Response Variance Estimation in Personal Interview Surveys with Several Interviewer Allocation Schemes

İnci Batmaz*
Middle East Technical University, Ankara, Turkey – ibatmaz@metu.edu.tr

Fidan Mahmut-Fahmi
Middle East Technical University, Ankara, Turkey – fidantelaferli@yahoo.com

Abstract

Linear and multiplicative models have been used to explain the factors effecting nonsampling errors, which naturally includes response error models. The response error models are generally evaluated as simple response errors while some others evaluated as correlated response errors. They naturally take the interaction between the interviewer and the respondent into account. In this work, the optimum interviewer allocation settings have been investigated by using different experimental design plans including nested, nested and factorial factors and split plot designs to study the sources of both types of errors. The proposed designs consider several stages of an interactive process

Keywords: survey design; response error; experimental design, face to face interviewing.

1. Introduction

Survey research is widely used in different areas such as medicine, astronomy, psychology, media for the purpose of gathering information about population of interest (Kish, 1995). To be comfortable with using survey data, we need to get information about the "quality of survey" by studying the sources of errors involved, which are classified as sampling and nonsampling errors (Lessler and Kalsbeek, 1992). While sampling errors occur due to observing a sample, nonsampling errors include all types of errors occur during survey activities other than sampling. Response error is a kind of nonsampling error that may arise from interviewer's effects on the respondents' answers. Survey researchers concerning survey errors are mostly interested in either measuring their effects on results or findings ways of eliminating them. Survey errors are described as deviations of survey results from the true values, and researches interested in estimating them such as Mahalanobis (1946), Hansen et al. (1961), Blair & Biemer, (1984), Biemer & Stokes (1991), Ayhan, (2003) and Alimohammadi & Navvabpour (2008), proposed models that incorporate variety of error sources at individual unit level. Assuming that two observations, y_{j1} and y_{j2}, are obtained from an interview-reinterview survey, differences within pairs of observations provide data for the reliability investigation. For categorical type of data, typical measures include *crude index*, *Kappa* and *measure of disagreement*. For metric data, *index of inconsistency*, *reliability of data*, and the *correlation* between the responses for the first and second interviews are among the most popular measures of reliability.

When interviewers cause errors in surveys, correlated errors are observed. The current practice in the area is to prevent the occurrence of nonsampling errors; this inevitably requires their detection and control. Under this motivation, in this study, a kind of nonsampling error, response error, is investigated. Under the assumption that "interviewers affect responses," several methodologies are proposed for developing response error models in complex interview-reinterview surveys, where observations are collected from each strata after the population is divided into stratums, to obtain efficient estimators for response error components by utilizing various design of experiments (DOEs). These include *nested design* (ND) (Ayhan, 2003), *nested and factorial design* (NFD) and *split plot design* (SPD) (Ayhan, 2012; Mahmut-Fahmi, 2013).

In the following section, three different methodologies, named as *allocation schemes*, involving experimental design plans in fieldworks for allocating interviewer and respondents as well as related variance component estimators are presented. Concluding remarks are stated in the last section.

2. Interviewer Allocation by Experimental Designs

The proposed methodologies involve schemes for allocating interviewer and respondents that can be achieved in two or more successive rounds by using various DOEs (Montgomery, 2012).

Allocation Scheme-1: Allocation by nested design. In some experiments the levels of one factor (e.g. B) are similar but not identical for the levels of another factor (e.g. A). Such an arrangement is called a ND with the levels of factor B nested under the levels of factor A. The response error model involving three factors, namely, controller, interviewer and respondent, can be written as

$$y_{ijk} = \mu_{ijk} + \tau_i + \beta_{j(i)} + \gamma_{k(ij)} + \varepsilon_{ijk}, \qquad \begin{cases} i = 1,2,\dots,a \\ j = 1,2,\dots,b \\ k = 1,2,\dots,c \end{cases} \quad (1)$$

where μ_{ijk} represents the true value, τ_i is the i^{th} controller error, $\beta_{j(i)}$ is the j^{th} interviewer error nested under the i^{th} controller, $\gamma_{k(ij)}$ is the k^{th} respondent error nested under the j^{th} interviewer and the i^{th} controller, $\varepsilon_{ijk}$is the NID$(0,\sigma^2)$ random error term (Ayhan, 2003; Mahmut-Fahmi, 2013). Let μ and σ^2 be the mean and variance, of y respectively. Then,

$$\mu = \frac{\sum_i^a \sum_j^b \sum_k^c \mu_{ijk}}{abc}, \text{ and } \sigma^2 = E\{\frac{1}{an}\sum_i^a \sum_j^b \sum_k^c (y_{ijk} - \mu)^2\} = \sigma_s^2 + \sigma_r^2, \quad (2)$$

where σ_s^2 is the sampling variance defined as,

$$\sigma_s^2 = \frac{1}{a}\sum_{i=1}^a \sigma_{si}^2 + \frac{1}{a}\sum_{i=1}^a (\mu_i - \mu)^2. \quad (3)$$

