DE GRUYTER OPEN

Journal of Official Statistics, Vol. 33, No. 3, 2017, pp. 659–686, http://dx.doi.org/10.1515/JOS-2017-0031

Fieldwork Monitoring for the European Social Survey: An illustration with Belgium and the Czech Republic in Round 7

Caroline Vandenplas[1], *Geert Loosveldt*[2], *and Koen Beullens*[3]

Adaptive and responsive survey designs rely on monitoring indicators based on paradata. This process can better inform fieldwork management if the indicators are paired with a benchmark, which relies on empirical information collected in the first phase of the fieldwork or, for repeated or longitudinal surveys, in previous rounds or waves. We propose the "fieldwork power" (fieldwork production per time unit) as an indicator for monitoring, and we simulate this for the European Social Survey (ESS) Round 7 in Belgium and in the Czech Republic. We operationalize the fieldwork power as the weekly number of completed interviews and of contacts, the ratio of the number of completed interviews to the number of contact attempts and to the number of refusals. We use a repeated measurement multilevel model, with surveys in the previous rounds of the European Social Survey as the macro level and the weekly fieldwork power as repeated measurements to create benchmarks. We also monitor effort and data quality metrics. The results show how problems in the fieldwork evolution can be detected by monitoring the fieldwork power and by comparing it with the benchmarks. The analysis also proves helpful regarding post-survey fieldwork evaluation, and links effort, productivity, and data quality.

Key words: Adaptive/responsive design; productivity metric; effort metric; data quality metric.

1. Background

The implementation of adaptive (e.g., Wagner 2008; Schouten et al. 2012; Luiten and Schouten 2013) or responsive (Groves and Heeringa 2006) survey designs rely on the availability of data. Tailoring data collection is based on information, which can be categorized as information about the sampling units and information about the fieldwork process as a whole. The former concerns the sampling units or the remaining active cases and this information also splits into two categories: (1) auxiliary data from the sampling frame or matched administrative data, and (2) paradata from previous waves in longitudinal surveys or gathered during the first data collection phase. Adaptive designs (e.g., Peytchev

[1] Centre for Sociological Research, KU Leuven, Parkstraat 45 - box 3601, 3000 Leuven, Belgium. Email: Caroline.Vandenplas@kuleuven.be
[2] Centre for Sociological Research, KU Leuven, Parkstraat 45 - box 3601, 3000 Leuven, Belgium. Email: Geert.Loosveldt@kuleuven.be
[3] Centre for Sociological Research, KU Leuven, Parkstraat 45 - box 3601, 3000 Leuven, Belgium. Email: Koen.Beullens@kuleuven.be
Acknowledgment: We thank the associate editor and the two reviewers for their thorough review and helpful comments.

et al. 2010; Lundquist and Särndal 2013; Schouten and Shlomo 2015) tailor the survey protocols to (groups of) sampling units based on characteristics determined by the available information, typically via a response propensity model. The main aims of such designs are to increase response rates (for example by prioritizing high response propensity groups), to decrease nonresponse bias (for example by prioritizing low response propensity groups), to reduce survey costs (for example by limiting the number of calls), or any combination of the previous, sometimes competing, goals. The decision on which of these goals should be the focus of the adaptive design should be taken in consultation with survey sponsors and the fieldwork agency, taking into consideration the main aims of the survey, the available budget, and balancing the different survey error components, such as sampling error and nonresponse bias or nonresponse bias and measurement error.

The second type concerns information about the fieldwork (or data collection) process as a whole. In this case, the survey manager assesses the progress of the fieldwork through different indicators, usually based on "live" paradata. These indicators are sometimes referred to as Key Performance Indicators (KPIs). Jans et al. (2013) define different types of KPIs: Effort metrics, the status of active cases, productivity metrics, the balance and representativeness of data, survey output, and measurement process quality. Moreover, Groves and Heeringa (2006) talk about "phase capacity" as an indicator for the need to change the design. A fieldwork phase with a specific design has reached its maximum capacity when it does not bring new cases that change the current estimates of key survey statistics. Fieldwork monitoring (Laflamme 2009; Malter 2013) refers to following up the evolution of the chosen KPIs during the fieldwork. Although the ultimate aims remain to increase response rates, decrease nonresponse bias, and reduce survey costs, the focus of monitoring fieldwork performance indicators is shifted to increasing the process performance (Kirgis and Lepkowski 2013), process efficiency (Zhang et al. 2013), or survey performance (Jans et al. 2013), where efficiency and performance can have different definitions but usually reflect a quality-cost or productivity-effort tradeoff.

However, managing fieldwork and the concept of responsive design go beyond monitoring key indicators; monitoring should lead to interventions that increase the process and data quality given a fix budget. Laflamme and colleagues (2008) identify six steps to *active management*: Planning, monitoring, identifying problems and finding solutions, taking appropriate corrective actions, communicating, and evaluating and documenting. In this article we will focus on monitoring and identifying problems. So far, we have introduced the idea of KPIs to be monitored. To identify problems, some inspiration can be taken from statistical process control and statistical quality control tools (Kreuter et al. 2010; Jans et al. 2013), which have been applied in recent decades to regulate manufacturing processes (Shewhart 1931, 1939; Juran and Gryna 1988). Shewhart proposes the use of time-series control charts to allow the distinction between the signal (real problem) and the noise (natural variability in the process). To do this, limits have to be defined for the indicators being monitored. Jans and colleagues (2013) distinguish *control limits* and *specification limits.* The former are based on data from the process itself (paradata). A typical example is monitoring interview duration (Jans et al. 2013). The length of each interview by one interviewer can be compared with the average interview duration and its standard deviation. These values are calculated based on previous rounds or waves, or over the first phase of the fieldwork. On the one hand, a single

divergence from the calculated mean larger than three standard deviations would be considered as problematic. On the other hand, two sequential divergences larger than two standard deviations or three sequential divergences larger than one standard deviation (Jans et al. 2013, 202–204) would equivalently signal a failure in the interview process. Such monitoring rules comprise both static (single divergence) and dynamic (sequential divergences) components. By contrast, specification limits originate from technical issues or from the judgement of a stakeholder. They are not based on empirical information about the process.

In any case, regardless of the KPIs or the charts used to monitor them, defining limits and planning the rules for process failure detection need to be based on information that may not be available at the start of the fieldwork. Groves and Heeringa (2006) accordingly define a "first phase," when the evolution of the fieldwork based on one or more designs can be observed. Although in practice, interventions may already occur during this first phase, the original idea was to adapt the design in a second phase. Depending on the survey, the fieldwork duration may be relatively short, and having to first learn from the process will involve a loss of time. Schouten et al. (2013, 54) note that "adaptive survey designs lend themselves best to settings where surveys are run repeatedly or for a longer period . . . The designs also lend themselves to survey institutes that conduct many surveys that are relatively similar in topics and budget." In turn, Laflamme (2009, 3) states that one of his research objectives is "to learn about the data collection survey process within and *across* surveys." Repeated cross-national surveys such as the European Social Survey (ESS) collect paradata about the fieldwork process across rounds and participating countries, which can be used to develop control limits for fieldwork KPIs in future rounds or waves. They are therefore suited for the development of such control limits, even though the lack of centralized fieldwork supervision may hinder the practical implementation of monitoring.

2. Research Objectives

In this article, we concentrate on performance indicators, their evolution, and how problems in their development can be detected through "flagging" rules. Moreover, we suggest possible interventions when problems in the fieldwork evolution are detected.

In the next section, we introduce the concept of the *fieldwork power*, defined as the productivity of the fieldwork per unit of time, and explain how the evolution of the fieldwork power can be modeled using past similar surveys in order to create a benchmark for fieldwork monitoring. Such a benchmark and its confidence bands can act in a similar way as control limits can, although instead of only detecting deviations from an average in certain acceptable bounds, we also seek to detect deviating patterns from the benchmark in terms of the evolution. Another advantage of the modeled benchmark is that it evolves over time; the expected values of the fieldwork power depend on the fieldwork week, whereas control limits are usually static. The benchmark is not a fixed value with boundaries through the fieldwork period, but evolves over time to reflect the dynamic of the fieldwork. For example, the weekly number of completed interviews cannot be expected to be constant, but will show a decreasing trend as the fieldwork progresses and some cases are classified. Moreover, the fieldwork power can detect both deviations in a

specific week and deviations from a pattern. This can less easily be accomplished when monitoring cumulative response rates – which is common practice – because information in cumulative response rates is aggregated.

In the following section, we propose three types of benchmarks that can be constructed for the countries participating in the ESS Round 7. The first benchmark is based on the six first rounds of the ESS, and all the participating countries in these rounds. We define an ESS survey as being a combination of a round and a participating country in that round. On the one hand, this benchmark has the advantage of including a large number of surveys sharing some key survey characteristics – including the mode of data collection, topic, or contact procedures – helping us to develop a good model for the benchmark. On the other hand, the ESS surveys can also have divergent characteristics – including sampling design or survey climate – that may influence the evolution of the fieldwork power. Hence, this benchmark would generally only be used if little information about the "to be monitored" survey characteristics was available. Nevertheless, constructing this benchmark is useful, as it allows us to gain some insight into the general shape of the fieldwork power evolution. As the model includes many surveys, it can be used to validate the hypothesized general shape (see Section 4).

