Fahmi, Fidan Mahmut; Ayhan, H. Öztaş; Batmaz, İnci

Article

Interviewer allocation through interview-reinterview nested design for response error estimation in sample surveys

Statistics in Transition New Series

Provided in Cooperation with:
Polish Statistical Association

Suggested Citation: Fahmi, Fidan Mahmut; Ayhan, H. Öztaş; Batmaz, İnci (2021) : Interviewer allocation through interview-reinterview nested design for response error estimation in sample surveys, Statistics in Transition New Series, ISSN 2450-0291, Exeley, New York, Vol. 22, Iss. 2, pp. 67-93,
https://doi.org/10.21307/stattrans-2021-017

This Version is available at:
https://hdl.handle.net/10419/236829

STATISTICS IN TRANSITION new series, June 2021
Vol. 22, No. 2, pp. 67–93, DOI 10.21307/stattrans-2021-017
Received – 29.06.2020; accepted – 08.01.2021

Interviewer allocation through interview–reinterview nested design for response error estimation in sample surveys

Fidan Mahmut Fahmi[1], **H. Öztaş Ayhan**[2], **İnci Batmaz**[3]

ABSTRACT

In surveys, non-sampling errors, due to their complex nature, are more challenging to quantify compared to sampling errors. Avoiding the release of these errors, however, results in biased survey estimates. In our previous paper, we devised the best interviewer allocation technique by using a nested experimental design to study response error estimation. In this study, in order to illustrate the effectiveness of this methodology in a different context, we apply it in interview-reinterview surveys relating to the time use and life satisfaction of academicians at Middle East Technical University, Turkey. An analysis of the pilot survey data showed that only half of the data was reliable, while the other half revealed interviewer effects. Prior to the main survey, interviewers underwent training in the course of which particular emphasis was put on the above-mentioned questions. In effect, the previously observed response variances which accounted for the total variance and data unreliability, were reduced considerably, increasing the quality of the main survey.

Key words: correlated response error, interviewer allocation assignments, quality of survey research, reinterview procedure, sample survey design.

1. Introduction

Research in the survey area mostly concerns the errors involved during the survey (Biemer *et al.* 2004); while some are dealing with the ways of eliminating errors, others try to measure the effect of them on the results by estimating the components of Total Survey Errors (TSE) involving both sampling and nonsampling errors (Kalton 1983, Salant and Dillman 1994). There are many different types of nonsampling errors. McNabb (2014) has defined nonsampling errors covering; frame error, measurement error, response error, interviewer error, and nonresponse error. McNabb (2014) also defines response error as basically respondent error. Measurement errors occur when

[1] Graduate student of METU, Ankara, Turkey. E-mail: fidantelaferli@yahoo.com.
[2] Professor Emeritus, Department of Statistics, METU, Ankara, Turkey. Corresponding author. E-mail: oayhan@metu.edu.tr. ORCID: https://orcid.org/0000--0003-3818-483X.
[3] Professor, Department of Statistics, METU, Ankara, Turkey. E-mail: ibatmaz@metu.edu.tr. ORCID: https://orcid.org/0000-0002-0821-3786.

the observed value differs from the true value according to the definition of the variable (Biemer and Lyberg, 2003). In several studies, response errors and measurement errors are used interchangeably (see; Hansen *et al.* 1951 and 1961).

The interviewer error, which may cause correlated errors in surveys, is one of the nonsampling errors. Since there exist only a few methods for compensating correlated errors caused by nonsampling errors, the current practice is to prevent the occurrence of them in data analysis (Biemer 2012). Due to the complex nature of the nonsampling error components, they are not usually examined in every survey report causing considerable bias in the survey estimates. Naturally, this study aims to highlight the importance of releasing such results, which covers a higher portion of the TSE.

Nonsampling error research was originally initiated by the U.S. Census Bureau Methodology Section around the middle of the last century. An early study of the methodology of response errors in surveys was published by Hansen, Hurwitz, Marks and Mauldin (1951). U.S. Bureau of the Census' survey model was described in Hansen, Hurwitz and Bershad (1961) and also in Hansen, Hurwitz and Pritzker (1964). Estimating the response variance components of the U.S. Bureau of the Census' Survey model was also evaluated by Bailar and Dalenius (1969). Later studies from the former Bureau researchers were done by Biemer and Stokes (1985, 2004) on the modelling of measurement error.

On the European side, response error related research was initiated by World Fertility Survey (WFS) Methodology Team, who worked on the single round high quality data collection. By this team, the methodology of the response errors was evaluated by O'Muircheartaigh and Marckwardt (1980) for the assessment of the reliability of WFS data. Sampling issues for national fertility surveys were also covered by Verma (1980). Computation methodology of response errors was proposed by O'Muircheartaigh (1984) for Peru Fertility Survey. Response errors for attitudinal surveys was also formulated by O' Muircheartaigh (1976, 1977). Later, simple response variance estimation methodology was overviewed by O' Muircheartaigh (2004). An excellent overview of the response error methodology is also covered by Moser and Kalton (1979). Estimation in the presence of measurement error is also evaluated by Fuller (1995).

Reinterview methodology was developed for the first time by the U.S. Census Bureau Methodology Division during 1950's (Hansen et al. (1951, 1961). Since then, several other researchers have improved on the methodology (Bailar and Dalenius 1969, Biemer and Stokes (1985, 1991). Identical approaches have also been taken for the WFS Methodology Division by O'Muircheartaigh (1977, 1984). Using design of experiments (DOEs) in interviewer allocation is relatively new in the survey research. O'Muircheartaigh (1984) has applied the methodology of interviewer-reinterviewer allocation to Peru Fertility Survey using a DOE, called Latin Square Design.

Estimators of nonsampling errors in interview–reinterview supervised surveys with interpenetrated assignments were proposed by Bassi and Fabbris (1997). Especially, O'Muircheartaigh (1977, 1984), Bassi and Fabbris (1997), Ayhan (2003) and Fahmi (2013) have also highlighted the use of supervised interview-reinterview design in their methodological research. It has been widely used since then in a variety of surveys (Biemer *et al.* 2004). It is an important tool to evaluate field work and to estimate and reduce the error components in a survey.