Here, $\mu_i = \sum_j^b \sum_k^c \frac{\mu_{ijk}}{bc}$ and $\sigma_{si}^2 = E\left(\frac{1}{bc}\sum_j^b \sum_k^c c_{ijk}^2\right)$, where $c_{ijk} = \mu_{ijk} - \mu_i$ is the sample deviation of the same unit. Also, the response variance is defined as

$$\sigma_r^2 = \frac{1}{a}\sum_{i=1}^a \sigma_{ri}^2, \quad (4)$$

where $\sigma_{ri}^2 = E\left(\frac{1}{bc}\sum_j^b \sum_k^c d_{ijk}^2\right)$ is the uncorrelated response variance, and d_{ijk} is the response deviation of unit k, enumerated by interviewer j, controlled by the controller i. Here, the response deviation is determined as $d_{ijk} = \tau_i + \beta_{j(i)} + \gamma_{k(ij)}$. Besides, δ_{2i} is the correlation coefficient between the response deviations within interviewers' assignments, and it is defined as

$$\delta_{2i} = \frac{1}{\sigma_{ri}^2} E\{\frac{1}{n(c-1)}\sum_j^b \sum_{k\neq k'}^c d_{ijk} d_{ijk'}\}, \quad (5)$$

and $\delta_{2i}\sigma_{ri}^2$ is the correlated interviewer variance. Here, the superscript refers to a different respondent, interviewer or controller as their location in the study. In addition, δ_{3i} is the correlation coefficient between response deviations within controllers' assignments, and defined as

$$\delta_{3i} = \frac{1}{\sigma_{ri}^2} E\{\frac{1}{nc(b-1)}\sum_{j\neq j'}^b \sum_{k,k'}^c d_{ijk} d_{ij'k'}\}, \quad (6)$$

where $\delta_{3i}\sigma_{ri}^2$ is the correlated controller variance. Also, $\delta_{4ii'}$ is the correlation coefficient among response deviations in two different domains, and defined as

$$\delta_{4ii'} = \frac{1}{\sigma_{ri}\sigma_{ri'}} E\{\frac{1}{b^2c^2}\sum_{j,j'}^b \sum_{k,k'}^c d_{ijk} d_{i'j'k'}\}. \quad (7)$$

To consider the response error associated with the reinterview survey, it is assume that the response model given in (1) is applied to the reinterview data. The variance components are indexed by 1 if it associated with the parent (main) survey and by 2 if associated with the reinterview survey. Moreover, the correlations of the variance model errors between surveys can be defined as follows. β_{1i} is the correlation between response errors of the same unit in different surveys and defined as

$$\beta_{1i} = \frac{1}{\sigma_{ri1}\sigma_{ri2}} E\{\frac{1}{n}\sum_j^b \sum_k^c d_{ijk1} d_{ijk2}\}. \quad (8)$$

β_{2i} is the correlation coefficient between the response errors of two different units interviewed by the same interviewer but in different surveys and given as

$$\beta_{2i} = \frac{1}{\sigma_{ri1}\sigma_{ri2}} E\{\frac{1}{nc}\sum_j^b \sum_{k\neq k'}^c d_{ijk1} d_{ijk'2}\}. \quad (9)$$

β_{3i} is the correlation coefficient between response errors of units belonging to the same controller in the two surveys and given as

$$\beta_{3i} = \frac{1}{\sigma_{ri1}\sigma_{ri2}} E\{\frac{1}{n(c-1)}\sum_j^b \sum_{k\neq k'}^c d_{ijk1} d_{ij^0k'2}\}, \quad (10)$$

where j^0 denotes the interviewer in the reinterview survey, and j denotes the interviewer in the parent survey. β_{4i} is the correlation between response errors of different units assignd to different interviewers in different surveys, and given by

$$\beta_{4i} = \frac{1}{\sigma_{ri1}\sigma_{ri2}} E\{\frac{1}{b(b-1)c^2} \sum_{j' \neq j, j^0}^{b} \sum_{k \neq k'}^{c} d_{ijk1} d_{ij'k'2}\}. \tag{11}$$

