be analysed, cross-classified multilevel models are the modelling of choice (e.g. Fielding and Goldstein 2006). Here, contacts are clustered in sample cases, but sample cases are not clustered in interviewers (see Figure 2). Finally, the cooperation and the refusal avoidance indices are modelled as logistically distributed, with the optimisation index as a binomially distributed variable.

Data

We use call (process) data from two 'multi-purpose' household panel surveys, conducted in Switzerland during the years 2005 and 2006. More specifically we use data from:

1. the Swiss Household Panel (SHP), an ongoing, nationwide, yearly conducted centralised CATI panel survey, which started in 1999 with slightly more than 5000 households;
2. the Swiss pilot of the Europe-wide Survey on Income and Living Conditions (SILC).

In each year, both surveys first ask the household composition together with the relationships between all household members, and the basic socio-demography of the household reference person in the grid questionnaire. Preferably, the household reference person should be the same individual across years. If, however, the previous year's reference person is not available, another adult person in the household who is knowledgeable enough about the household can replace him/her. The grid questionnaire takes three to ten minutes to complete, depending on household size and complexity of relationships. After filling the grid, a household related questionnaire is to be completed (about 10 minutes), again by the reference person. After the household related information is given, each household member from the age of 14 years on has to complete his/her own individual questionnaire (about 35 minutes). We restrict our analysis to the first step, i.e. the household grid level response, leaving aside the subsequent household and individual questionnaire responses.

Due to high attrition of former respondents (Lipps 2007a), the SHP recruited a refreshment sample in 2004, representative of the Swiss residential population. For the Swiss SILC pilot, the first wave was conducted in 2004 in parallel to the SHP, by the same survey agency, also using CATI mode, with a partial overlap of the interviewers involved. The questionnaires of the SILC and the SHP are almost the same with the grid and household questionnaires almost completely, and around 60% of the questions of the individual questionnaire being identical. A random half of the pilot SILC households sampled and first interviewed in 2004 was asked to take part a second time in the subsequent year. Also in 2005 a new, smaller SILC sample was drawn and interviewed. The main difference between the two surveys from the sample members' point of view is twofold:[7]

1. the SHP sample members are informed about the structure, but not the exact duration of the survey. According to funds available, they are told that the survey will go on at least for another two years.
2. the sponsors of the SHP are the Swiss National Science Foundation and the University of Neuchâtel, which are both research institutions. By contrast, the Swiss Federal Statistical Office acts as both organiser and sponsor of the SILC survey. The SILC can therefore primarily considered to be government based.

Each year, after the regular fieldwork phase is finalised, an attempt is made to convince the sample members who refused to answer the survey to complete it during the refusal

[6] `http://www.cmm.bristol.ac.uk/MLwiN/index.shtml`

[7] See Graf and Tillmann (2005) for details.

conversion phase. Generally, all refusals at the first stage are re-contacted unless a written refusal is sent to the Swiss Household Panel, or the centre's survey manager considers recontacting to be hopeless.

The number of contacts on a household until final disposition (cooperation or refusal) is in principle not limited in both survey stages, but it is also at the discretion of each centres[8] survey manager to decide not to make further attempts to contact a household. Thus some households remain 'unworked' in the sense that either they cannot be contacted or that a vague or fixed appointment is still pending. The latter can be considered a (soft) refusal. These are, however, very rare cases; in the data used the maximum number of contacts in order to work a household grid is 70 during the regular fieldwork phase and 28 during refusal conversion.

To summarise the 'pre-field' variables in the model, we distinguish the following samples and survey years. First in the survey year 2005:

- the original SHP sample, then in its seventh wave (SHP I)
- the SHP refreshment sample, then in its second wave (SHP II)
- the original SILC sample, then in its second (and last) wave (SILC I)
- the SILC refreshment sample, then in its first (and last) wave (SILC II)

and in the survey year 2006:

- the original SHP sample, then in its eighth wave (SHP I)
- the SHP refreshment sample, then in its third wave (SHP II)

Because we expect both different random and fixed effects for the regular and for the refusal conversion fieldwork phase, we build separate models. Interviewers who conduct less than ten contacts during a respective fieldwork phase are omitted from the analysis. During the regular fieldwork phases, 39,194 contacts were made on 8,745 households by a total of 202 interviewers; during the refusal conversion phase, 5,919 contacts were made on 2,509 households by 69 interviewers. We can assume that interviewers who are appointed to conduct refusal conversion attempts are those who had already proved good performance with the SHP/SILC responding households during the regular phases.