Interview–reinterview designs are necessary to estimate response variance. Hox et al. (2004) stated that "Several authors have criticized the existing studies on interviewer effects (Hagenaars and Heinen, 1982). A central criticism concerns the adequacy of the statistical models used. The structure of the data to be analyzed is hierarchical, since respondents are nested within interviewers." Lyberg and Kasprzyk (2004) also stated that "Interviewer errors and interviewer variability can be measured in various ways. Basically, different systems for reinterviews (replication; McCarthy, 1966) and interpenetration (Mahalanobis, 1946) are used." Based on these, is it possible to say that the researchers here use a reinterview design to estimate interviewer effects, because such a design eliminates the nested structure where one respondent is interviewed by one interviewer only. The rationale for the current study is the nested design of respondents within interviewers.

In this article, one kind of nonsampling error, called response error, is investigated in a personal interview in sample surveys, where multiple stages of sampling are employed. Response errors may occur because of the respondent error, interviewer error, or their interactions causing correlated response error. Under this assumption, in this article, a nested experimental design (ND) is utilized for developing response error models and obtaining efficient estimators for response such as simple and correlated response variance in interview-reinterview surveys as suggested by Ayhan (2003, 2012).

Ayhan (2003) used ND to make interviewer allocation for the interview-reinterview process. Ayhan (2012) also mentioned that experimental settings of the interviewer allocation can be based on the Nested and Factorial Design (NFD), or the Split Plot Design (SPD). Then, Fahmi (2013) investigated these designs for interviewer allocation in personal interview surveys. Next, Batmaz and Fahmi (2015) established theoretical backgrounds of the simple and correlated response error estimation procedure.

In general, to make the analysis simple, the expected value of the interviewer effect is assumed to be zero although it is not in reality. To measure it, the survey is designed to have different interviewers for the respondents within the main and reinterview survey. The advantage of the ND allocation is that it provides different respondents for the same interviewer in both the pilot and main survey, enabling to compute the response variance independently for each survey. It also provides flexibility in the field

allocation and application. Note here that there are some important issues to be dealt with in designing reinterviews such as selection of sample, reinterviewer, respondent, mode of interview as well as designing of reinterview questionnaire (Biemer and Stokes 2004).

Purposes of reinterview were extensively covered by Forsman and Schreiner (2004) where they have proposed "purposes of interview–reinterview designs", which are classified as:

(1) *To evaluate fieldwork*: (a) reinterview is used to identify interviewers who are falsifying data, and (b) reinterview is also used for misunderstood procedures and require remedial training.
(2) *To estimate error components in a survey model*: (a) reinterview is also used to estimate simple response variance, and (b) reinterview is also used to estimate response bias.

In the current study, reinterview is used to estimate simple response variance (Design 2a). In addition, this study also estimates correlated response variance and interviewer variance.

Forsman and Schreiner (2004) further cover other reinterview design issues, evaluating interviewer performance, model-based analysis of reinterview data, and the use of computer assisted interviewing. Correlates of reinterview response inconsistency are also examined by O'Muircheartaigh (1986) in the Current Population Survey.

A very recently edited book by Olson *et al.* (2020 a) covers interviewer effects from a total survey error perspective. An overview of research on interviewer effects is covered by Olson *et al.* (2020 b) within the book. They state that the errors introduced by interviewers can take the form of bias or variance. Early research also found that interviewers vary in how they administer survey questions and their effects were similar to sample clusters in both face-to-face (Hansen *et al.* 1961, Kish 1962) and telephone surveys (Groves and Magilavy 1986). In particular, similar to a design effect for cluster samples, interviewers increase the variance of an estimated mean as a function of their average workload and the intra-interviewer correlation.

Given the nesting of respondents within interviewers, following Kish's ANOVA-based model (Kish 1962), hierarchical or random effects models have long been used for the study of interviewer effects (Dijkstra 1983, Hox 1994, O'Muircheartaigh and Campanelli 1998).

Recently, Edwards *et al.* (2020) studied behaviour change techniques for reducing interviewer contributions to total survey error. Modelling interviewer effects in the National Health Interview Survey is also investigated by Dahlhamer *et al.* (2020). On the other hand, West (2020) designed studies for comparing interviewer variance in two groups of survey interviewers.

In this study, the methodology previously developed by Ayhan (2003, 2012), Fahmi (2013) and Batmaz and Fahmi (2015) is applied to an interview-reinterview survey for inquiring about the time use and life satisfaction of academicians working at the Middle East Technical University (METU), Turkey. This study contributes to the literature

in various aspects. Keeping in mind that nonsampling errors are equally important as sampling errors, it guides the survey researchers to measure the response errors, particularly interviewer effects on the responses as well as to measure the total, sampling, response and the correlated interviewer variances, efficiently. Thus, an important concern regarding the estimation of nonsampling errors is overcome. However, nonsampling errors are neglected in most surveys, due to "hardness in quantifying" as well as "additional data coasting," and consequently, error based information on most of the TSE components cannot be obtained completely. In the case the researchers want to compute and report these errors along with the survey results, it definitely will increase the validity and reliability, and hence, quality of surveys. This way, usefulness and limitations of surveys conducted will be appreciated better. Moreover, these tools will provide feedbacks regarding errors involved, particularly in surveys conducted periodically. By evaluating the experiences gained, quality of the survey can be continuously improved.

This article is organized as follows: the response reliability measures are defined in Section 2. In Section 3, the methodology for interviewer allocation by ND and its application are presented. In Section 4, findings of the applications are presented and the work is concluded in Section 5.

2. Measures of response reliability

We know that for each individual covered by the survey, there is an individual true value. The difference between an individual true value and the value recorded on the schedule is the individual response error. There is always a possibility that true values may change. To determine the optimum period between interview and reinterview, we followed the guide suggested by Biemer and Stokes (2004), and it is mentioned clearly in Section 3.2 Pilot Survey Application. We hope only few such changes in true values left, and they are represented by the residual error in the model.

In investigating the reliability of data, we can focus on two different but related aspects of the data: bias and variance. For each individual j, we have for each variable y, the results of two separate observations, y_{j1} and y_{j2}. In this case, they are assumed to be obtained from an interview-reinterview survey, respectively. The differences within the pairs of observations provide the raw material for the reliability investigation. Measures of reliability used depend on the types of data.