$\beta_{5ii'}$ is the correlation coefficient between response deviations of units assigned to different controllers in surveys and given by

$$\beta_{5ii'} = \frac{1}{\sigma_{ri1}\sigma_{ri'2}} E\{\frac{1}{nbc} \sum_{j,j'}^{b} \sum_{k\prime k'}^{c} d_{ijk1} d_{i'j'k'2}\}. \tag{12}$$

An estimator for uncorrelated response variance of stratum I, σ_{ri}^2, suggested by Fellegi (1964) is

$$s_{ri}^2 = \frac{1}{2b(c-1)} \sum_j^b \sum_k^c \left(y_{ijk1} - y_{ij^0k2} - \bar{y}_{ij1} + \bar{y}_{ij^02}\right)^2. \tag{13}$$

He showed that without a second measurement for a subsample of parent survey respondents, the uncorrelated response variance is confounded with the sampling variance, and is therefore, not separately estimable. An estimator for the sum of sampling and uncorrelated response variance, that is, $\sigma_s^2 + \sigma_r^2$ is given by Kish (1962) as

$$s_{si}^2 + s_{ri}^2 = \frac{1}{b(c-1)} \sum_j^c \sum_k^d (y_{ijk} - \bar{y}_{ij})^2. \tag{14}$$

When the reinterview data is available, the estimator of the interviewer variance in stratum I, $\delta_{2i}\sigma_{ri}^2$, is given by Fellege (1964) as

$$I\hat{\delta}_{2i}s_{ri}^2 = \frac{1}{2(b-1)} \sum_j^b (\bar{y}_{ij1} - \bar{y}_{ij^02} - \bar{y}_{i1} + \bar{y}_{i2})^2 - \frac{1}{2bc(c-1)} \sum_j^b \sum_k^c (y_{ijk1} - y_{ij^0k2} - \bar{y}_{ij1} + \bar{y}_{ij^02}).^2 \tag{15}$$

And, without reinterview data, it is given by Kish (1962) as

$$II\hat{\delta}_{2i}s_{ri}^2 = \frac{1}{b-1} \sum_j^b (\bar{y}_{ij} - \bar{y}_i)^2 - \frac{1}{bc(c-1)} \sum_j^b \sum_k^c \left(y_{ijk} - \bar{y}_{ij}\right)^2.$$

An estimator for the controller variance, which combines information from all strata in both surveys is given as

$$I\hat{\delta}_3 s_r^2 = \frac{1}{2(A-1)} \sum_i^a (\bar{y}_{i1} - \bar{y}_{i2} - \bar{y}_1 + \bar{y}_2)^2 - \frac{1}{2Ab(b-1)} \sum_i^a \sum_j^b \left(\bar{y}_{ij1} - \bar{y}_{ij^02} - \bar{y}_{i1} + \bar{y}_{i2}\right)^2. \tag{16}$$

If reinterview is not conducted, then the estimator is given by Fabbris (1991) as

$$II\hat{\delta}_3 s_r^2 = \frac{1}{A-1} \sum_i^a (\bar{y}_i - \bar{y})^2 - \frac{1}{Ab(b-1)} \sum_i^a \sum_j^b \left(\bar{y}_{ij} - \bar{y}_i\right)^2.$$

Allocation Scheme-2: Allocation by nested and factorial factors design. In some experiments, some factors are arranged in a factorial layout while others are nested. The linear model of such a design with three factors can be written as

$$y_{ijk} = \mu + \tau_i + \beta_j + \gamma_{k(j)} + (\tau\beta)_{ij} + (\tau\gamma)_{ik(j)} + \varepsilon_{ijk}, \begin{cases} i = 1,2, \dots, a \\ j = 1,2, \dots, b \\ k = 1,2, \dots, c \end{cases} \tag{17}$$

where μ represents the true value, τ_i is the i^{th} controller error, β_j is the j^{th} interviewer error, $\gamma_{k(j)}$ is the k^{th} respondent error nested under the j^{th} interviewer, $(\tau\beta)_{ij}$ is the interaction error between the i^{th} controller and the j^{th} interviewer, $(\tau\gamma)_{ik(j)}$ is the interaction error between the i^{th} controller and the k^{th} respondent nested under the j^{th} interviewer, ε_{ijk} is NID($0,\sigma^2$) random error term. The associated layout of the fieldwork is presented in Figure 1. For this design, the response deviation of unit k, enumerated by the interviewer j and controller i is calculated by

$$d_{ijk} = \tau_i + \beta_j + \gamma_{k(j)} + (\tau\beta)_{ij} + (\tau\gamma)_{ik(j)}. \tag{18}$$