Modelling Results

The results of the MCMC estimated multiple membership and the cross-classified multilevel regression models of the three interviewer performance measures are listed in Table 6 and Table 7. We discuss the modelling results of the first two performance indicators, and use the results of the optimisation indicator primarily for comparison purposes. Looking at the deviance statistics development, we realise immediately that both the models of the regular and the refusal conversion fieldwork phase improve significantly when the two covariate blocks ('prefield' and 'postfield' variables) are added.[9] This effect is especially strong in the refusal avoidance model during the refusal conversion phase after the inclusion of the postfield variables block.

The first independent variable ("Swiss German Part") distinguishes the two interview centres with the language regions. As to the sample considered, and as expected, contacts in the original SHP sample (seventh/eight wave) show the highest performance, and contacts in the SILC II sample (first wave) the worst. This is due to the much longer panel membership ('panelisation') of the SHP I survey members. There are some differences between the SHP II (second/third wave) and the SILC I (second wave) samples; however, it is not the case that one of these samples performs better on both indices. This shows that the fact that the sample member knows about the structure of the survey (SHP II), or the kind of sponsor does not significantly affect contact performance. The survey year variable coefficients emphasise the importance of panelisation effects on contact performance.

It is the third models (post-field) that we are mostly interested in. In all models contact performance significantly worsens with fieldwork time. This is to be expected since the more difficult cases are usually reached later and they take longer to be worked.

Contact time of day is more important during refusal conversion; a contact during the evening shift is in general less successful, while contacting a household on afternoons has positive effects on refusal avoidance. Evening contacts affect refusal avoidance in a negative way during the regular phase. We speculate that the effects of time of interview on performance are a consequence both of reaching differently predisposed households at certain times and of the different performance quality of interviewers working the different shifts. We test this hypothesis by including the time of interview in the pre-field models, and compare interviewer and household random effects with those from the pre-field models. Surprisingly, at least in the regular phase, only the interviewer random effects decrease, while the sample case random effects remain the same. This means that the effects from different times of day are entirely due to the different performance of the interviewers working the different shifts.

The total number of contacts on a household during the regular fieldwork phase has a highly significant negative effect on contact cooperation results, and a highly significant positive effect on refusal avoidance results. This latter finding holds especially for the refusal conversion phase, and is in line with the "maintaining interaction" concept. It is probably the case that some interviewers might have followed the "stepping back" strategy. The negative effect on the cooperation indicator is most probably due to the higher difficulty

[8] The interviews are conducted from two centres: Berne, mainly responsible for the Swiss-German speaking area, and Lausanne, mainly responsible for the French and Italian speaking parts of Switzerland.

[9] The difference of the deviance (= -2^* Log Likelihood) statistics is approximately chi^2 distributed with the number of additional variables as a degree of freedom. Note that the likelihood estimate is only approximate for discrete models.

to convince cases who are reluctant and thus require more contacts.

Using the same interviewer for the next contact on a sample case has no effect during the regular phase, and a positive effect during refusal conversion. It is probably not until the more problematic refusal conversion phase that respondents begin to have confidence in the interviewers given the few possible tools of communication available over the phone.

The last three variables in the third variable block measure interviewer experience made during the two panel waves considered (contact number and second year) and the total workload (total number contacts). They have rather small effects both in the regular and the refusal conversion fieldwork phases. While the contact performance slightly improves with each contact, the effect of panel experience is not consistent. Also a high total workload does not necessarily pay off.