The second benchmark is built based on ESS surveys that have (more) design characteristics that are similar to the "to be monitored" survey. In practice, the fieldwork capacity depends on survey designs that can differ from one ESS survey to the other: Individual frame or not, survey climate, and refusal conversion percentage, among others. We therefore model the evolution of the fieldwork power for surveys having the same type of frame, a similar expected response rate, and a similar planned refusal conversion rate. This will result in a less powerful model due to the reduced number of surveys considered, but will be more realistic as the surveys will share a large number of characteristics.

The third benchmark is constructed based on previous rounds in a specific country, reducing the fit of the model but increasing the resemblance to the survey to be monitored.

The benchmarks are increasingly refined (the model being evaluated on ESS surveys that increasingly resemble the survey of interest) but decreasingly precise (the number of surveys entering the model reducing at each step).

In the next section, we propose a number of "flagging" rules that should trigger actions: Finding the reasons for the deviating pattern, intervening by changing the protocol, or increasing the effort.

Lastly, we simulate the monitoring of the fieldwork power using the three described benchmarks for Belgium and the Czech Republic in the ESS Round 7. In combination with the fieldwork power – which can be seen as a *productivity metric* – we examine the mean interviewer productivity, and monitor the number of contact attempts and the number of active interviewers (an interviewer having performed at least one contact attempt during the week examined), which can be seen as *effort metrics*. We also monitor *data quality metrics* to link the process to data quality (Kreuter et al. 2010). Based on the survey data, we first calculate the cumulative estimated mean age and alcohol consumption during weekdays as reported by the respondents during the interview, and their sampling errors. This is in line with the idea of phase capacity as suggested by Groves and Heeringa (2006). We therefore observe when the estimates become stable and when sampling errors fall beyond a certain pre-set value. Many variables from the ESS questionnaire could

eventually be chosen. The selection should contain as many variables as relevant but an excess would render the monitoring clutter. In our case, we restricted the choice to two variables to stay succinct: age and alcohol consumption, which are known to suffer from nonresponse bias. Moreover, alcohol consumption could also suffer from some measurement bias due to, among other things, social desirability, whereas age should be more accurate. Finally, age is a variable that can be found in almost all social surveys whilst alcohol consumption was specific for ESS Round 7.

Second, we calculate the cumulative percentage of women among people living with a partner, which should be 50% following the idea of "internal criteria of representativeness" (Kohler 2007). Indeed, assuming partnerships comprise a man and a woman, there should be as many women as men in the subsample of people living with a partner.

3. Fieldwork Power as an Indicator for Monitoring

The fieldwork yields different "products" that can vary depending on the data collection mode: including completed interviews or returned questionnaires, refusals, contacts, ineligible cases, response rate, and so on. Mimicking a principle of classical mechanics that states that the production or "work" (e.g., the total number of completed interviews) of a machine is its power (e.g., the number of completed interviews per time unit) multiplied by its running time:

$$Work = Power * Time \tag{1}$$

or elaborating this equation for time-varying power:

$$Work = \int_{t=0}^{t=end} Power(t)dt \tag{2}$$

we study the fieldwork power and the production of the fieldwork per time unit (Vandenplas et al. 2015; Vandenplas and Loosveldt 2017). Concentrating our attention on face-to-face surveys, we specify the fieldwork power as the weekly number of completed interviews, the weekly number of contacts, the weekly ratio of the number of completed interviews to the number of contact attempts (*efficiency*), and the weekly ratio of the number of completed interviews to the number of refusals (*performance*). We chose the time units as weeks for two main reasons: Allowing the processing of paradata generated during the fieldwork, and avoiding differences in fieldwork production during weekdays and weekends that are a natural variation of the process (noise). The time unit can, however, differ depending on the survey design, especially the mode of data collection. Telephone surveys, for instance, might use days or shifts as time units.

For the first two specifications of the fieldwork power, the "work" can be interpreted as the total number of completed interviews and the total number of contacts established at the end of the fieldwork; the sum of the power over all the fieldwork weeks. It should be noted that the same individual or household can be contacted more than once. The interpretation of the "work" in the two last cases is not straightforward.

Moreover, to compare different surveys and their respective fieldwork, we need to standardize the "input" (i.e., the sample size) to 100 units. From here onward, we consider the standardized number of completed interviews and the standardized number of contacts

each week, dividing the original values by the sample size and multiplying them by 100. The standardization of the sample size does not affect the efficiency or the performance, because they are ratios.

These four specifications of the fieldwork power can be seen as KPIs that need to be monitored. Efficient monitoring, however, requires a benchmark or a reference against which the KPIs can be tested. We therefore propose to model the fieldwork power evolution of surveys that are similar in terms of data collection mode, sampling frame and design, budget, response-enhancement strategies, target population, and so on, and to use the parameters estimated in the model to create a fieldwork power evolution curve that can be used as a benchmark. As explained, we propose three possible benchmarks; the best fitting one depends on the information available before the start of the fieldwork. A new country in the survey may have to be compared with the general ESS curve, whereas a country with many previous rounds might better be compared with its "previous rounds" benchmark. Nonetheless, a possible change in survey design (e.g., the frame) may imply that the "similar ESS survey" curve is better suited.

4. Modeling the Fieldwork Power to Create a Benchmark: The Example of the ESS

4.1. Data

The ESS is an academically driven, cross-national survey that measures attitudes, beliefs, and behavior patterns across Europe. The survey has been repeated every two years since its first round in 2002. Data are collected through standardized computer-assisted or paper-assisted personal interviews lasting about 60 minutes. In addition to its goal to reflect changes in Europe's social culture, the ESS aims to achieve and disseminate rigorous methodological standards for cross-national research in terms of sampling, translation, questionnaire design, and so on. Among other matters, participating countries have to present a sampling design that ensures a representative random sample is selected to participate in the survey. Data about contact attempts is also collected by the interviewers and recorded in what are termed contact forms. Information about each contact attempt is recorded: Date, time, day of the week, outcome, mode, interviewer, and so on. These contact forms are a rich source of paradata that permit us to calculate – for each round and each participating country – the fieldwork duration in weeks, the sample size, the weekly standardized number of completed interviews, the weekly standardized number of contacts, the weekly ratio of the number of completed interviews to the number of contact attempts (efficiency), and the weekly ratio of the number of completed interviews to the number of refusals (performance).

In order to create the fieldwork power evolution curves to be used as benchmarks, we need to develop a model. Therefore, we consider each ESS survey, round-country combination, and their associated fieldwork as a unit for which we measure the weekly fieldwork power. In this way, we can construct a dataset with a nested structure, in which for each survey and each week we have four measurements; one for each specification of the fieldwork power. Even though adding up the weekly standardized number of interviews over the duration of the fieldwork results in the response rate when disregarding

ineligibles, this way of thinking of the fieldwork is different to comparing weekly response rates. To compare weekly response rates across surveys, we either have to consider cumulative response rates, which differ from the power (that is one measurement in time), or to compare the number of completed interviews to the number of worked cases per time unit, which are more difficult to define and different to the weekly standardized number of interviews.

4.2. The Basic Model

Given the nested nature of our dataset, we use a repeated measurement multilevel model with the fieldwork power measurement nested in the ESS surveys. Based on the shape of the fieldwork power in Round 6 of the ESS (Vandenplas and Loosveldt 2017), we propose a cubic shape for the fieldwork power, having four important characteristics (see Figure 1): The fieldwork power in the first week, the rate of increase/decrease of power at the start, the rate of acceleration of the increase/decrease (how fast the decrease becomes larger or the increase becomes smaller), and the start of the "tail" when the fieldwork power levels off.

To model the evolution of the fieldwork power separately for each specification with four characteristics, we need at least a cubic expression. The model studied is a multilevel model with repeated measurements (the power in each week of the fieldwork) as follows:

$$P(s,w) = \beta_0(s) + \beta_1(s)w + \beta_2(s)w^2 + \beta_3 w^3 + \varepsilon_{s,w}, \qquad \text{(Model 1)}$$

$$\beta_0(s) = \gamma_{00} + u_{0s},$$

$$\beta_1(s) = \gamma_{10} + u_{1s},$$

$$\beta_2(s) = \gamma_{20} + u_{2s},$$

$$\beta_3 = \gamma_{30},$$

where w represents the weeks of the fieldwork ($0 =$ week 1), s stands for the surveys that we consider, and the residual term is normally distributed $\varepsilon_{s,w} \sim N(0, \sigma^2)$. The surveys s

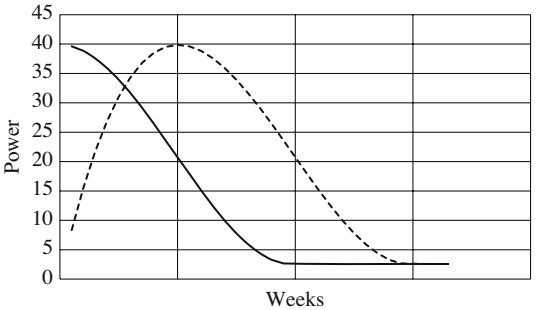

Fig. 1. The shapes of the fieldwork power evolution.

are the units at the macro level. For the model to be meaningful, the surveys should have enough similarities in terms of their fieldwork design: Data collection mode, contact procedures, and so on. The parameters γ_{00}, γ_{10}, γ_{20}, and γ_{30} are the overall fixed effects. The parameters u_{0s}, u_{1s}, and u_{2s} are the survey-specific intercept, linear, and quadratic parameters that express the survey-specific deviations from the overall evolution. The parameter β_3 is necessary to model the tail of the curve, but there is no substantive argument to specify survey-specific tails. Moreover, adding a random cubic effect is almost never significant and, for some models, does not lead to convergence. Therefore, we do not specify a survey-specific (random) cubic parameter. The level 2 covariance matrix for the random intercept (u_{0s}) and the random slopes (u_{1s}, u_{2s}) is given by

$$\begin{bmatrix} \sigma_0^2 & \sigma_{0,1} & \sigma_{0,2} \\ \sigma_{0,1} & \sigma_1^2 & \sigma_{1,2} \\ \sigma_{0,2} & \sigma_{1,2} & \sigma_2^2 \end{bmatrix}$$

Based on this model, we create three different benchmark curves. The benchmarks are as defined in the previous section, and the curves are used for the simulation of monitoring the ESS Round 7 for Belgium and the Czech Republic. Although the application would be interesting for all the 21 participating countries in Round 7, the two counties used serve as illustrations. We wanted to use two countries that display different survey characteristics: Individual/non-individual sampling frame, differences in response rate (57% versus 67% in Round 7), and different percentages of refusal conversions in previous rounds. Moreover, it was possible to construct the three types of benchmarks for these two countries, as they both participated in multiple previous rounds. Lastly, at least for Belgium, we have some insights into how the fieldwork agency organized the data collection process. It is important to note that for these two countries we simulate the monitoring using the contact data after the fieldwork had been completed. The feasibility of "live" monitoring still has to be determined.