Measures of response reliability and response error estimation are proposed by Hansen et al. (1961). The methodology was also extended by O'Muircheartaigh (1977, 1984) and O'Muircheartaigh and Marckwardt (1980). They have covered the methodology for several data measurement scales, which can be categorized as in the following sections.

2.1. For categorical data

In comparing the responses obtained for a particular variable, the data may be represented by the square matrix $\{n_{ij}\}$, where n_{ij} is the number of elements classified in category i according to the first interview, and in category j according to the second interview, i.e. reinterview. The diagonal of this square matrix, with entries, n_{ii}, contains the cases of exact agreement. The simplest measure of reliability (bivariate agreement) is the *index of crude agreement* (or *crude index*), which can be written as

$$A = \frac{1}{n}\sum_i n_{ii}. \tag{1}$$

It represents the proportion of correctly classified units. Another simpler one is the *index of crude disagreement*

$$D = 1 - A. \tag{2}$$

It represents the proportion of incorrectly classified units. Here, values of A and D close to one (1) and zero (0), respectively, indicate good agreement.

However, the *crude index A* in formulae (1) has a fairly serious drawback; it does not take into account the fact that some agreements will occur by chance even if the measurement is completely unreliable (random). To overcome the problem, Cohen (1960) define an *index of consistency*, called *kappa*, of the following form:

$$\mathrm{K} = 1 - \frac{1-Po}{1-Pe} = \frac{Po-Pe}{1-Pe}, \tag{3}$$

where $P_o = \sum_{i=1}^{L}\left(\frac{n_{ii}}{n}\right)$ and $P_e = \sum_{i=1}^{L}\left(\frac{n_{i.}}{n}\right)\left(\frac{n_{.i}}{n}\right)$. Here, P_o is the sum of the observed proportions reflecting agreement, and P_e is the sum of the expected proportions reflecting agreement. Under the assumption of independence between the two observations, formulae (3) can be written as

$$\widehat{K} = 1 - \frac{\sum_{i\neq j} n_{ij}}{\frac{1}{n}\sum_{i\neq j} n_{i.}n_{.j}} = \frac{\sum_{i=1}^{L}\left(\frac{n_{ii}}{n}\right) - \left(\frac{n_{i.}}{n}\right)\left(\frac{n_{.i}}{n}\right)}{1-\sum_{i=1}^{L}\left(\frac{n_{i.}}{n}\right)\left(\frac{n_{.i}}{n}\right)}, \tag{4}$$

where L represents the number of categories. For evaluating the magnitude of *kappa*, *K*, the standards proposed by Landis and Koch (1977) are utilized. Note also that *CI*'s non containing 0 (zero) indicate significant *kappa* values.

2.2. For ordinal data

When the scales are ordinal, interval or ratio, any measure of agreement should take into account the degree of disagreement, which is a function of the difference between scale values. Accordingly, formula (1) is modified by redefining *agreement* to mean that the two interviews obtain values within some acceptable distance (k units) of each other. Then, agreement can be written as

$$A_k = \sum_{|i-j|\leq k} n_{ij} = 1 - D_k. \tag{5}$$

Cohen (1968) introduce a modified form of *K*, which allows for the scaled disagreement or partial credit in terms of weights $\{W_{ij}\}$, which reflect the contribution of each cell in the table to the degree of disagreement.

$$K_W = \frac{Po^* - Pe^*}{1 - Pe^*} = 1 - \frac{\sum_{i \neq j} w_{ij} n_{ij}}{\frac{1}{n}\sum_{i \neq j} w_{ij} n_{i.} n_{.j}}. \tag{6}$$

Here, any monotonically decreasing function of the differences between the values *i* and *j* can be used as weights. Cicchetti (1972) suggests the use of the following weights for the ordinal and metric data, respectively.

$$w_{ij} = 1 - |i - j|/(L - 1); \;\; w_{ij} = 1 - (i - j)^2. \tag{7}$$

2.3. For interval and ratio scale data

For metric measurements, Hansen *et al.* (1961) proposed the basic mathematical or response error given below. The same methodology was reformulated by O'Muircheartaigh (1977) and O'Muircheartaigh and Marckwardt (1980). For simplicity, the discussion is restricted to the estimation of the population mean, where

$$\mu = \frac{1}{N}\sum_{j=1}^{N} \mu_j. \tag{8}$$

Assume that an observation for the *j*th element in the survey for trial *t* is denoted by y_{jt}. An estimator of μ obtained from a survey (one trial) is

$$\bar{y}_t = \frac{1}{n}\sum_{j=1}^{n} y_{jt}. \tag{9}$$

Here, the population consists of N individuals from which a sample of size *n* is sampled.

The total variance of the survey estimator is

$$\sigma_t^2 = E(\bar{y}_t - \bar{Y})^2, \tag{10}$$

where $\bar{Y} = E(\bar{y}_t)$, and the mean square error (MSE) of the estimator becomes

$$MSE = \; E(\bar{y}_t - \mu)^2 = \sigma_t^2 + \beta^2. \tag{11}$$

The expected value over all possible trials for the element *j* is

$$E\left(y_{jt}|j\right) = Y_j. \tag{12}$$

The difference between the observation on the *j*th unit of a particular survey (say trial *t*) and the expected value is

$$d_{jt} = y_{jt} - Y_j \tag{13}$$

2.3.1. Simple response variance

This is the *response deviation*, which is measured from the expected value. For the estimator obtained from the survey, the total variance can be partitioned as follows:

$$\sigma_t^2 = \; E(\bar{y}_t - \bar{Y})^2$$
$$= E(\bar{y}_t - \bar{y})^2 + 2E(\bar{y}_t - \bar{y})(\bar{y} - \bar{Y}) + E(\bar{y} - \bar{Y})^2, \tag{14}$$

where $\bar{y}_t = \frac{1}{n}\sum_{j=1}^{n} y_{jt}$ and $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$.