As to the interviewer random effects, they are quite substantial in all models. We find a strong decrease of the interviewer cooperation performance random effect after the inclusion of the post-field variables during the regular phase. Probably a large portion of interviewer variance stems from the fieldwork time s/he is employed: interviewers working later are more likely to be contacting more difficult cases. Regarding refusal avoidance, the fieldwork progress and the number of contacts on a household have opposite effects on performance. Therefore, one cannot definitely say that fieldwork progress is positively correlated with a higher refusal rate of contacts. A correlation analysis confirms this: while the correlation coefficient between the number of days of fieldwork and the refusal of a contact amounts to a positive value of .09 (significant on 1%), the correlation with the cooperation index is a high negative value of -.34 (significant on 1%).

We try to further decrease the unexplained interviewer model variance by the inclusion of variables collected with the help of a paper and pencil interviewer questionnaire. This questionnaire contains, amongst other things, interviewer socio-demography and socioeconomy, job satisfaction, variables on attitudes towards trying to convince or persuade a sample member to participate (de Leeuw et al. 1998), job motivation (Sonnentag and Frese 2002), perceived burden and to what degree one is able to adapt to people or situations (Japec 2005). None of these variables proved significant in the (fieldwork variables) controlled models, neither during the regular nor during the refusal conversion fieldwork period. This finding reinforces previous results that interviewer main effects do not have an impact on their performance at convincing sample members to participate in surveys.

Application Example 1: Residual Analysis of Interviewer Performance

Similarly to the work in Pickery and Loosveldt (2004), we are able to identify exceptional interviewers in a residual analysis. For the survey agency this might be an appropriate tool to assess interviewer performance in an equitable way. For example, if it turns out that an interviewer performs badly before post-fieldwork quantities are controlled, and better *after* controlling for these, it can be concluded that his/her fieldwork assignment might have produced bad fieldwork results. For example, in the cooperation model during the regular fieldwork phase, we find the residual plots of interviewer performance depicted in figure 3 after controlling for the pre-field variables and in figure 4 after the inclusion of all variables:

The highlighted interviewer shows a relatively bad performance in figure 3. However, controlled for the fieldwork variables, the outlier problem almost vanishes. The reason for the highlighted interviewer to have performed so badly was his/her late fieldwork period with a difficult sample to be worked: while on average interviewers worked 43.3 (s.e.=.22) days after the fieldwork started, the interviewer concerned has a value of 143.4 (s.e.=6.0). Also the number of contacts on the households contacted by this interviewer is comparatively high. It is very likely that this interviewer joined the fieldwork staff quite late and had a high workload, and thus only obtained hard to convince households. Of course this special case is quite easy to detect and has only illustrative purposes. More sophisticated reasons might be responsible for a bad (or good) interviewer performance. The instrument described above can nevertheless help to find a reason for under/over performance using the different performance indices.

Application Example 2: Intermediate Contact Results

In this example, we consider intermediate contact results, i.e. vague or fixed appointments, as regards to the probability of completing a household, averaged by interviewer and contact result. We are interested in the question of whether interviewers achieving appointments X on a finally successfully administered household, are also successful with appointments Y. In addition, we would like to answer the question of whether interviewers who are successful with appointments X during the regular fieldwork phase, are also successful with appointments X during the refusal conversion. As above, X and Y may be fixed appointments with the target person, vague appointments with the target person, or appointments with another person in the household.

Here, we use the cooperation index. However, we do not model in a multilevel way but use the interviewer specific weighted[10] means of the household cooperation, distinguished by intermediate contact result (see Table 1 and Table 2). We calculate simple correlation coefficients between the mean household cooperation, averaged for each interviewers intermediate contact results. In addition, we depict the number of interviewers having obtained the corresponding contact results, and the significance level of the correlation coefficient (see table 4).

Interviewers who obtained a vague appointment during the regular phase, and 'whose' households finally cooperate,

[10] Similarly to the weights in the multiple membership multilevel models we use the inverse of the contact number on the household.