4.3. Benchmark Curves

To create a *first benchmark* – the "ESS curve" – we applied Model 1 to all the surveys in the first six rounds of the ESS (Vandenplas and Loosveldt 2017), obtained by considering combinations of rounds and countries participating in each round for a total of 149 surveys. The surveys in the ESS have enough similarities to allow modeling them together: Data are collected through face-to-face interviews, and participating countries have to follow the ESS specifications in terms of sampling, translation, use of contact forms, data deposits, and so on (ESS 2011). This first benchmark also allows us to validate our cubic model, given the large number of surveys and measurements that are entered in the estimation of the model.

Even if we consider the ESS surveys to be similar enough to be modeled together, there are also many differences among their designs: Not all the surveys could use an individual sampling frame and hence sampling designs may differ, the degree of compliance to the contact procedure guidelines (Matsuo and Loosveldt 2013) also vary, the use of incentives

is not standardized across countries and rounds, and so on. Therefore, we propose a second, more refined, benchmark curve based on a limited number of ESS surveys that are more similar in terms of design to the survey that we intend to use to monitor the fieldwork.

In contrast with the first benchmark, this *second benchmark* "similar surveys" curve is tailored to the type of survey that we wish to monitor. Many characteristics could be taken into account to define what a similar survey is. Based on the results of Vandenplas and Loosveldt (2017), we separate countries with an individual sampling frame and we take into consideration the percentage of initial refusers re-approached for refusal conversion by a different interviewer. Both characteristics have been found to influence the shape of the fieldwork evolution in global analysis. Moreover, we also consider the main outcome of the fieldwork – the response rate – as being a differentiating characteristic among ESS surveys. As the sampling design needs to be approved before the start of the fieldwork, information on the type of frame is available before the fieldwork. The percentage of re-approached refusals and the final response rates are obviously not known before the start of the fieldwork; we therefore base our classification on the results of previous rounds.

A *third benchmark* curve is constructed based on "previous rounds in the same country". Different countries can have different cultures and attitudes toward surveys that could influence the survey climate, and hence, the implementation and the evolution of the fieldwork. Measurements for the survey climate are difficult to develop, therefore it may be interesting to focus on previous rounds in the same countries for the benchmark, when this is possible. We elaborate further on these possible benchmarks in Section 6.

5. Flagging Rules and Fieldwork Management

We suggest the following rules for detecting problems in the fieldwork progress. Immediate action should be taken if the fieldwork power (any of the four specifications):

- is below the confidence band of the benchmark in two subsequent weeks,
- is below the benchmark for three weeks in a row,
- or, reduces for three weeks in a row – is smaller than the previous weeks twice in a row.

The first action is focused on finding out the reason for the low power. Several relevant questions can be asked: "Are there fewer completed interviews than expected because of reduced effort or because of more reluctance of the sample?", "Is the power lower because the yield was very high in the first week?", "Are there fewer contacts because the fieldwork period started in a holiday period or were the contact attempts badly timed?", "Is the efficiency lower because of noncontacts, refusals, or other reasons?", "Have a lot of appointments been made?", "Is the performance poor because the active interviewers are less experienced?" Once the cause(s) for the low power have been determined, possible solutions can be sought, such as sending more interviewers into the field, providing them with guidelines about when to attempt contact based on the data available about respondents, re-training some interviewers in refusal conversion, or increasing incentives.

It is important to note that our focus is on increasing response rates. Nevertheless, if the power is repeatedly higher than a benchmark, this should also trigger some concerns,

unless it is the result of a conscious survey design feature (higher budget for a specific round, for example). Indeed, a power that is higher than expected could be a sign of falsification or a less strict application of quality rules.

With regard to the data quality metrics, we apply the following rule: The ESS specifies that the responding sample size should be equivalent to an effective sample of 1,500 people. This corresponds to a sampling error for any continuous variable of $SE = \sigma/\sqrt{1500}$. As already argued in the discussion of the research questions (Section 2), age and alcohol consumption are considered as relevant variables in this context. The standard deviation for age can be estimated by the standard deviation of age in the previous round or other data sources (e.g., census data). The standard deviation for alcohol consumption should be estimated after two weeks, based on the sample that responded up to that time. We then consider that a fieldwork design phase has attained its full capacity if the sampling error is lower than the pre-set value $SE_{pre} = \sigma/\sqrt{1500}$, for which σ is calculated as explained above (for two weeks in a row), and if the absolute difference in the estimate of a week from that of the previous one is lower than SE_{pre} for two weeks in a row.

6. Simulation of Fieldwork Monitoring in Round 7

To illustrate the use of the modeled fieldwork power evolution curve as a benchmark for fieldwork monitoring, we apply the principle to Round 7 of the ESS. This round is not included in the model to develop the ESS benchmark curve. We use the second release of the contact data (ESS 2014) containing 21 countries, and focus on two countries: Belgium and the Czech Republic.

We are well aware that using post-survey contact data is very different to using incoming information every week "live" during the fieldwork. In our case, the data has already been processed and there are no (or few) missing cases. During the fieldwork, data may be messier and some cases may not be reported in a timely manner by interviewers. This method, however, shows the importance of regular updates from the field.

These examples provide a good illustration of the potential and possible applications of the fieldwork power. We examine deviations from the benchmarks for the fieldwork in Round 7 in Belgium and in the Czech Republic. The benchmarks are the ESS global curve, the "similar surveys" curve, and the "previous rounds" curve. This not only allows us to illustrate how the fieldwork could have been monitored, but also to evaluate the fieldwork in Round 7 in those two countries.

6.1. The ESS Benchmark Curve

The estimated fixed parameters of Model 1 are shown in Table 1, and support the proposed shape. The number of fieldwork weeks considered are limited to 32, first because few surveys exceeded this fieldwork duration – only Ireland in Round 6 (46 weeks) and the Netherlands in Round 4 (39 weeks) – and second to allow the model to converge. Moreover, for the four specifications, the model fit improves for each step of the model building, including adding the cubic term (see Appendix online at: http://dx.doi.org/10.1515/JOS-2017-0031).

The intercepts in Table 1 show us that, over all the ESS surveys, in the first week, 4.80 interviews are completed out of 100 cases, 13.70 contacts are established, 22% of the

Table 1. Fixed effects for the multilevel model describing the shape of the fieldwork power evolution for all ESS countries (Model 1).

Effect	Completed interviews	Contacts	Efficiency	Performance
γ_{00} – intercept	$4.80\ (0.35)^{***}$	$13.70\ (0.79)^{***}$	$.223\ (.012)^{***}$	$2.59\ (0.17)^{***}$
γ_{10} – linear term	$0.23\ (0.07)^{**}$	$0.40\ (0.18)^{*}$	$.016\ (.003)^{***}$	$0.18\ (0.08)^{*}$
γ_{20} – quadratic term	$-0.05\ (0.01)^{***}$	$-0.12\ (0.01)^{***}$	$-.002\ (.000)^{***}$	$-0.03\ (0.01)^{***}$
γ_{30} – cubic term	$1^{E\text{-}3}\ (1^{E\text{-}4})^{***}$	$3^{E\text{-}3}\ (3^{E\text{-}4})^{***}$	$1^{E\text{-}3}\ (6^{E\text{-}6})^{***}$	$1^{E\text{-}3}\ (2^{E\text{-}4})^{***}$

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

contact attempts result in a completed interview, and 2.59 more interviews are completed than there are refusals. The positive sign of the linear term means that overall the fieldwork power (all specifications) increases at the start of the fieldwork, whereas the negative sign of the quadratic terms points to a levelling-off of that increase after a few weeks. Further, the parameters in Table 1 allow us to create an ESS curve that can be used as a benchmark for countries participating in further rounds of the survey. The following four expressions can be derived for the weekly standardized number of completed interviews, the weekly standardized number of contacts, the efficiency, and the performance:

$$\textit{Completed interviews}\,(w) = 4.80 + 0.23\,w - 0.05\,w^2 + 0.001\,w^3$$

$$\textit{Contacts}\,(w) = 13.70 + 0.40\,w - 0.12\,w^2 + 0.003\,w^3$$

$$\textit{Efficiency}\,(w) = 0.223 + 0.016\,w - 0.002\,w^2 + 0.001\,w^3$$

$$\textit{Performance}\,(w) = 2.59 + 0.18\,w - 0.03\,w^2 + 0.001\,w^3$$

6.2. Belgium

In the first step, we construct the benchmarks for the evolution of the fieldwork power based on (a subset of) surveys from the first six ESS rounds. In the second step, we present graphs that are aimed at monitoring the evolution of the fieldwork power. The graphs also display the metrics for "effort" (the number of contact attempts and the number of active interviewers) as well as survey data quality metrics. Simultaneously monitoring effort, productivity, and data quality metrics puts the three dimensions of process and data quality into perspective.