The first term in equation (14) is the response variance, $\sigma^2_{\bar{d}_t}$; the second term involves the covariance between $\bar{d}_t$ and $\bar{y}$, and the third term is the sampling variance, $\sigma^2_{\bar{y}}$. The response variance can be restated as

$$\sigma^2_{\bar{d}_t} = E\left(\bar{d}_t^2\right) = \frac{\sigma_d^2}{n}[1 + (n-1)\rho], \tag{15}$$

where σ_d^2 is simple response variance and ρ is the interclass correlation coefficient among the response deviations within a trial. Fellegi (1964), permits in principle the estimation of a number of components of the correlated response variance, $(\sigma_d^2(n-1)\rho)/n$. The sample variance can be written as

$$\sigma^2_{\bar{y}} = E(\bar{y} - \bar{Y})^2 = \sigma_s^2[1 + \rho(n-1)], \tag{16}$$

where σ_s^2 is the population variance, and ρ is the intracluster correlation coefficient.

Then, index of inconsistency, IOI, is defined to be

$$IOI = \frac{\sigma_d^2}{\sigma_s^2 + \sigma_d^2}, \tag{17}$$

which measures the proportion of the total element variance due to the response variability. To measure how much the data are reliable, another statistic, called *reliability of data*,

$$r = 1 - IOI \tag{18}$$

has been proposed by Yu et al. (2000). Note that values of *IOI* and *r* close to zero (0) and one (1), respectively, indicate that data are consistent and reliable.

2.3.2. Correlated response variance

The analysis of response deviations presented above treats them as uncorrelated. The basic model of the response process for individual *j* is

$$y_{jt} = \mu_j + \beta_j + d_{jt}, \tag{19}$$

where y_{jt} is the response obtained from the individual *j* on the occasion *t*; μ_j is the true value for the individual; β_j is the individual response bias and d_{jt} is the response deviation. Since we cannot, unless we have external validating information for the individual, estimate β_j, we rewrite equation (13) as

$$y_{jt} = Y_j + d_{jt}, \tag{20}$$

where Y_j is the expected value of the observation for individual *j* over a large number of trials under the same essential survey conditions. However, if the interviewers cause a systematic distortion of the responses, we can write

$$y_{ijt} = Y_j + \alpha_i + \varepsilon_{ijt},$$

where the subscript *i* is added to denote the interviewer and split the response deviation d_{jt} into two additive components α_i and ε_{ijt}. Here, the α_i represents the net systematic

effect of interviewer *i* on the responses; it is the net bias introduced by the interviewer *i*. The ε_{ijt} is the residual response deviation, which is assumed to be unrelated to the interviewer. For making the relations simpler among the interviewers as a whole, it is assumed that the expected value of the interviewer error is taken as zero [$E(\alpha_i) = 0$]. This assumption naturally makes life easy, instead of computing interviewer error. Note also that that the interviewer effect can also be measured by the correlation between the responses for the first and the second interview as follows:

$$\alpha = \text{corr}(y_{j1}, y_{j2}). \tag{21}$$

Values of α close to 1 (one) indicate no interviewer effect at all.

Making the usual assumptions about the variances and covariances, we can write the variance of a single observation y_{ijt} as

$$\text{Var}\,(y_{ijt}) = \sigma_S^2 + \sigma_\alpha^2 + \sigma_\varepsilon^2. \tag{22}$$

If $\bar{y}_t$ is the sample mean for the survey, then its variance is

$$\text{Var}\,(\bar{y}_t) = \frac{\sigma_S^2}{n} + \frac{\sigma_\alpha^2}{k} + \frac{\sigma_\varepsilon^2}{n}, \tag{23}$$

where *k* is the number of interviewers. And it is

$$= \frac{\sigma_S^2}{n} + \frac{\sigma_d^2}{n}(1 + \rho(m-1)), \tag{24}$$

where σ_S^2 is the population variance of the $\{y_j\}$; $\sigma_d^2 (= \sigma_\alpha^2 + \sigma_\varepsilon^2)$ is the simple response variance; m is the average workload per interviewer; and ρ is the intra-interviewer correlation coefficient $(= \sigma_\alpha^2/\sigma_d^2)$.

The usual estimate of the survey variance will include both σ_S^2 and σ_d^2, and if the sample is a simple random sample, it will be $(\sigma_S^2 + \sigma_d^2)/n$. Thus, the survey variance will be underestimated by an amount equal to

$$\frac{\sigma_d^2}{n}\rho(m-1). \tag{25}$$

2.3.3. Simple response variance, correlated response variance, and interviewer error

Computations of response variance and correlated interviewer variance are based on the following estimators. Let us denote μ and σ_t^2, the mean and the variance of y, respectively. Then, the total variance is the combination of sampling and response variances.

$$\mu = \frac{\sum_i^a \sum_j^b \mu_{ij}}{ab}, \tag{26}$$

$$\sigma_t^2 = E\{\frac{1}{n}\sum_i^a \sum_j^b (y_{ij} - \mu)^2 = \sigma_s^2 + \sigma_r^2, \tag{27}$$

where σ_s^2 is the sampling variance and σ_r^2 is the response variance. The sampling variance is defined (Fellegi 1964; Bassi and Fabbris 1997) as

$$\sigma_s^2 = E\left(\frac{1}{ab}\sum_i^a \sum_j^b c_{ij}^2\right), \tag{28}$$

where $c_{ij} = \mu_{ij} - \mu_i$ is the sample deviation of the same unit. Also, the response variance is defined as

$$\sigma_r^2 = E\left(\frac{1}{ab}\sum_i^a \sum_j^b d_{ij}^2\right), \tag{29}$$

where d_{ij} is the *response deviation* of unit *j*, enumerated by interviewer *i*. Here, the response deviation is determined as $d_{ij} = \tau_i + \beta_{j(i)}$. Values of *d* close to zero (0) and/or *CI for d* containing zero (0), indicate no significant deviation of the response between two interviews. Besides, δ_2 is the correlation coefficient between the response deviations within interviewer's assignments, and it is defined as

$$\delta_2 = \frac{1}{\sigma_r^2} E\left\{\frac{1}{n(b-1)} \sum_i^a \sum_{j=j'}^b d_{ij} d_{ij'}\right\}, \tag{30}$$

and $\delta_2 \sigma_r^2$ is the correlated interviewer variance.

The response error analysis is conducted and response reliability measures are calculated for each question listed in Table 2 by using the formulas shown in Section 2. Particularly, the response reliability statistics, A, D, A_k, IOI, r, K, K_w are calculated by using the formulae (1), (2), (5), (17), (18), (4), and (6), respectively. In addition, the interviewer effect, total, sampling, response and correlated interviewer variances are obtained by using the formulae (21), (27), (28), (29), and (29) and (30), respectively. Note here that to calculate the response reliability statistics for the Likert scale type questions, the agreement proportions between two interviews are obtain first (see Fahmi, 2013).