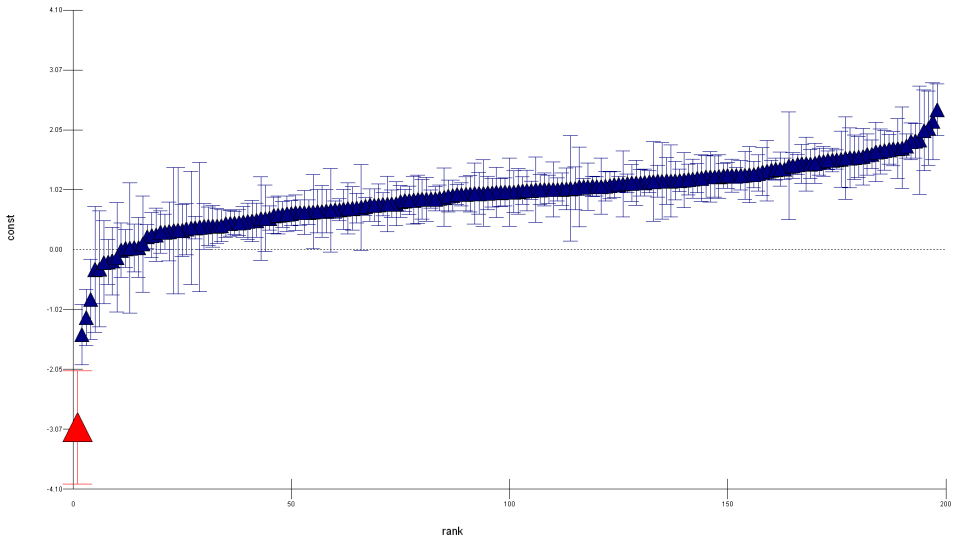

Figure 3. Residual of Interviewers in Cooperation Model, Regular Fieldwork Phase. Vertical Lines Standard Deviations. Negative Outlier Highlighted in Pre-field Model

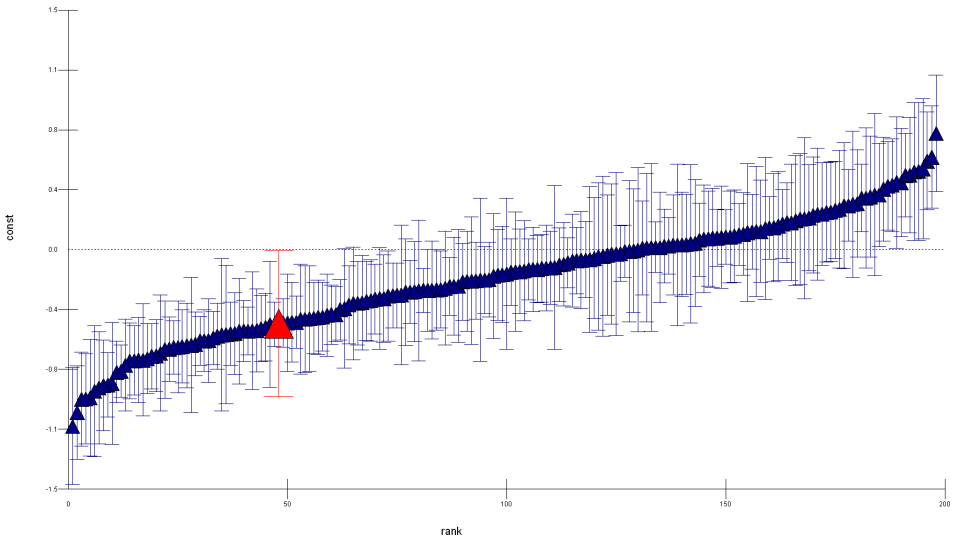

Figure 4. Residual of Interviewers in Cooperation Model, Regular Fieldwork Phase. Vertical Lines Standard Deviations. Negative Outlier Highlighted in Post-field Model

tend to also be successful with fixed appointments with the target person (corr = .28). There are no further significant correlations during the regular phase. To the contrary, there is even a negative, albeit insignificant, correlation between vague appointments and fixed appointments with household members other than the target person.
There are two correlations worth mentioning for the refusal conversion phase (when there are a much smaller number of interviewers): finally successful vague appointments positively correlate with finally successful fixed appointments made both with the target person and with other persons.