6.2.1. Benchmark Curves for Belgium

The first benchmark curve, the "ESS curve" was already constructed in the previous section (Model 1). To evaluate the second benchmark curve, we have to define which ESS surveys in previous rounds have a similar design to that of Belgium in Round 7. We therefore use the design characteristics of Belgium in that round.

In Round 7, the sampling design for Belgium was a two-stage cluster sample design. In the first step, municipalities were selected, and in the second step, individuals within each

Table 2. Fixed effects for the multilevel model describing the shape of the fieldwork power evolution for ESS surveys similar to Belgium round 7 (Model 1).

Effect	Compl. interviews	Contacts	Efficiency	Performance
γ_{00} – intercept	4.07 (0.91)***	16.13 (2.48)***	.154 (0.025)**	1.99 (0.35)***
γ_{10} – linear term	1.27 (0.28)***	3.28 (0.80)***	.027 (0.008)**	0.20 (0.11)
γ_{20} – quadratic term	−0.22 (0.03)***	−0.63 (0.07)***	−.003 (0.001)***	−0.03 (0.01)*
γ_{30} – cubic term	0.008 (0.001)***	0.02 (0.003)***	1^{E-4} (3^{E-5})***	1^{E-3} (5^{E-4})*

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{****}p < 0.001$.

selected municipality. The response rate was 57.0%, and varied between 53.4% and 61.2% in previous rounds. The percentages of initial refusals that were re-approached by a different interviewer in Rounds 2 to 6 vary between 12.1% and 20.1% (Round 1 was 0.8%). The fieldwork lasted for 17 weeks in Round 3 and 6, and up to 30 weeks in Round 5. The second benchmark curve for the simulation of fieldwork monitoring of Belgium is based on similar ESS surveys, defined as surveys in previous rounds with an individual sampling frame, response rates above 55%, and refusal conversion above 10%. These surveys are: Belgium, Rounds 2, 3, 4, and 6; Estonia, Rounds 2, 3, 4, 5, and 6; Finland, Rounds 4, 5, and 6; Hungary, Rounds 4, 5, and 6; Norway, Rounds 3, 4, and 5; Poland, Rounds 4 and 5; Sweden, Rounds 3 and 4; and Slovakia, Rounds 2 and 4. Table 2 displays the fixed parameters for the multilevel model (Model 1) with repeated measurements limited to these surveys. For reasons of convergence, we limit the analysis to 19 weeks, excluding the last weeks of fieldwork for: Belgium, Round 2 (20 weeks); Finland, Rounds 4 and 6 (22 weeks and 27 weeks); Norway, Rounds 4 and 5 (23 and 24 weeks); and Sweden, Rounds 3 and 4 (both 20 weeks). All these surveys had reached a 55% response rate in their nineteenth week of fieldwork. It should be noted that we apply the same cubic model as for the ESS benchmark curve. Our aim here is not to find the best fitting model but the best fitting cubic curve to the data. Nevertheless, the model fit statistics (see Appendix online at: http://dx.doi.org/10.1515/JOS-2017-0031) show that the cubic model is still the best fitting one for the standardized number of completed interviews and contacts.

Comparing the fixed parameter estimate of this benchmark with the overall ESS benchmark shows that the "similar surveys" have a slower start with a lower standardized number of completed interviews (even though there are more contacts), a lower efficiency, and lower performance. The linear term, however, describes a large initial increase of power, whilst the quadratic term shows a faster leveling-off of this initial increase, leading quickly to a decrease of power. We can hence conclude that for similar surveys to the Belgium ESS Round 7, in general, the first part (before the start of the tail) of the fieldwork lasts for a shorter period than the overall ESS mean but has a higher weekly yield.

The benchmark curves for the "similar surveys" can be derived for the weekly standardized number of completed interviews, the weekly standardized number of contacts, the efficiency, and the performance from the fixed effects in Table 2.

Table 3 displays the fixed parameters for the third benchmark based on the previous rounds in Belgium. The analysis is limited to 15 weeks for convergence reasons.

From the fixed parameters in Table 3, we can expect that the fieldwork power in the first week is lower than expected from the ESS curve or the similar survey curve. Only the

Table 3. Fixed effects for the multilevel model describing the shape of the fieldwork power evolution for previous rounds in Belgium (Model 1).

Effect	Completed interviews	Contacts	Efficiency	Performance
γ_{00} – intercept	2.89 (1.14)	12.00 (4.02)*	.173 (0.021)**	1.48 (0.30)**
γ_{10} – linear term	1.19 (0.47)	2.45 (2.45)	.021 (.012)	0.33 (0.25)
γ_{20} – quadratic term	− 0.19 (0.05)*	− 0.41 (0.17)	− .004 (.002)	− 0.07 (0.04)
γ_{30} – cubic term	0.007 (0.002)*	0.015 (0.007)	2^{E-4} (1^{E-5})*	4^{E-3} (2^{E-3})*

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{****}p < 0.001$.

value for efficiency lies between the benchmark curve and the similar survey curve. The estimated linear term for all power specifications lies between the estimated linear term of the ESS curve and the similar survey curve, showing that the improvement in power at the start of the fieldwork in the previous rounds is better than the overall ESS curve, but not as good as the similar survey one. Lastly, the quadratic terms show that the leveling-off appears faster than the leveling-off of the ESS curve, and than the similar surveys for the efficiency and the performance.

The covariance parameters (results not shown) are a lot smaller, showing that the fieldwork evolution in the different rounds in Belgium are more closely related than when considering all ESS surveys. However, the significance of fixed and random parameters is lower, because only six surveys were involved and hence a limited number of fieldwork power measurements are available. The benchmark curves for the "previous round" can be derived for the weekly standardized number of completed interviews, the weekly standardized number of contacts, the efficiency, and the performance from the fixed effects in Table 3.

6.2.2. Fieldwork Power: Standardized Number of Completed Interviews and Contacts

Figures 2a and 2b display the three benchmark curves for ESS Belgium – the ESS curve (short dashed line), similar surveys (long dashed line), and previous rounds (solid line) – with their 95% confidence band as well as the achieved weekly standardized number of completed interviews (Figure 2a) and contacts (Figure 2b) (black dots) in Round 7. In Figure 2c, the solid line represents the weekly standardized number of contact attempts divided by ten and the dashed line represents the weekly standardized number of active interviewers. The diamonds represent the mean number of contact attempts per interviewer.

In line with the flagging rules from Section 6, week 2 should (possibly) have been flagged; the standardized number of completed interviews and of contacts are under the benchmark bands, even though in week 2 the power enters the "previous rounds" benchmark band, it stays below the expected yield. Looking at "effort" (Figure 2c), we see that in week 1, the number of active interviewers, the number of contact attempts, and the mean number of contact attempts per interviewer is very low. In week 2, the number of active interviewers increased only slightly, but the active interviewers performed more contact attempts (twice as many). This should have led, in our opinion, to the fieldwork agency contacting their inactive interviewers to ask them to start working on their cases as soon as possible, with the aim of avoiding the fieldwork period being dragged out.

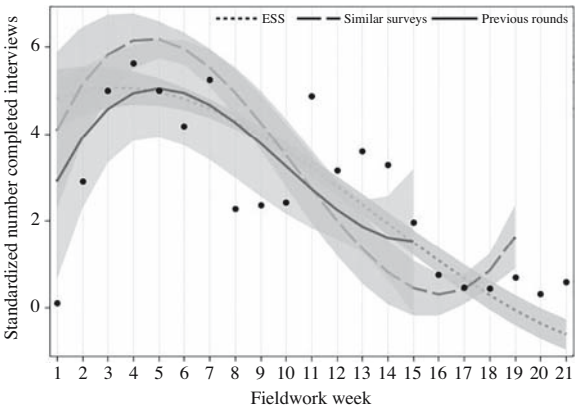

Fig. 2a. Monitoring the weekly number of completed interviews – BE R7.

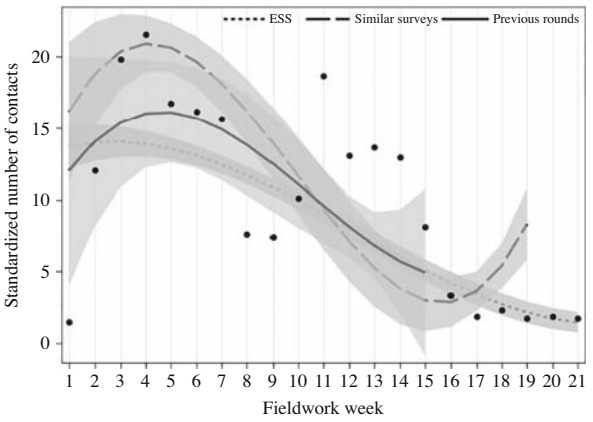

Fig. 2b. Monitoring the weekly number of contacts – BE R7.