3. Application of the methodology

The main purpose of this study is to provide an insightful application whose results shed new light on the success of the methodology already developed by Ayhan (2003 & 2012), Fahmi (2013), and Batmaz and Fahmi (2015). The use of DOE technique for allocating interviewers provides a novel approach for estimating response error variance. In order to apply the methodology, an interview-reinterview survey is designed and conducted at Middle East Technical University (METU), Ankara, Turkey, inquiring about time use and life satisfaction of its academicians.

3.1. Interviewer allocation by a nested design

The experiments involved two or more random factors and the levels of at least one factor are similar but not identical for different levels of another factor is generally designed as nested experiments, and are commonly used to determine the sources of variation in the system (Box *et al.* 2005, Montgomery 2012). To illustrate this, suppose that the levels of a factor (e.g. B) are similar but not identical for the levels of another factor (e.g. A). Such an arrangement is called an ND with the levels of factor B nested

under the levels of factor A. A linear statistical model for analyzing such an experiment, a two-stage ND, is written as

$$y_{ijk} = \mu + \tau_i + \beta_{j(i)} + \varepsilon_{(ij)k} \begin{cases} i = 1,2,\dots,a \\ j = 1,2,\dots,b \\ k = 1,2,\dots,n \end{cases}. \quad (31)$$

There are a levels of factor A, b levels of factor B nested under each level of A. The subscript $j(i)$ indicates that the jth level of factor B is nested under the ith level of factor A. The replicates, if exist, are assumed to be nested within the combination of levels of A and B; so that the subscript $(ij)k$ is used for the white noise error term. In addition, this is a balanced ND because there are an equal number of levels of B within each level of A and an equal number of replicates. Because not every level of factor B appears with every level of factor A, there can be no interaction between A and B. In our particular case, this implies that respondents in different domains can only be visited by different interviewers, hence, data collected by ND can only be analyzed under the assumption that there is no interaction between interviewer and respondent factors. To measure this interaction, factorial designs can be used. However, in such allocations, the number of interviewers to be allocated for the field application may be combinatorically problematic.

Although ND does not allow measuring the interaction between the interviewer and respondent factors, it provides flexibility in allocating interviewers to respondents. Therefore, when compared with the factorial design, ND is much more time and cost efficient. Moreover, due to the fact that the factors involved are assumed to be random here, ND naturally provides estimates for the variance components, which are sample, interviewer and response variance in this case.

3.2. Pilot survey application

As the first step, a pilot survey is applied to a METU department. The main purpose of conducting pilot surveys is diverse; it includes pretesting the questionnaires, estimating the duration of interview, and planning the timing of reinterviews. In addition, data obtained from pilot studies (Fahmi 2013) are also analyzed to get feedback on the applicability of the methodology considered. Here, we present the pilot survey, in which interviewer allocation is done by using ND in a 10 question life satisfaction survey for the academicians in two rounds (i.e. pilot parent survey interview and reinterview).

Twenty-two academicians are involved as respondents in this survey; 10 of them are faculty members and 12 of them are research assistants. They are randomly clustered into four domains, where two contain five and other two contain six respondents. The interviewers selected randomly from the graduate students of the department are randomly assigned to one of these domains. As a result, an unbalanced

ND design is formed. The nested layout of the pilot fieldwork interview is shown in Figure 1. Also, the fieldwork allocation of the interviewers to respondent groups for the pilot survey is given in Table 1. Here, all domains are assigned to one supervisor, who controls the completion errors within the completed questionnaire, in the field, after its field data collection. Interviewers match with the respondents according to the preplanned survey design and interview allocation schemes. An interviewer is not allocated to the same respondent in the interview and the reinterview. Thus, we do not have replications in ND in our case (i.e., k=1). Note that the interviewers have training for sample respondent selection and questionnaire execution for few days. In case of nonresponse, a new respondent is determined by random substitution. The same approach is also used during the reinterview. Timing of interview, reinterview, and reconciliation survey was proposed by the World Fertility Survey Methodology Division for their 42 country surveys (WFS, 1977).

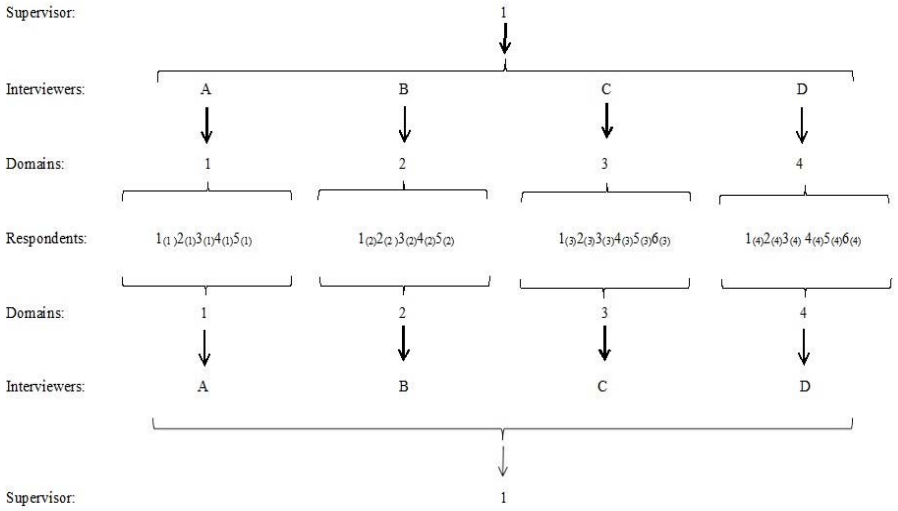

Figure 1. Pilot parent survey layout of the fieldwork

Table 1. Fieldwork allocation of the interviewers to respondent groups for the pilot survey

Domain Number	Respondent Number	Interview	Reinterview
1	$1_{(1)};2_{(1)};3_{(1)};4_{(1)};5_{(1)}$	Interviewer A	Interviewer B
2	$1_{(2)};2_{(2)};3_{(2)};4_{(2)};5_{(2)}$	Interviewer B	Interviewer A
3	$1_{(3)};2_{(3)};3_{(3)};4_{(3)};5_{(3)};6_{(3)}$	Interviewer C	Interviewer D
4	$1_{(4)};2_{(4)};3_{(4)};4_{(4)};5_{(4)};6_{(4)}$	Interviewer D	Interviewer C

By following the suggestions (Biemer and Stokes, 2004), after an interval of one month, the second round of the survey, the reinterview, is applied to the same respondents by exchanging the interviewers' domains using the same questionnaire paper (see Figure 2 and Table 1). Since the purpose of the reinterview is to estimate response variance and response bias, the original questions are repeated in their exact forms as suggested by Kish (1965). Note also that the execution of the questionnaires in both interview and reinterview is made on a voluntary basis to the respondent.