To summarise, interviewers who obtained a fixed or vague appointment on a finally cooperating household, are not necessarily also successful with other appointment types. In addition, final 'successes' on appointments work differently during the regular and the refusal conversion fieldwork phase.
In table 5 we depict the correlations which result after the same appointment type across the fieldwork phases: Regarding cooperation of households with an intermediate contact result across fieldwork phases, there are positive correlations, of which only one is significant (at the 6% level). Interviewers who are (un)successful with households after obtaining appointments with other persons during the regular phase, are also rather (un)successful after the same contact result during the refusal conversion phase.

Table 4: Correlations of Interviewer Specific Mean Household Cooperation after Appointments. Regular and Refusal Conversion Stage Separated

Correlation Coefficient (N Interviewers) Significance Level	Regular Fieldwork Phase (RE)			Refusal Conversion Phase (RF)		
Contact Result	FT	VT	FO	FT	VT	FO
Fixed Appointment with Target Person (FT)	1 200			1 64		
Vague Appointment with Target Person (VT)	.28 200 .01	1 202		.24 62 .06	1 67	
Fixed Appointment with Other Person (FO)	.09 45 .56	-.04 162 .61	1 163	.09 45 .56	.30 47 .04	1 48

Table 5: Correlations of Interviewer-Specific Mean Household Cooperation after Appointments, across Regular and Refusal Conversion Stages

		Regular Fieldwork Phase (RE)		
Phase	Correlation Coefficient N (Interviewers) Significance Level	FT	VT	FO
	Fixed Appointment with Target Person (FT)	.06 64 .64		
Refusal Conv. Phase	Vague Appointment with Target Person (VT)		.16 67 .19	
	Fixed Appointment with Other Person (FO)			.28 48 .06

Summary and Conclusion

The article has investigated the question of how to measure interviewer performance and effects as to sample case participation in CATI surveys, in which sample cases are not completely assigned to single interviewers, but where several interviewers work the same phone number, through a randomised allocation of contacts. Existing approaches mostly focus on the first contact, whose result implies the highest interviewer effect. This approach makes the investigation of the performance of interviewers impossible, who work later contacts on a sample case without a final disposition after the first contact. Others measure single contact outcomes.

Both approaches suffer from the problem of how to assign values to intermediate results (mostly appointments), i.e. contact results other than a completed interview or a refusal. What is it worth if the interviewer obtains, for example, a fixed appointment with an agreed date and time, rather than a vague appointment? How should one take the outcome of a possible previous contact into account? In addition, both existing approaches do not take into consideration the complex clustering of contacts within sample cases within interviewers, which might or might not be hierarchical. To model interviewer performance effects makes complex multilevel models necessary.

In this article we propose and model two interviewer performance measures for centralised CATI surveys, built on existing theories of cooperation in CAPI surveys, in which each sample case 'belongs' to one interviewer:

- the 'cooperation' index measures the binary outcome of the sample member; all contacts on a sample case after which the treated sample member *finally* cooperates are given a value of 1, and of 0 if s/he does not *finally* cooperate. The idea behind this measure is that it is not so much the individual contact outcome which is decisive, but that interviewers who work the sample case follow the common target 'cooperation' of the sample member.
- the 'refusal avoidance' index measure is derived from the well known theory elaborated by Groves and Couper (1996, 1998), with the strategy of maintaining interaction with the (reluctant) sample case and of avoiding refusals, rather than trying to push a sample case and to risk a final refusal. Binary success in this context is defined as 1 if the contact outcome is not a refusal.

In addition, we define and model a 'naive' interviewer performance measure, which is a direct conversion of the contact result into a real number:

- the 'optimisation' index: this measure directly as-

Table 6: Household Grid Completion, MCMC Estimates, Regular Fieldwork Phase. All Coefficients 'significant' (at least twice their Standard Error). All Interviewer Random Effects at least three Times their Standard Error.[a] In italics: at least 10x their s.e.. in Brackets: not significant Effects.