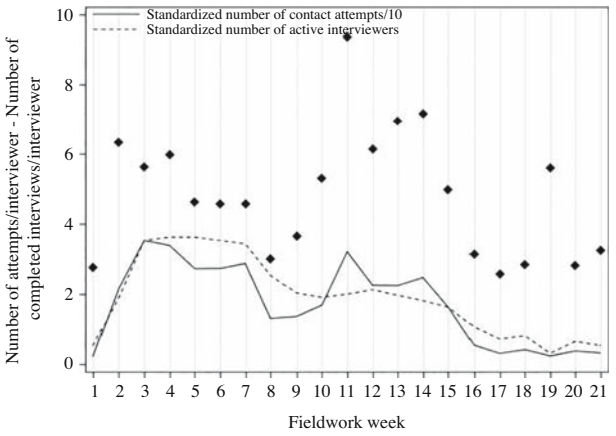

Fig. 2c. Monitoring the effort metrics BE-R7.

Subsequently, the fieldwork power in weeks 3, 4, and 5 is satisfactory, and even higher than expected based on the "previous rounds" benchmark (which is probably the best fitting). In week 6, however, we observe a third successive week of drop in power (both contacts and completed interviews), coinciding with a lower mean number of contact attempts but the number of active interviewers staying the same. Week 6 should therefore have been flagged, and the reason for the reduced effort at the interviewer level should have been investigated. We can hypothesize that some of the interviewers had worked through most of their assigned addresses and should have been given another set.

In week 7, the standardized number of completed interviews displays a sudden increase, whilst the number of contacts keeps dropping (this corresponds to a relatively high performance in week 7. See Figure 3a).

Weeks 8, 9, and 10 display a substantial drop in the fieldwork power, which falls under all the benchmark bands (except week 10), corresponding to a sharp reduction in the number of active interviewers and contact attempts (Figure 2c). This corresponds to a transition between the first release of sample addresses and the second release of sample addresses paired with the start of refusal conversion. In light of this monitoring, we believe that the second phase of the fieldwork should have been started sooner (in week 6, the third consecutive week of reduction).

Week 11 displays a peak in both power and effort, indicating the start of the second phase. However, there is no peak in the number of active interviewers, which is consistent with our knowledge of the organization of the fieldwork agency and of the fieldwork. Only a subset of interviewers is given a second set of sample addresses or reassigned to refusal conversion during the fieldwork period.

In the following three weeks (12, 13, and 14), the power is lower than in week 11, higher than all the benchmark bands, and stable. The power decreases again in weeks 15 and 16, to start to form the fieldwork tail.

In Figure 2d, the fine and medium dashed lines represent the cumulative estimates for the mean age (years) and the mean alcohol consumption during weekdays (grams), together with their 95% confidence intervals (grey bands). The long dashed line represents the cumulative percentage of women among respondents living with a partner.

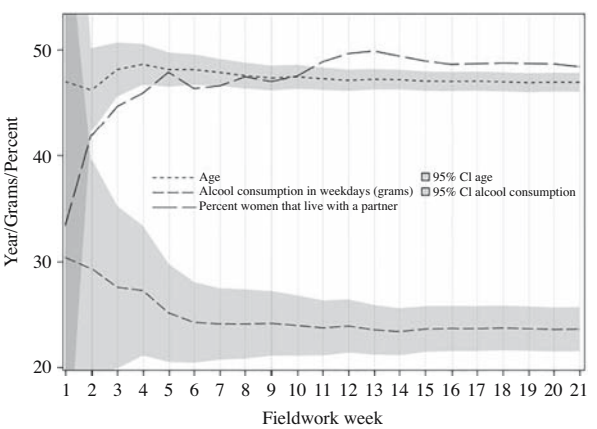

Fig. 2d. Monitoring the data quality metrics BE-R7.

For age, the pre-set sampling error is $SE_{pre} = 19.082/\sqrt{1500} = 0.493$. The estimates of SE_{pre} for the alcohol consumption each week are given in the following Table 4:

Table 4. Pre-set sampling errors for alcohol consumption based on the previous week – BE R7.

Week	SE_{pre} alcohol consumption
1	.
2	1.20
3	1.41
4	1.53
5	1.34
6	1.23
7	1.21
8	1.23
9	1.19
10	1.15
11	1.12
12	1.12
13	1.09
14	1.07
15	1.07
16	1.05
17	1.05
18	1.05
19	1.05
20	1.05
21	1.04

The results show that the estimates for age and for alcohol consumption are stable from week 7. In week 7, for the second week in a row, the difference in the estimates is lower than the SE_{pre}. This corresponds to the end of the first phase (in fieldwork power), before the transition period in weeks 8, 9, and 10. The sampling error for age becomes smaller than SE_{pre} in week 14 and remains smaller in week 15. This could be an indication that the fieldwork had reached its aims in week 15, and corresponds to the findings from the previous graphs that the fieldwork reaches its tail after week 15. However, the sampling error of alcohol does not become smaller than SE_{pre}.

The percentage of women living with a partner is systematically under 50% and peaks in weeks 5 and 8. During weeks 11, 12, 13, and 14, an increase in the percentage of women that live with a partner can be observed. From our knowledge, the corresponding peak in power is due partially to refusal conversion. This observation therefore suggests that women are more likely to initially provide "soft refusals," which can be more easily converted.

After week 15, the main goal is only to reach the minimum required number of completed interviews, but this does not contribute to higher data quality.

6.2.3. Efficiency and Performance in Belgium

Figures 3a and 3b display the three benchmark curves, as well as the achieved efficiency (Figure 3a) and performance (Figure 3b) in Belgium Round 7.

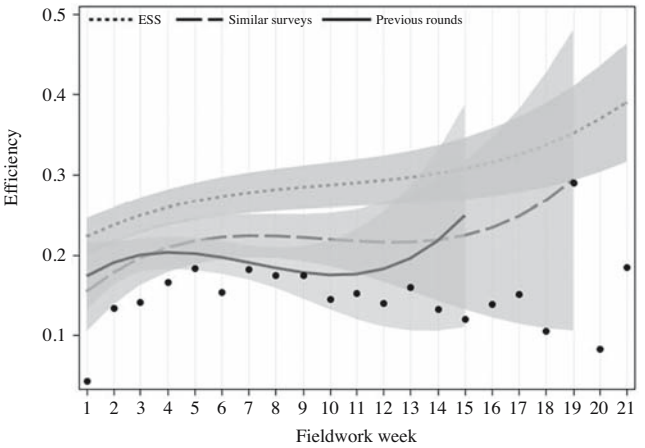

Fig. 3a. Monitoring the weekly efficiency – BE R7.

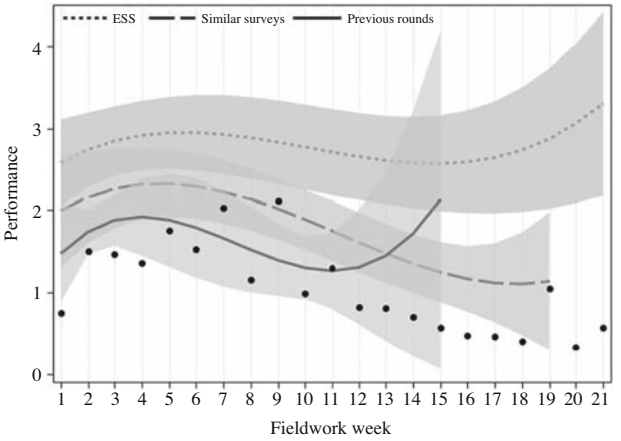

Fig. 3b. Monitoring the weekly performance – BE R7.

Both the efficiency and the performance are, in general, systematically lower than the ESS overall, the "similar surveys" and the "previous rounds" benchmark curves, so almost all weeks of the fieldwork should have been flagged. The only exceptions are the performance in weeks 7, 9, and 11. In weeks 1, 2, 3, 4, and 6, the efficiency is lower than all the confidence bands, showing a low number of contact attempts converted to interviews. In weeks 1, 3, and 4, the performance is also lower than all the confidence bands. This confirms the previous conclusion that the start of the fieldwork was very slow: Not only was the effort in the first two weeks low, but also the success of the contact attempts.

Noticeably, the efficiency and performance fall in the confidence band for the "previous round" curve in weeks 8, 9, and 10, corresponding to a low number of completed interviews and contacts. This demonstrates again that the lower power during this period was due to reduced effort, rather than due to difficult circumstances. The observed peak in

completed interviews and contacts in week 11 also corresponds to a slight increase in performance.

The performance decreases substantially after week 11, when most worked cases are refusal conversions. After week 14, the number of contact attempts is small and we should be careful about interpreting the values of efficiency and performance.

In conclusion, we should pay attention to the way the fieldwork agency organizes the fieldwork. These observations are very important in the post-evaluation of the fieldwork and useful for the implementation of a new round.

6.3. The Czech Republic

We apply the same analysis strategy to the fieldwork evolution of the Czech Republic Round 7 as we did for Belgium. However, we have less information about the organization of the fieldwork, and therefore we need to speculate more with regard to the reasons for deviations from the expected evolution pattern. We believe this is an interesting illustration, as the survey conditions in the Czech Republic are very different to those in Belgium: The obtained response rate is often higher, even though the survey duration is shorter.

6.3.1. Benchmark Curves for the Czech Republic

In Round 7, the sampling design for the Czech Republic is a four-stage cluster sample design. In the first step, basic settlement units were selected and then six addresses were sampled within each selected settlement. In the third and fourth stages, one household was selected from each sampled address and one person was selected within each selected household. The response rate is 67.9% for Round 7 and between 43.3% and 70.2% in previous rounds. The percentages of initial refusals that were re-approached by a different interviewer never exceed 3% in the previous rounds. The fieldwork duration in previous rounds is between 6 weeks in Round 4, and 15 weeks in Rounds 1 and 2.