Allocation Scheme-3: Allocation by split plot design. An experiment where the first factor is applied to whole (main) plots and each whole plot is divided into subplots (or split-plots), to which the second factor is applied, is called an SPD. A response error model involving three factors namely, controller, interviewer and respondent, can be stated as follows

$$y_{ijk} = \mu + \tau_i + \beta_j + \gamma_k + (\tau\beta)_{ij} + (\tau\gamma)_{ik} + (\beta\gamma)_{jk} + (\tau\beta\gamma)_{ijk} + \varepsilon_{ijk}, \begin{cases} i = 1,2, \dots, a \\ j = 1,2, \dots, b \\ k = 1,2, \dots, c \end{cases} \tag{19}$$

where μ represents the true value, τ_i is the i^{th} controller error, β_j is the j^{th} interviewer error, γ_k is the k^{th} respondent error, $(\tau\beta)_{ij}$ is the interaction error between the i^{th} controller and the j^{th} interviewer, $(\tau\gamma)_{ik}$ is

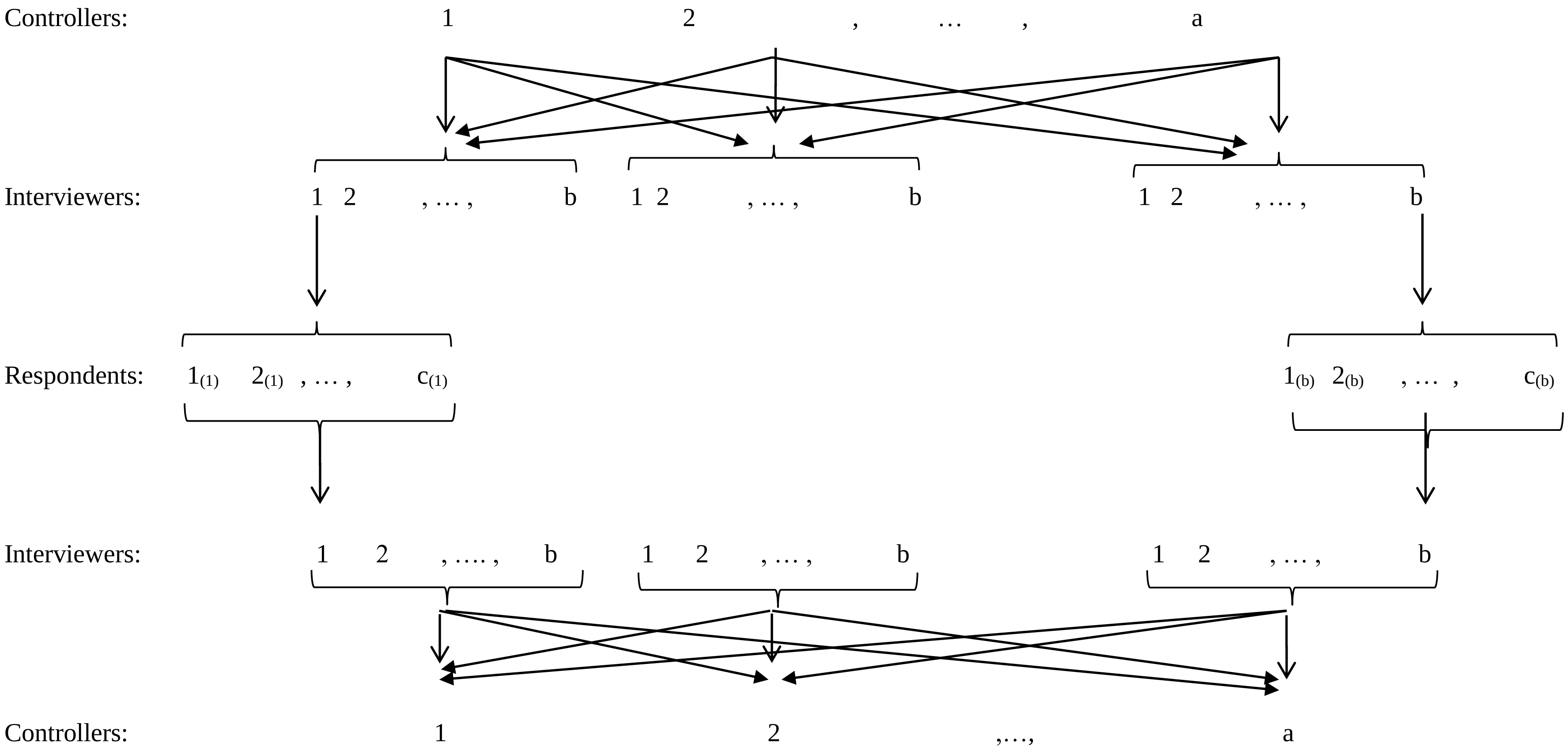

Figure 1. General nested and factorial factors layout of the fieldwork: respondents are nested under interviewers

the interaction error between the i^{th} controller and the k^{th} respondent, $(\beta\gamma)_{jk}$ is the interaction error between the j^{th} interviewer and the k^{th} respondent, $(\tau\beta\gamma)_{ijk}$ is the interaction error between the i^{th} controller, the j^{th} interviewer, and the k^{th} respondent, and ε_{ijk} is the NID$(0,\sigma^2)$ distributed random error term. The layout related with the fieldwork is presented in Figure 2. For this design, the response deviation of unit *k*, enumerated by the interviewer *j* and controller *i* is calculated by

$$d_{ijk} = \tau_i + \beta_j + \gamma_k + (\tau\beta)_{ij} + (\tau\gamma)_{ik} + (\beta\gamma)_{jk} + (\tau\beta\gamma)_{ijk}. \quad (20)$$

While implementing all schemes, we assume that there are *A* domains, and *n* (=*bc*) units are drawn without replacement from each domain, where *b* is the number of interviewers and *c* is the size of each interviewer's assignments. Also, we assign each controller to one of the *A* domains. Moreover, in *Allocation Scheme-2* and *Scheme-3*, similar to *Allocation Scheme-1*, to calculate the mean and total, sampling and the uncorrelated response variances, equations from (2) to (4) can be used. Besides, for the correlation coefficients, the correlated interviewer, controller variances and estimators of error components, equations from (5) to (16) can be utilized.