Regular Fieldwork Phase	Binary Result for Household (1=Completed): 'Cooperation' Multiple Membership Model			Binary Result for Contact (1=No Refusal): 'Refusal Avoid' Cross classified Multilevel			Cooperation Probability for HH (Binomial): 'Optimisation' Cross classified Multilevel		
Model	Only Intercept	+ Pre-Field	+ Post-Field	Only Intercept	+ Pre-Field	+ Post-Field	Only Intercept	+ Pre-Field	+ Post-Field
N (Contacts on Households)					39,194				
N (Interviewers)					202				
Intercept	*.297*	*1.041*	*2.418*	*2.679*	*2.910*	*3.174*	*2.676*	*2.941*	*2.075*
+ *Pre-Field Variables*									
Swiss German Part		(.059)	.410		.155	(-.091)		(.129)	(.015)
SHP I sample		Base	Base		Base	Base		Base	Base
SHP II sample		*-.809*	*-.701*		*-.607*	*-.638*		*-.617*	*-.714*
SILC I sample		*-1.091*	*-.710*		-.625	-.447		-.641	-.650
SILC II sample		*-1.731*	*-1.525*		*-1.267*	*-1.213*		*-1.277*	*-1.199*
Survey Year 2006		.087	-.166		.544	.387		.534	.611
+ *Post-Field Variables*									
Number of Day of Fieldwork			*-.014*			*-.023*			*-.018*
Contact Time of Day: 9 am 1 pm			Base			Base			Base
Contact Time of Day: 1 pm 5 pm			(-.055)			(-.111)			(-.017)
Contact Time of Day: 5 pm 10 pm			(.020)			-.308			(.056)
Household Contact Number									*-3.165*
Total Number of Contacts on Household			*-.065*			*.190*			*3.164*
Same Interviewer as in previous Contact			(.025)			(.146)			.210
Interviewer Contact number on Household			.001			.002			.003
Total Number of Contacts of Interviewer			(.000)			.001			-
Interviewer second year at SHP/SILC (only 2006)			.239			(-.025)			(-.090)
Previous Contact: none (fresh sample line)									Base
Previous Contact: fixed Appointment									*1.900*
Previous Contact: vague Appointment									(.047)
Previous Contact: Appointment by other Person									(.095)
Random Intercept *Sample member*	3.290[b]	3.290	3.290	3.290	3.290	3.290	3.290	3.290	3.290
Random Intercept *Interviewer*	2.251	1.433	.274	.509	.439	.375	.506	.443	.241
Deviance (MCMC) Statistic	45,362	43,526	39,750	18,586	18,145	16,349	18,597	18,146	11,235

[a] See Fielding and Goldstein (2006): " more than 3 times their standard errors. As such if it were desired to refer them to the appropriate test null distribution they would all be significantly different from zero beyond the 1% level." (p. 30).

[b] In logit models the variance at the lowest level can be interpreted as the area under the logistic curve ($\pi^2/3 \approx 3.29$); see Snijders and Bosker (1999).

Table 7: Household Grid Completion, MCMC Estimates, Refusal Conversion Fieldwork Phase. All Coefficients 'significant' (at least twice their Standard Error). In italics: at least 10x their s.e.. in Brackets: not significant Effects.