The second benchmark curve is based on "similar surveys," defined as surveys in previous rounds with no individual sampling frame, response rates above 65%, and refusal conversion lower than five percent. These surveys are: Bulgaria, Round 4; Cyprus, Rounds 3, 4, 5, and 6; the Czech Republic, Rounds 4, 5, and 6; Greece, Rounds 1, 4, and 5; Lithuania, Round 6; the Netherlands, Round 1; Portugal, Rounds 1, 2, 3, 4, 5, and 6; Romania, Round 6; Russia, Round 6; the Slovak Republic, Rounds 5 and 6; Ukraine, Rounds 2 and 3; and Kosovo, Round 6. We limited the analysis to 19 weeks for convergence reasons, excluding the last weeks of the fieldwork for: Cyprus, Round 5 (24 weeks); the Netherlands, Round 1 (23 weeks); and Portugal Rounds 2, 4, and 6 (respectively 22, 20, and 22 weeks). Only Round 4 in Portugal reached a response rate of 65% in week 19, the other surveys had not reached this criterion and are therefore not considered for the estimation of the "similar surveys" benchmark.

Table 5 displays the fixed parameters for the multilevel model (Model 1) with repeated measurements limited to the above surveys. The model fit for the successive nested model can be found in Appendix (online at: http://dx.doi.org/10.1515/JOS-2017-0031), even though we aim to have the best fitting cubic curve and not the best fitting model.

Table 5. Fixed effects for the multilevel model describing the shape of the fieldwork power evolution for ESS surveys similar to the Czech Republic round 7 (Model 1).

Effect	Completed interviews	Contacts	Efficiency	Performance
γ_{00} – intercept	5.69 (1.32)***	10.88 (2.55)***	.39 (0.04)***	5.01 (0.85)***
γ_{10} – linear term	1.97 (1.52)**	2.78 (0.92)**	.014 (.020)	0.34 (0.35)
γ_{20} – quadratic term	−0.39 (0.07)***	−0.63 (0.10)***	-3^{E-4} (.003)	−0.08 (0.05)
γ_{30} – cubic term	0.02 (0.002)***	0.03 (0.004)***	4^{E-5} (1^{E-4})	3^{E-3} (2^{E-3})

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{****}p < 0.001$.

The yield of standardized completed interviews is higher than the ESS overall, even though the standardized number of contacts is lower. The increase in the first week is also steeper (linear term), and the leveling-off leading to the decrease in yield also comes earlier than in the ESS curve. This shows that the fieldwork production in the Czech Republic is expected to have a high and short peak, based on the similar surveys. This holds for the performance as well, whilst the efficiency seems to have a flatter evolution curve.

The covariance parameters are all significant at the 0.05 level for the number of contacts. The variance of the intercepts for the number of completed interviews, the efficiency, and the performance, and the covariance of the intercept and the linear term for the number of completed interviews are also significant at the 0.05 level (results not shown).

Table 6 displays the fixed parameters for the benchmark based on the previous ESS rounds in the Czech Republic (the model fit can be found in Appendix online at: http://dx.doi.org/10.1515/JOS-2017-0031). We limit the analysis to seven weeks for convergence reasons.

In the first week, the yield of the fieldwork is lower (completed interviews and contacts) than for the ESS overall and for similar surveys, but increases more rapidly at the beginning of the survey (linear term). The quadratic term is large in absolute value, showing that the rapid increase quickly levels off. This is typical for very short fieldwork duration. The efficiency and the performance are lower than for similar surveys but higher than the ESS overall. The general shape, however, is different to the fieldwork power evolution that we have observed so far. The efficiency and performance decrease at first (negative linear term) to then increase (quadratic linear term) and subsequently level off.

Table 6. Fixed effects for the multilevel model describing the shape of the fieldwork power evolution for ESS surveys previous rounds in the Czech Republic (Model 1).

Effect	Completed interviews	Contacts	Efficiency	Performance
γ_{00} – intercept	3.68 (2.41)	7.98 (5.46)	.341 (.064)**	3.31 (1.02)*
γ_{10} – linear term	4.68 (3.84)	16.51 (7.59)	−.112 (.087)	−3.45 (1.12)*
γ_{20} – quadratic term	−0.93 (1.60)	−5.58 (3.16)	.062 (0.24)	1.64 (0.63)
γ_{30} – cubic term	0.03 (0.18)	0.47 (0.35)	−.006 (.003)*	−0.11 (0.04)*

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{****}p < 0.001$.

The covariance parameters (results not shown) are much smaller, showing that the fieldwork evolution in the different rounds in the Czech Republic are more closely related than when comparing with other surveys in the ESS. The lower significance of fixed and random parameters are, however, mainly due to the smaller number of surveys involved – only five – and as a consequence, fewer measurements of the fieldwork power.

6.3.2. Fieldwork Power: Standardized Number of Completed Interviews and of Contacts

Figures 4a and 4b show the three benchmark curves and confidence bands, and the achieved weekly numbers of completed interviews (Figure 4a) and contacts (Figure 4b) in the Czech Republic, Round 7.

Following the flagging rules, week 3 should have been flagged as the second week in a row in which the fieldwork power is below the confidence band for all the benchmarks.

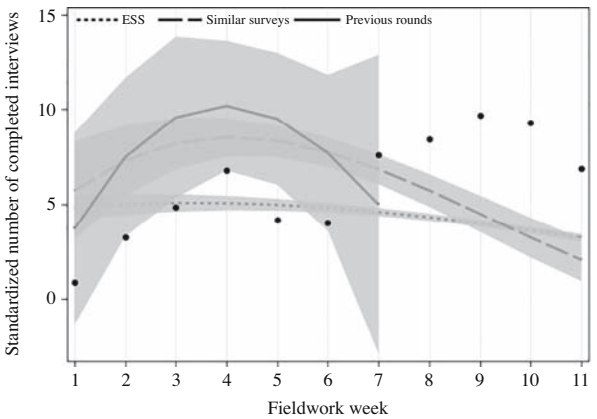

Fig. 4a. Monitoring the weekly number of completed interviews – CZ R7.

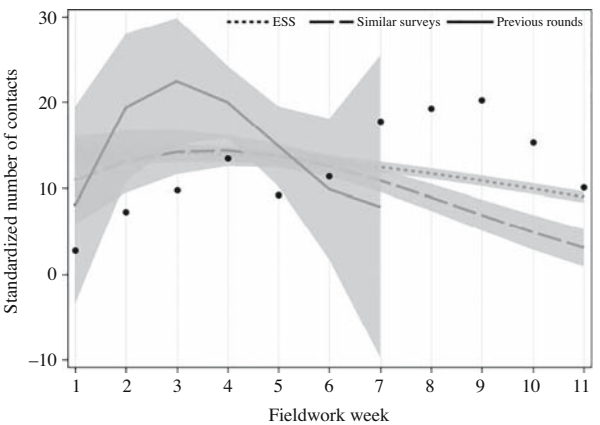

Fig. 4b. Monitoring the weekly number of contacts – CZ R7.

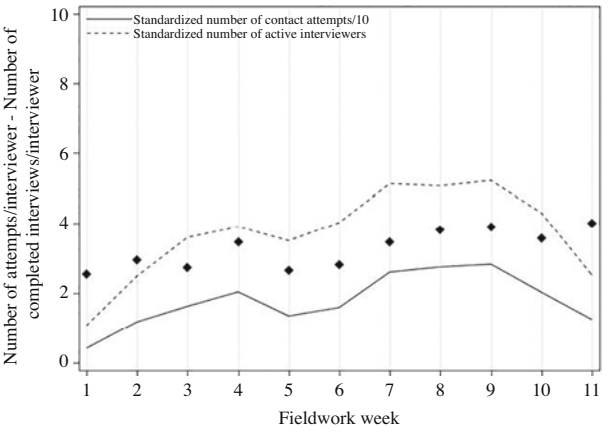

Fig. 4c. Monitoring the effort metrics – CZ R7.

Further, we see that, even though week 4 has a higher power, it is still under the benchmark and week 5 is again below the confidence band. Week 6 falls inside the confidence band of the "previous rounds" benchmark, but still below all the benchmarks for the standardized number of completed interviews. Looking at Figure 4c, we can observe that the standardized number of active interviewers increases over the first three weeks, but the intensity of the work carried out by them remains stable, leading to a less steep increase in the number of contact attempts. Looking further at the efficiency and performance curve, it does not seem that the low yield in the first three weeks results from difficulties in the field, as the efficiency and performance are within the confidence bands of the previous rounds benchmarks. The effort, in terms of the number of active interviewers and their number of contact attempts, is probably lower than planned, although this is speculative as we do not have information about the fieldwork organization.

In weeks 7 to 11, the fieldwork power increases, being above the confidence bands of the ESS overall and the similar surveys curve. The previous rounds curve stops in week 8, because the fieldwork period had been shorter in previous rounds. This illustrates the potential need for different benchmarks to the previous rounds one, although we would expect this to provide a better projection. The weekly fieldwork power in this second phase is higher than in the first phase (week 1 to 6), which is a clear deviation from the expected pattern, namely a peak in the first part followed by a lower tail. In weeks 7, 8, and 9, this is paired with an increase in the standardized number of active interviewers and of contact attempts. Both effort metrics, however, drop in weeks 10 and 11.