4. Concluding Remarks

In this study, we aim to investigate response errors, a kind of nonsampling errors, under complex interview-reinterview settings. Assuming that interviewers have influence on the respondents, three methodologies are proposed for developing response error models that utilize various DOE's for allocation of interviewers and respondents. They present experimental design plans in fieldworks for allocating interviewer and respondents, and also the related variance component estimators. Note here that the implementations of *Allocation Scheme-1* and *Scheme-2* seem to be feasible although the implement of *Allocation Sheme-3* does not seem to be so since it requires too many visits during the survey research.

References

Alimohammadi, R. & Navvabpour, H. (2008). Response error modeling in face-to-face surveys. *Research Journal of Science*, 33, 1-14.

Ayhan, H.Ö. (2003). Models of response error components in supervised interview–reinterview surveys. Journal of Applied Statistics, 30(9), 1047-1054.

Ayhan, H.Ö. (2012). Response reliability and response error estimation in personal interview surveys. AISC International Conference on Advances in Interdisciplinary Statistics and Combinatorics, Greensboro, NC, October 5-7.

Bailar, B. & Biemer, P. (1984). Some methods for evaluating nonsampling error in household censuses and surveys. In: P.S.R.S Rao & J. Sedransk (Eds.), W.G. Cochran's impacton statistics. New York: John Wiley & Sons. 253–274.

Biemer, P.P. & S.L. Stokes (1991). Approaches to the modeling of measurement errors. In: Biemer et al. (Eds.), Measurement errors in surveys. New York: John Wiley & Sons. 487-516.

Fabbris, L. (1991). The pairing of error sources and measures of estimates variance. *SIS Bulletin*, 22, 19-54.

Fellegi, I.P. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, 69, 19-54.

Hansen, M.H., Hurwitz, W.N., & Bershad, M.A. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 38(2), 359-374.

Kish, L. (1995). Survey sampling. New York: John Wiley & Sons.

Kish,L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.