Refusal Conversion Fieldwork Phase	Binary Result for Household (1=Completed): 'Cooperation' Multiple Membership Model			Binary Result for Contact (1=No Refusal): 'Refusal Avoid' Cross classified Multilevel			Cooperation Probability for HH (Binomial): 'Optimisation' Cross classified Multilevel		
Model	Only Intercept	+ Pre-Field	+ Post-Field	Only Intercept	+ Pre-Field	+ Post-Field	Only Intercept	+ Pre-Field	+ Post-Field
N (Contacts on Households)					5,919				
N (Interviewers)					69				
Intercept	-.166	-.276	1.054	.593	.252	(.363)	*-1.163*	-.945	-.273
+ Pre-Field Variables									
Swiss German Part		(.199)	(.071)		.383	(.003)		(.171)	(.190)
SHP I sample		Base	Base		Base	Base		Base	Base
SHP II sample		-.512	-.490		-.183	-.302		-.546	-.409
SILC I sample		-.319	-.436		-.392	-.755		-.836	-.783
SILC II sample		-.395	-.571		-.655	-.892		-.815	-.745
Survey Year 2006		.339	(.215)		.484	.566		(.096)	(.311)
+ Post-Field Variables									
Number of Day of Fieldwork			-.012			*-.016*			-.005
Contact Time of Day: 9 am 1 pm			Base			Base			Base
Contact Time of Day: 1 pm 5 pm			(.175)			.922			(.210)
Contact Time of Day: 5 pm 10 pm			-.563			-.561			(-.113)
Household Contact Number									(.023)
Total Number of contacts on Household			.027			*.773*			-.105
Same Interviewer as in previous contact			.361			.492			.260
Interviewer Contact number on Household			.001			.002			(.000)
Total Number of Contacts of Interviewer			(.000)			-.001			-
Interviewer second year at SHP/SILC (only 2006)			(.049)			-.511			(-.355)
Previous Contact: none (fresh sample line)									Base
Previous Contact: fixed Appointment									*1.885*
Previous Contact: vague Appointment									(.096)
Previous Contact: Appointment by other Person									(-.022)
Previous Contact: (soft) Refusal									-.347
Random Intercept *Sample member*	3.290	3.290	3.290	3.290	3.290	3.290	3.290	3.290	3.290
Random Intercept *Interviewer*	.856	.515	.520	.524	.306	.306	.271	.263	.289
Deviance (MCMC) Statistic	7,496	7,440	7,266	7,293	7,239	5,306	6,386	6,304	5,816

sesses the contact result by calculating the rate of *finally* cooperating sample members, by contact result. Trivial contact results are cooperation (=1) and refusal (=0), but it can be shown by means of contact data that fixed appointments result in a higher mean number of finally cooperating sample members than vague appointments. The idea is that each interviewer tries to optimise the outcome result of the contact as to finally try to convince the sample member to participate.

In the empirical part of the paper, we model the three indices using data from two waves of two Swiss general panel surveys, distinguished by regular and refusal conversion fieldwork phases. Due to the complex clustering structure, we model the cooperation index using multiple membership multilevel models, and the refusal avoidance and the optimisation indices using cross-classified multilevel models. It first turns out that the interviewer effects during refusal conversion measured by the optimisation index are rather small. This is probably caused by defining the index as the contact results averaged over all sample cases. Therefore this index is not suitable to measure interviewer effects.

Second, we find for the two remaining indices that both fixed and random effects differ; while sample effects are comparable, fieldwork effects are sometimes quite different. However the effects are mostly consistent with the underlying theoretical concepts, e.g. 'maintaining interaction' or 'stepping back'. We are able to substantially reduce interviewer variance by adding fieldwork variables, especially in the models which use regular fieldwork data. Importantly, we show the importance of controlling for fieldwork time in order to assess interviewer performance. This is most important when analysing the cooperation performance index.

The different results obtained for the two indices call for a more sophisticated treatment of how interview performance and effects should be measured and modelled in centralised CATI surveys, possibly also considering special survey characteristics and performance targets. A tentative application of the indices considered might be tried here: The refusal avoidance performance measure could be used in surveys in which it is of crucial importance to have as many sample members as possible turned into respondents. Examples are panel surveys, whose long-term existence depends crucially on a low attrition of the sample members. The cooperation performance measure could be used in any other random sample survey, in which one important target is to maximise response rate, and where teamwork rather than single contact results are to be improved.

The proposed measures still need to be evaluated on other surveys. The next step could be to conduct experiments in which the measures are tested in varying survey specific conditions.

Acknowledgements

Support by the Swiss National Science Foundation is gratefully acknowledged. I thank two anonymous reviewers for comments on a previous version of this paper.

References

Browne, W. (2005). *MCMC Estimation in MLwiN. Version 2.0.* Centre for Multilevel Modelling, University of Bristol.