The reason why the real core of the fieldwork effort only occurred after week 6 should be investigated. This late start could have been planned, but was possibly also a reaction to the low productivity in the first weeks that triggered the activation of more interviewers. Closer monitoring would have detected the problem at an earlier stage. Two main reasons could have caused the fieldwork to start slowly. First, not enough interviewers started in the first week because not enough of them were available, or alternatively, because they were left free to start working on their set of addresses whenever it suited them. In the latter case, interviewers should have been asked to start working straight away. In the first

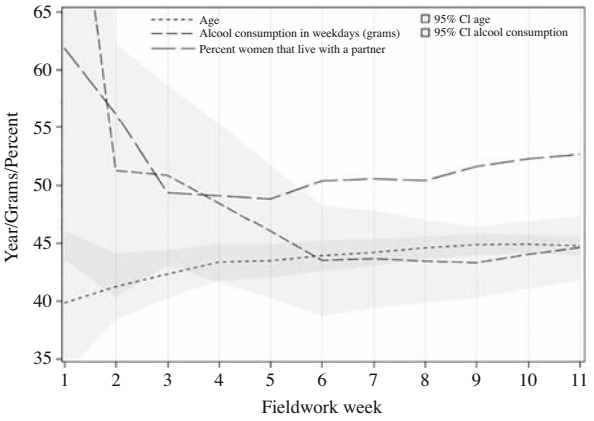

Fig. 4d. Monitoring the data quality metrics – CZ R7.

case, this is more a planning issue. The second reason could be that the active interviewers were expected to work more intensively than they did, in which case they should have been contacted and asked to work more intensively.

Observing Figure 4d, the percentage of women living together with a partner increases in weeks 8, 9, and 10 to above 50%, showing that more women than men in households completed interviews in this period.

For age, the pre-set sampling error is $SE_{pre} = 17.146/\sqrt{1500} = 0.443$. The estimates of SE_{pre} for the alcohol consumption each week are given in the following Table 7:

Table 7. Pre-set sampling errors for alcohol consumption based on the previous week – CZ R7.

Week	SE_{pre} alcohol consumption
1	.
2	1.34
3	1.34
4	1.49
5	1.42
6	1.35
7	1.39
8	1.33
9	1.26
10	1.33
11	1.33

The difference in age estimates from one week to the other comes below the pre-set sampling error in weeks 5 and 6, however, we see a constant increase in the estimated value toward the end of the fieldwork. The sampling error becomes smaller than the pre-set value in the week before the final week and remains smaller in the final week. The difference in alcohol estimates from one week to another only comes below the pre-set sampling error in

weeks 7 and 8, and the sampling error of alcohol consumption never reaches the pre-set value. This shows that the stabilization phase of the estimates was not completely reached.

6.3.3. Efficiency and Performance in the Czech Republic

Figures 5a and 5b show the three benchmark curves and their confidence bands, and the achieved efficiency and performance for the Czech Republic, Round 7.

In weeks 1 to 6, the weekly efficiency (Figure 5a) is systematically lower than in previous rounds and in "similar surveys" but, in general, higher than the overall ESS efficiency. Week 3 could therefore have been flagged. In week 6, the efficiency is at the lower edge of the previous round confidence band and lower than the ESS benchmark, which is surprising. This should have been flagged, together with the low yield of the fieldwork in the first weeks.

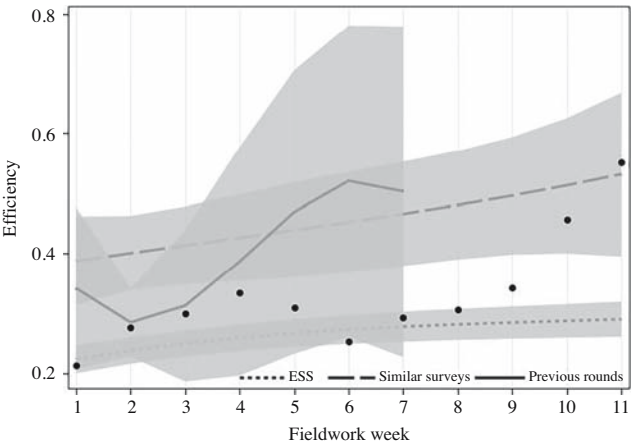

Fig. 5a. Monitoring the weekly efficiency – CZ R7.

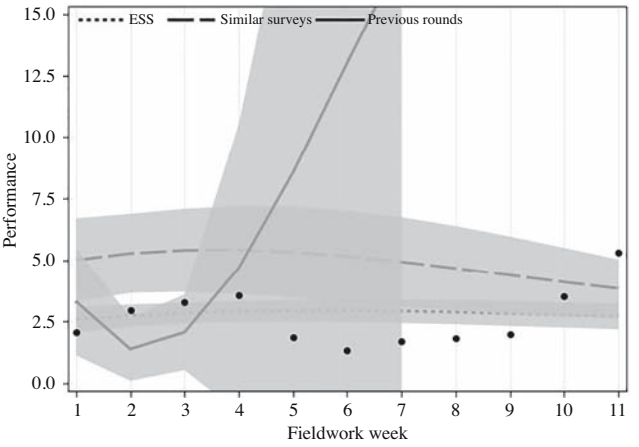

Fig. 5b. Monitoring the weekly performance – CZ R7.

The efficiency reaches surprisingly high levels in week 10 (more than 40%) and week 11 (more than 50%), meaning that between 40 and 50% of the contact attempts were converted into completed interviews during these weeks. This is certainly a large deviation from the ESS curve. It is highly unusual for such efficiency to be achieved, and hence these results raise questions. However, the last two weeks of fieldwork could have been used to finalize open cases that were likely to be converted into completed interviews, such as soft refusals or appointments.

The weekly performance (Figure 5b) also follows a somewhat unexpected pattern. In weeks 1, 2, 3, and 4, it is within or higher than all the benchmark bands but falls below the benchmark bands in weeks 5 to 9 (with the exception of the very large "previous round" band) when the number of completed interviews and contacts was higher, showing that the increase in effort produced more completed interviews but also more refusals. Again, the performance is at its highest in week 10 and 11; in the last week it even reached higher than all the benchmark bands.

Comparing the achieved fieldwork power with the benchmark raises specific and precise questions. In this example, one could question how such a high level of efficiency could be reached at the end of the fieldwork. The use of incentives or follow-ups to appointments could be an explanation, but this unexpected result could also point to deviations from the required sampling and contact procedures.

7. Conclusion

Inspired by a principle of classical mechanics, we introduce the "fieldwork power" as a fieldwork performance indicator to be monitored. Fieldwork power can have different specifications in terms of productivity per time unit, from which we focus on four: The weekly (standardized) number of completed interviews, the weekly (standardized) number of contacts, the efficiency (ratio of completed interviews to contact attempts), and the performance (ratio of completed interviews to refusals). These four specifications can be seen as Key Performance Indicators (KPIs) that have to be monitored during the fieldwork period. As we argue, in order to be monitored meaningfully an indicator needs to have *control limits* – boundaries between which values are acceptable- or a *benchmark* – giving an evolution pattern against which it can be compared in a dynamic way. To develop a benchmark, we need empirical information. Repeated surveys such as the ESS lend themselves well to this type of analysis. To monitor Round 7 of the ESS in Belgium and in the Czech Republic, we created three benchmarks based on a multilevel model with repeated measurements of the four KPIs. The macro-level units are surveys from previous ESS rounds, and the repeated measurements within the surveys are the fieldwork power. The set of surveys entered into the model were increasingly refined in terms of design similarities with the survey to be monitored: A global ESS benchmark curve based on all ESS surveys from Round 1 to 6, a "similar surveys" benchmark curve based on ESS surveys that shared design characteristics (sampling frame, percentage of refusal conversion, and response rate), and a "previous rounds" benchmark curve based on surveys in the same country in earlier ESS rounds. The different benchmarks can be useful in different situations. A new country may want to join the ESS, for which the overall ESS benchmark could serve as the reference, even though large deviations may be normal due

to specific survey design characteristics in that country. As we see in the examples, the fieldwork can be longer than the "similar surveys" and "previous rounds" benchmarks allow us to model for, and the ESS curve can then become the reference curve. Nevertheless, when available, the "previous rounds" curve is probably the most appropriate one. If a survey design changes, the "similar surveys" curve may be more suitable. Moreover, the confidence bands of the previous rounds are generally large given the low number of measurements entering the model estimation. Therefore, it may be useful to consider the three benchmarks simultaneously. The "similar surveys" – and more particularly the "previous rounds" – benchmarks do not seem to be such good predictors of the fieldwork power. Changes in survey design characteristics might explain this, but the lack of long-term country-specific accumulated knowledge might also be the origin of the problem. We may expect our benchmarks based on previous rounds to be increasingly precise, due to the higher number of measurements entering the model and also less influenced by a specific round deviation.

We then simulated monitoring the evolution of the four specifications of the fieldwork power, together with some metrics for effort (the number of active interviewers and contact attempts overall and mean per interviewer) and data quality (mean estimates, and standard errors of age and alcohol consumption and percentage women among respondents living with a partner).

The results of simulated monitoring of the ESS Round 7 in Belgium and the Czech Republic show that the benchmarks help to detect deviating patterns during the fieldwork and help in post-fieldwork evaluation. This type of analysis also allows us to raise specific questions about the implementation of the fieldwork.