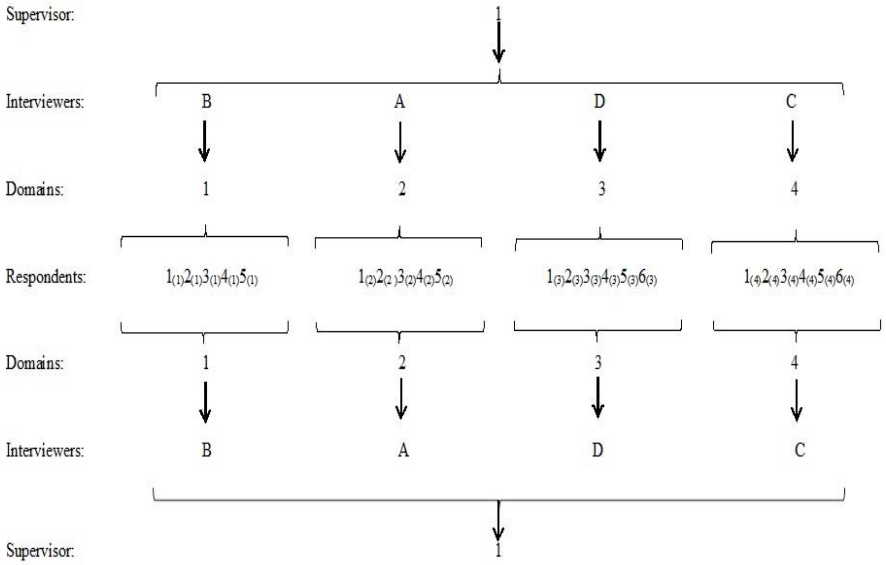

Figure 2. Pilot reinterview survey layout of the fieldwork

As a field operation, reinterviews are expensive, in face to face surveys. Because of its complex methodology, some survey designers would like to neglect this operation. On the other hand, nonsampling errors cover a larger amount of the total error, when compared with sampling errors. One should make a decision on the error versus cost of the survey operation. The reinterview is always conducted on a subsample of the original survey sample, the costs can be moderate. However, with the use of computer assisted interviewing, operational costs can be kept minimal while the usefulness of the reinterview is increased. The survey contained 10 basic questions, which are designed to cover a different range of data measurement levels such as dichotomy, polytomy, ordinal, interval (see Table 2).

The random effects model used for this survey is developed from model (31), and written as

$$y_{ij} = \mu + \tau_i + \beta_{j(i)} + \varepsilon_{(ij)} \begin{cases} i = 1,2,\dots,4 \\ j = 1,2,\dots,22 \end{cases}. \tag{32}$$

Here, μ represents the true value, τ_i the ith interviewer error, $\beta_{j(i)}$ is the jth respondent error nested under the ith interviewer, $\varepsilon_{(ij)}$ is the NID $(0,\sigma^2)$ random error term. Thus, in this design, we assume that there are four domains and from each domain a sample of size five, five, six and six respondents are drawn, respectively, without replacement.

This is an unbalanced design because the sizes of each interviewer's assignment are not the same. The response error analysis is conducted and response reliability measures are calculated for each question listed in Table 2 by using the formulas shown in Section 2, and the results obtained are presented in Table 3 and Table 4. Note here that to calculate the response reliability statistics for the Likert scale type questions, the agreement proportions between two interviews are obtain first (see Fahmi 2013).

Table 2. Pilot study questions and related information

Pilot Survey Question Number	Measurement scale	Variable name	Main Survey Question Number
1	Dichotomy	Gender of respondent	1
2	Interval	Age of respondent	2
3	Interval	Height of respondent	3
4	Likert scale	Last degree owned	5
5	Dichotomy	Title of respondent	6
6	Interval	Working duration in years in the university	8
7	Interval	Payment on clothing in TL* per month**	15
8	Interval	Payment on cultural activities in TL* per month**	16
9	Likert scale	Job satisfaction	17
10	Likert scale	Salary satisfaction	18

* Note that TL refers to Turkish Lira as currency; ** "per month" refers to any average month within the year.

Table 3. Response reliability statistics for the pilot survey

Ques. No.	A	D	$A_{k=1}$	$D_{k=1}$	IOI	IOI Eval.	r	K	K_w	K Eval.	CI for K
1	1.000	0.000	1.000	0.000	NA	NA	NA	1.000	1.000	AP	(1.00, 1.00)*
2	1.000	0.000	1.000	0.000	0.000	L	1.000	1.000	1.000	AP	(1.00, 1.00)*
3	1.000	0.000	1.000	0.000	0.000	L	1.000	1.000	1.000	AP	(1.00, 1.00)*
4	1.000	0.000	1.000	0.000	0.000	L	1.000	1.000	1.000	AP	(1.00, 1.00)*
5	1.000	0.000	1.000	0.000	NA	NA	NA	1.000	1.000	AP	(1.00, 1.00)*
6	1.000	0.000	1.000	0.000	0.001	L	0.999	1.000	1.000	AP	(1.00, 1.00)*
7	0.636	0.363	0.818	0.182	0.399	M	0.601	0.480	0.528	M	(0.20, 0.76)*
8	0.454	0.545	0.864	0.136	0.685	H	0.315	0.248	0.313	F	(-0.03, 0.53)
9	0.864	0.136	0.908	0.092	0.469	M	0.531	0.749	0.891	S	(0.53, 0.97)*
10	0.682	0.318	0.955	0.045	0.326	M	0.674	0.566	0.596	M	(0.32, 0.82)*

Notes: 1. NA indicates that this statistic is irrelevant for this type of variable; M: Moderate; L: Low; H: High; F: Fair: AP: Almost Perfect; S: Substantial 2. * denotes statistically significant kappa value for that particular type of question at α=0.05 level of significance by paired-t test. 3. Indices of *A*: crude agreement; *D*: crude disagreement; *IOI*: inconsistency; *r*: reliability of data; *K:* consistency (kappa) 4. Values of *A*, *r* and *K* close to one (1) indicate consistent and reliable data.