Lessler, J.T. & Kalsbeek, W.D. (1992). Nonsampling error in surveys. New York: Wiley.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian statistical institute. *Journal of the Royal Statistical Society*, 109, 325-378.

Mahmut-Fahmi, F. (2013). Estimation of response errors in complex sample surveys. Ph.D. thesis, Middle East Technical University, Ankara, Tukey.

Montgomery, D. C. (2012). Design and analysis of experiments. New York: John Wiley & Sons.

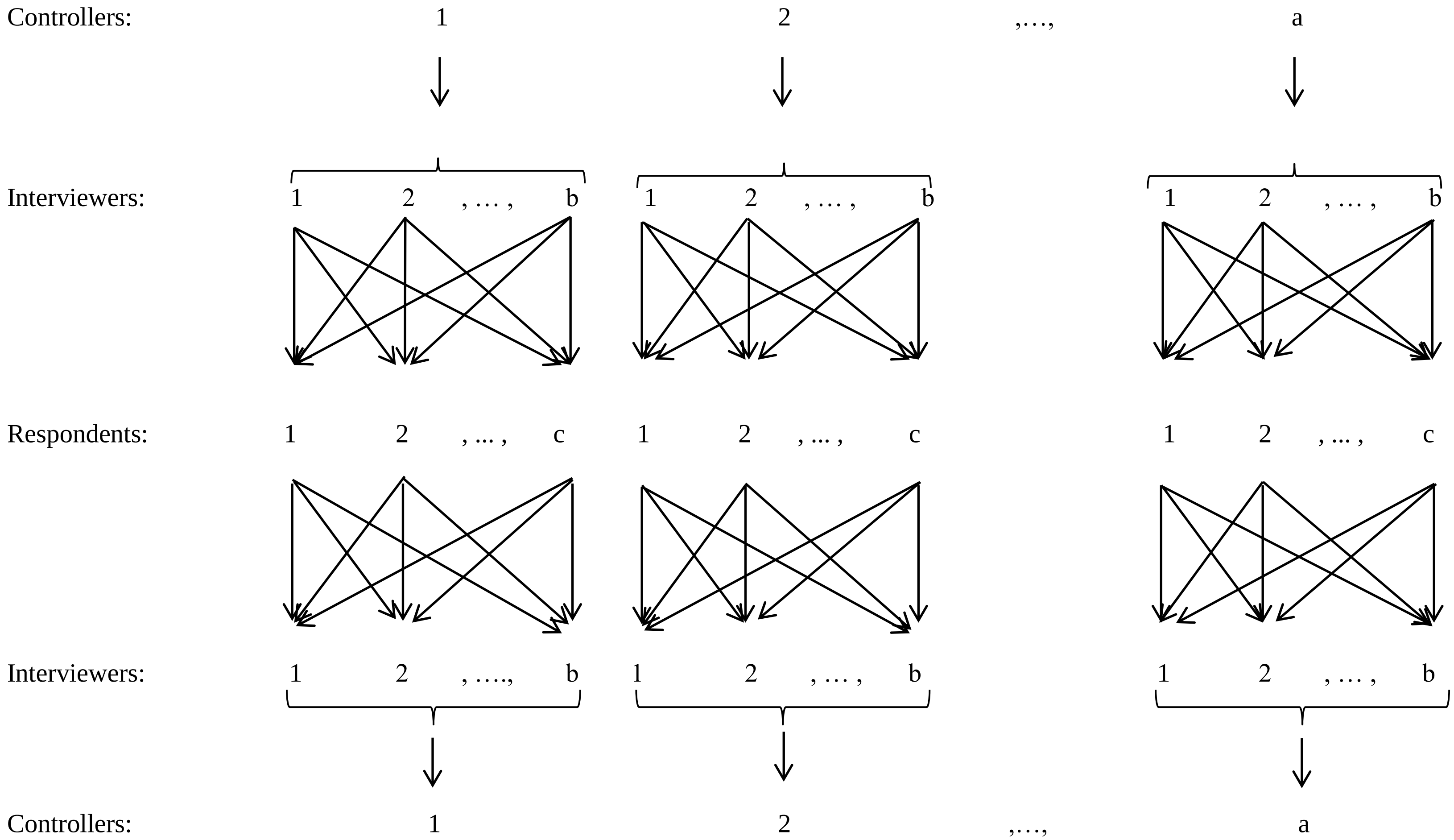

Figure 2. Split plot layout of the fieldwork