Campanelli, P., & O'Muircheartaigh, C. (1999). Interviewers, Interviewer Continuity, and Panel Survey Nonresponse. *Quality & Quantity*, *33*, 59-76.

De Leeuw, E., Hox, J., Snijkers, G., & De Heer, W. (1998, August). Interviewer Opinions, Attitudes and Strategies Regarding Survey Participation and Their Effect on Response. *ZUMA Nachrichten Spezial*, 239-248.

Durand, C. (2005). Measuring Interviewer Performance in Telephone Surveys. *Quality & Quantity*, *39*, 763-778.

Fielding, A., & Goldstein, H. (2006). *Cross-classified and Multiple Membership Structures in Multilevel Models: An Introduction and Review.* (Research Report RR791, department for education and skills, University of Birmingham)

Graf, E., & Tillmann, R. (2005). *Comparaison du déroulement des enquêtes PSMI, PSMII et SILC (2004-2005).* (Working Paper 4-05, Swiss Household Panel, Neuchâtel)

Groves, R., & Couper, M. (1996). Contact-Level Influences on Cooperation in Face-to-Face Surveys. *Journal of Official Statistics*, *12*(1), 63-83.

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

Hox, J., & de Leeuw, E. (2002). The Influence of Interviewers' Attitude and Behaviour on Household Survey Nonresponse: An international Comparison. In R. Groves, D. Dillman, J. Eltinge, & R. Little (Eds.), *Survey nonresponse.* New York: Wiley.

Hox, J., de Leeuw, E., & Kreft, I. (1991). The Effect of Interviewer and Respondent Characteristics on the Quality of Survey Data: A Multilevel Model. In Biemer, Groves, Lyberg, Mathiowetz, & Sudman (Eds.), *Measurement Errors in Surveys.* New York: Wiley.

Hox, J., De Leeuw, E., & Snijkers, G. (1998, August). Fighting Nonresponse in Telephone Interviews; Successful Interviewer Tactics. *ZUMA Nachrichten Spezial*, 173-185.

Japec, L. (2005). *Quality issues in interviewer surveys: some contributions.* (Department of Statistics, Stockholm University. PhD Thesis)

Lipps, O. (2007a). Attrition in the Swiss Household Panel. *methoden - daten - analysen*, *1*, 45-68.

Lipps, O. (2007b). *Cooperation in centralised CATI Household Panel Surveys - a Contact based multilevel Analysis to examine Interviewer, Respondent, and Fieldwork Process Effects.* (Mimeo, University of Neuchâtel)

Mayer, T., & O'Brien, E. (2001). *Interviewer refusal-aversion training to increase survey participation.* (American Statistical Association, Alexandria VA)

O'Muircheartaigh, C., & Campanelli, P. (1999). A Multilevel Exploration of the Role of Interviewers in Survey Non-Response. *Journal of the Royal Statistical Society: Series A*, *162*(3), 437-446.

Pickery, J., & Loosveldt, G. (2004). A Simultaneous Analysis of Interviewer Effects on Various Data Quality Indicators with Identification of Exceptional Interviewers. *Journal of Official Statistics*, *20*(1), 77-89.

Pickery, J., Loosveldt, G., & Carton, A. (2001). The Effects of Interviewer and Respondent Characteristics on Response Behaviour in Panel Surveys. *Sociological Methods & Research*, *29*(4), 509-523.

Rendtel, U., Behr, A., Bellgardt, E., Neukirch, T., Pyy-Martikainen,

M., Sisto, J., et al. (2004). *Report on Panel Effects.* (CHINTEX Working Paper 16, European Commission)

Snijders, T., & Bosker, R. (1999). *Multilevel analysis*. Newbury Park, California: Sage.

Snijkers, G., Hox, J., & de Leeuw, E. (1999). Interviewer's tactics for Fighting Survey Nonresponse. *Journal of Official Statistics*, *15*(2), 185-198.

Sonnentag, S., & Frese, M. (2002). Performance concepts and performance theory. In S. Sonnentag (Ed.), *Psychological management of individual performance.* Chichester: Wiley & Sons.