3.3. Main survey application

In the main survey, the questionnaires of the pilot survey are extended with some additional questions, and executed to the randomly selected faculty members of METU, Turkey, under a preplanned schema according to an ND, again in two rounds, namely, main and reinterview surveys. Details of both response error applications are presented below. Note that main survey data can be found in Fahmi (2013).

In this part of the study, following the methodology proposed by Ayhan (2003, 2012), Fahmi (2013), and Batmaz and Fahmi (2015), an ND is applied to allocate the interviewers to respondents in a life satisfaction and time use survey for METU academicians. The survey contains 20 questions, and it is applied to 168 academicians. They are randomly selected from METU's five faculties which have 839 academicians. The sample corresponds to 20% of the total number of the academicians working at METU. The number of faculty members and the corresponding sizes of the selected samples are given in Table 5. Note here that since this is an academic research, the sample size is limited to 168 academicians.

Table 4. Other response error statistics for the pilot survey

Ques. No.	$\bar{y}$	$\bar{y}_1$	$\bar{y}_2$	d= $\bar{y}_1 - \bar{y}_2$	CI for d	α	s_t^2	s_r^2	s_s^2	$\hat{\delta}_2 s_r^2$
1	NA	NA	NA	NA	NA	1.000	NA	NA	NA	NA
2	32.86	32.86	32.86	0	(0.00, 0.00)	1.000	0.00	0.00	0.00	0.00
3	170.56	170.59	170.55	0.04	(-0.24, 0.33)	0.998	0.00	0.00	0.00	0.00
4	2.18	2.18	2.18	0	(0.00, 0.00)	1.000	0.00	0.00	0.00	0.00
5	NA	NA	NA	NA	NA	1.000	NA	NA	NA	NA
6	7.4	7.30	7.60	-0.30	(-0.48, -0.10)*	0.998	53.13	0.09	53.04	0.11
7	142.70	133.86	151.59	-17.73	(-61.91, 26.46)	0.624	13.459	5.380	8.079	979
8	90.45	77.50	103.41	-25.91	(-47.99, 3.83)*	0.802	1.863	1.277	587	789
9	4.25	4.27	4.23	0.04	(-0.24, 0.34)	0.599	0.49	0.23	0.27	0.094
10	3.00	3.04	2.95	0.09	(-0.21, 0.39)	0.820	1.38	0.45	0.93	0.052

Notes: 1. NA indicates that this statistic is irrelevant for this type of variable, 2. * denotes statistically significant difference between the parent and reinterview values for that particular type of question at α=0.05 level of significance by paired-t test. 3. *d*: response deviance (error); *α*: interviewer effect; s_t^2: total variance; s_r^2: response variance; s_s^2: sampling variance $\hat{\delta}_2 s_r^2$: correlated interviewer variance 4. Values of *D*, *IOI*, and *d* close to zero (0) and also *CI for K* and for *d* containing zero indicate consistent and reliable data.

Table 5. Number of academicians at each METU faculty and the selected sample sizes

Faculties	Total number of academicians	Selected number of academicians
Architecture	52	10
Economic and Administrative Sciences	91	18
Education	80	16
Engineering	384	77
Arts and Sciences	232	47
Total	**839**	**168**

The respondents are divided into eight domains, and one interviewer is sent to each domain, randomly. As in the case of the pilot study, the numbers of each interviewer's assignments are not the same. Hence, an unbalanced ND design is formed again.

One month later, the second round of the main survey is applied to six respondents from each domain (See Table 5), which are again selected randomly by exchanging the interviewers' domains, and using the same original questionnaire (See Figure 4). The nested layout of the first round main survey fieldwork interview is shown in Figure 3.

In the main survey, the questionnaire is expanded to 20 questions by including 10 more questions related to academicians' time use in addition to the life satisfaction questions covered in the pilot study. Main survey questions and the related information is presented in Table 6. The number of respondents at each sample department in the main survey first round (interview) and the second round (reinterview) is given in Fahmi (2013). Note here that the same random effects model given in model (32), which is developed from model (31), is used.

4. Findings and discussion

In order to exemplify the methodology considered, a pilot and also a main sample survey are executed both in two rounds (interview–reinterview). The response error analysis is conducted and response reliability measures are calculated using formulas given in Section 2 for each question listed in Table 6, and the results obtained are presented in Table 7 and Table 8.

Response reliability measures for questions based on different levels of measurement scales are obtained for data collected from the sample surveys. Simple and correlated response errors are also estimated for different measurement scaled data. In this Section, the data obtained from these applications are evaluated with respect to the reliability measures and other statistics for all variables.