In Belgium Round 7, the fieldwork power in terms of completed interviews and contacts was very low compared with the benchmarks in weeks 8, 9, and 10. This low power corresponds to lower efforts in terms of active interviewers and contact attempts, and not to difficulties in the field, as the efficiency and performance were not low during this period. This corresponds to a transition period between the first phase after the first release of addresses, and a second phase after the second release of addresses and the start of refusal conversion. One can observe that the transition period is too long and that the second phase could have been started earlier. Indeed, in week 11 and the following weeks, the fieldwork power (and effort) was higher than the benchmark curves – probably due to a reaction to the previous week – but the performance was lower. This also corresponds to a small peak in the percentage of women among respondents living with a partner, showing that the fieldwork process also influences data quality. After week 15, the fieldwork power levels off and the performance decreases, while the mean estimates and the width of the confidence interval stabilize. The last weeks of the fieldwork only serve to comply with the specification for the number of completed interviews, but bring little new information to the survey data.

In the Czech Republic Round 7, the evolution of the fieldwork power in terms of completed interviews and contacts displays a largely deviating pattern from the benchmarks: Low in the first six weeks and then high toward the end of the fieldwork. The effort metrics follow the same pattern. No tail in the evolution of the fieldwork power is observed. In addition, no stabilization of the estimates can be seen. The increase in efficiency and performance at the end of the fieldwork also raise questions: The very high

efficiency (more than 40%) in the last two weeks is intriguing, and additional analysis of the contact forms for the last two weeks (e.g., the interval between contacts) is recommended.

These illustrations for Belgium and the Czech Republic show how the fieldwork power and its benchmarks could be used as an indicator in a fieldwork monitoring context, although practical considerations – for example timely access to contact data and fast processing of it – have to be taken into account. We use the ESS, although we acknowledge that the lack of a centralized control center may make the implementation of fieldwork monitoring difficult. We therefore advocate better communication of fieldwork evolution indicators between the fieldwork agency, the national coordinators/representatives, and the survey headquarters. The fieldwork power together with its benchmarks is, however, also a powerful tool for the post-fieldwork evaluation process.

Further research may consider extending this analysis to more countries participating in the ESS, in order to obtain a better picture of the specific aspects of the fieldwork implementation in particular countries. The analysis strategy could also be expanded to other repeated or longitudinal surveys. Moreover, we concentrate here on four specifications of the fieldwork power, but more research should be carried out to assess if other specifications, for instance the number of new contacts made each week, could be useful. Although so far, the different specifications considered seem to validate and to complement each other: A low standardized number of completed interview paired with a "normal" efficiency, hints at a lack of effort. The effort metrics and data quality metrics could also be further developed: The weekly budget expended, number of hours worked per interviewer, or the evolution of other variable means or other relevant statistics may be of interest. Lastly, the feasibility of implementing the fieldwork follow-up "live" for monitoring should also be assessed.

8. References

European Social Survey (ESS). 2014. *European Social Survey Round 7 Data. Data file edition 1.0*. Norway: Norwegian Social Science Data Services. Data Archive and distributor of ESS data for ESS ERIC.

European Social Survey (ESS). 2011. *Round 6 Specification for Participating Countries*. London: Centre for Comparative Social Surveys, City University London. Available at: http://www.europeansocialsurvey.org/docs/round6/methods/ESS6_project_specification.pdf (accessed March 2017).

Groves, R.M. and S.G. Heeringa. 2006. "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society. Series A: Statistics in Society* 169(3): 439–457. Doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00423.x.

Jans, M., R. Sirkis, and D. Morgan. 2013. "Managing Data Quality Indicators with Paradata based Statistical Quality Control Tools: the Keys to Survey Performance." In *Improving Surveys with Paradata. Analytic Uses of Process Information*, edited by Frauke Kreuter, 191–229. Hoboken, New Jersey: John Wiley & Sons.

Juran, J.M. and F.M. Gryna. 1988. *Juran's Quality Control Handbook*, 4th edition. New York: McGraw-Hill.

Kirgis, N.G. and J.M. Lepkowski. 2013. "Design and Management Strategies for Paradata-driven Response Design: Illustration from the 2006–2010 National Design of Family Growth." In *Improving Surveys with Paradata. Analytic Uses of Process Information*, edited by Frauke Kreuter, 191–229. Hoboken, New Jersey: John Wiley & Sons.

Kohler, U. 2007. "Surveys from inside: An Assessment of Unit Nonresponse Bias with Internal Criteria." *Survey Research Methods* 1(2): 55–67. Doi: http://dx.doi.org/10.18148/srm/2007.v1i2.75.

Kreuter, F., M. Couper, and L. Lyberg. 2010. "The Use of Paradata to Monitor and Manage Survey Data Collection." In *Section on Survey Research Methods: American Statistical Association*, 282–296. Vancouver: American Statistical Association. Available at: http://ww2.amstat.org/sections/SRMS/Proceedings/y2010/Files/306107_55863.pdf (accessed March 2017).

Laflamme, F., M. Maydan, and A. Miller. 2008. "Using Paradata to Actively Manage Data Collection Survey Process." In *Section on Survey Research Methods: American Statistical Association*, 630–637. Denver: American Statistical Association. Available at: http://ww2.amstat.org/sections/srms/Proceedings/y2008/Files/300608.pdf (accessed March 2017).

Laflamme, F. 2009. "Data Collection Research Using Paradata at Statistics Canada." In *Proceeding of Statistics Canada's International Symposium Series*. Ottawa: Statistics Canada. Available at: http://www.statcan.gc.ca/pub/11-522-x/2008000/article/10997-eng.pdf (accessed March 2017).

Luiten, A. and B. Schouten. 2013. "Tailored Fieldwork Design to Increase Representative Household Survey Response: An Experiment in the Survey of Consumer Satisfaction." *Journal of the Royal Statistical Society. Series A: Statistics in Society* 176(1): 169–189. Doi: http://dx.doi.org/10.1111/j.1467-985X.2012.01080.x.

Lundquist, P. and C.-E. Särndal. 2013. "Aspects of Responsive Design with Applications to the Swedish Living Conditions Survey." *Journal of Official Statistics* 29(4): 557–582. doi: https://doi.org/10.2478/jos-2013-0040.

Malter, F. 2013. "Fieldwork Monitoring in the Survey of Health, Ageing and Retirement in Europe (SHARE)." *Survey Methods: Insights from the Field*, 1–8. http://surveyinsights.org/?p=1974. Doi: http://dx.doi.org/10.13094/SMIF-2014-00006.

Matsuo, H. and G. Loosveldt. 2013. *Report on quality assessment of contact data files in Round 5: Final report 27 countries*. Leuven: European Social Survey, University of Leuven. Available at: https://www.europeansocialsurvey.org/docs/round5/methods/ESS5_response_based_quality_assessment_e01.pdf (accessed March 2017).

Peytchev, A., S. Riley, J. Rosen, J. Murphy, and M. Lindblad. 2010. "Reduction of Nonresponse Bias in Surveys through Case Prioritization." *Survey Research Methods* 4(1): 21–29. Doi: http://dx.doi.org/10.18148/srm/2010.v4i1.3037.

Schouten, B., J. Bethlehem, K. Beullens, O. Kleven, G. Loosveldt, A. Luiten, K. Rutar, N. Shlomo, and C. Skinner. 2012. "Evaluating, Comparing, Monitoring, and Improving Representativeness of Survey Response through R-Indicators and Partial R-Indicators." *International Statistical Review* 80(3): 382–399. Doi: http://dx.doi.org/10.1111/j.1751-5823.2012.00189.x.

Schouten, B., M. Calinescu, and A. Luiten. 2013. "Article Optimizing Quality of Response through Adaptive Survey Designs." *Survey Methodology* 39(1): 29–58. Available at: http://www.statcan.gc.ca/pub/12-001-x/2013001/article/11824-eng.pdf (accessed March 2017).

Schouten, B. and N. Shlomo. 2015. "Selecting Adaptive Survey Design Strata with Partial R-Indicators." *International Statistical Review*, Doi: http://dx.doi.org/10.1111/insr.12159.

Shewart, W.A. 1931. *Economic Control of Quality of Manufactured products*. Princeton, New Jersey: van Nostrand Reinhold Co.

Shewart, W.A. 1939. *Statistical Methods from the View Point of Quality control*. Washington: The Graduate School, the department of Agriculture.

Vandenplas, C., G. Loosveldt, and K. Beullens. 2015. "Better or Longer? The Evolution of Weekly Number of Completed Interviews over the Fieldwork Period in the European Social Survey." Paper presented at the Nonresponse Workshop, Leuven 2015. Available at: https://lirias.kuleuven.be/handle/123456789/521789 (accessed March 2017).

Vandenplas, C. and G. Loosveldt. 2017. "Modeling the Evolution of Fieldwork in Terms of Quality Indicators: Six Rounds of the European Social Survey." *Journal of Statistics and Methodolgy* 5(2): 212–232. Doi: http://dx.doi.org/10.1093/jssam/smw034.

Wagner, J. 2008. *Adaptive Survey Design to Reduce Nonresponse Bias*. Unpublished Doctoral Thesis, University of Michigan. Available at: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/60831/jameswag_1.pdf?sequence=1 (accessed March 2017).

Zhang, L.C., I.B. Thomsen, and O. Kleven. 2013. "On the Use of Auxiliary and Paradata for Dealing with Non-Sampling Errors in Household Surveys." *International Statistical Review* 81(2): 270–288. Doi: http://dx.doi.org/10.1111/insr.12009.

Received March 2016
Revised March 2017
Accepted April 2017