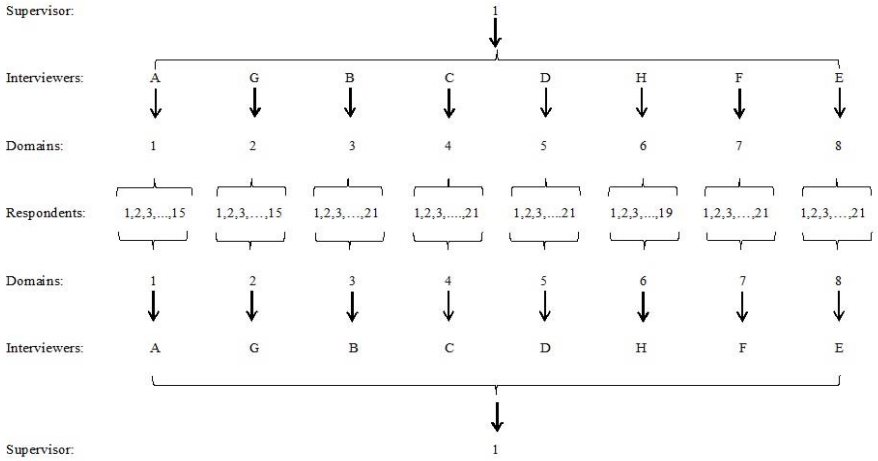

Figure 3. Parent main survey layout of the fieldwork

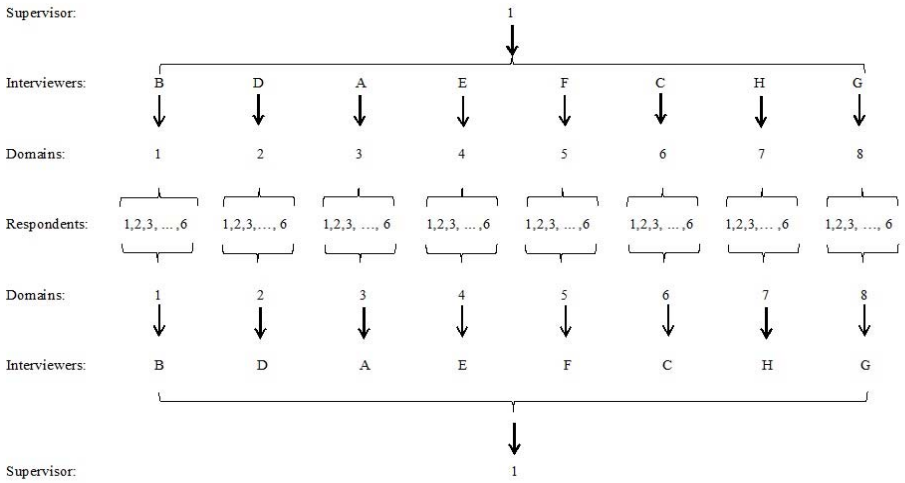

Figure 4. Main survey reinterview layout of the fieldwork

4.1. Findings of the pilot survey

When the results are examined, the following findings of the pilot study may be given as follows:

- For questions 1-5, we have completely reliable data with respect to all relevant indices considered; there exists ignorable response variance for question 6 with respect to *index of consistency* (*IOI*) and of *data reliability* ($r = 1 - IOI$) in Table 3, and also with respect to $\mathbf{s_r^2}$ in Table 4. For the rest of questions (7-10), there exist response variances with respect to *IOI*, r and also $\mathbf{s_r^2}$. Among them *interviewer effects* (α) are observed on questions 9, 7, 8 and 10, in decreasing order (Table 4).

Table 6. Main survey questions and related information

Main Survey Question Number	Measurement scale	Variable name	Pilot Survey Question Number
1	Dichotomy	Gender of respondent	1
2	Interval	Age of respondent	2
3	Interval	Height of respondent	3
4	Dichotomy	Marital status of respondent	
5	Likert scale	Last degree owned	4
6	Likert scale	Title of the respondent	5
7	Likert scale	Number of languages known	
8	Interval	Working duration in years in the university	6
9	Interval	Fixed working duration in a day in hours*	
10	Interval	Sleeping duration a day in hours	
11	Interval	Time spent for leisure and sports per week	
12	Interval	Time spent for eating and drinking per day	
13	Interval	Time spent with their family per week	
14	Dichotomy	Interested in cooking	
15	Interval	Payment on clothing in TL** per month***	7
16	Interval	Payment on cultural activities in TL** per month***	8
17	Likert scale	Job satisfaction	9
18	Likert scale	Salary satisfaction	10
19	Likert scale	Working duration satisfaction	
20	Likert scale	Current use for time satisfaction	

* Note that the time frame for the working duration may not be well defined for this variable, and may create limitation and potential reason of variability from one round to another. ** TL refers to Turkish Lira as a currency. *** "per month" refers to any average month within the year.

- There is no response error at all in questions 2, 3 and 4 of type interval, interval, and ordinal, respectively, according to response deviation (d), and *CI for d*, which includes zero (Table 4). Also, the associated almost perfect *kappa* values (K=1.0) are found to be statistically significant based on *CI for K* (Table 3). Besides, according to α, responses in the two interviews are perfectly correlated (Table 4).
- For questions 1 and 5 of the dichotomy type, *crude agreement (A), disagreement (D)* and *consistency* (K) index values indicate statistically significant with respect to *CI for K* and perfect agreement with respect to evaluation of K between two interviews (Table 3).
- There is very small but statistically significant response error for question 6 according to *response deviation* (d), and *CI for d* (Table 4). However, response variance is accounted for only 0.1% of the total variance with respect to *index of inconsistency* (*IOI*) (Table 3). For the question inquiring about the working period duration of respondents, the estimators of the *uncorrelated response variance* ($\mathbf{s_r^2}$) and *correlated interviewer variance* ($\widehat{\boldsymbol{\delta}}_{\mathbf{2}}\boldsymbol{s}_{\boldsymbol{r}}^{\mathbf{2}}$) are found to be low (Table 4), indicating smaller interviewer effect on the respondents.

- For questions 7 and 8 of the interval type, which ask the respondents about the amount of money spend on clothing and on cultural activities, respectively, *response errors* (*d*) and their associated *response variances* ($\mathbf{s_r^2}$) are the largest among the other questions (Table 4). However, there is a statistically significant difference in responses between two interviews only for question 8 with respect to *CI for d* (Table 4). The response variances are attributed to 40% and 69% of the total variance with respect to *IOI* for questions 7 and 8, respectively (Table 3). In addition, interviewer variances are attributed to 7% (=979/13.459) and 42% (=789/1.863) of the total variance for questions 7 and 8, respectively. According to the *kappa*, there is a moderate and fair agreement between responses of two interviews (Table 3), although correlation statistics do not indicate (α values are 0.775 and 0.947 for question 7 and 8, respectively (Table 4)).

Table 7. Response reliability statistics for the main survey

Ques. No.	A	D	A_k	D_k	IOI	IOI Eval.	r	K	K_w	K Eval.	CI for K
1	1.000	0.000	1.000	0.000	NA	NA		1.000	1.000	AP	(1.00, 1.00)*
2	0.937	0.063	1.000	0.000	0.011	L	0.989	0.908	0.952	AP	(0.81, 1.01)*
3	0.915	0.085	1.000	0.000	0.019	L	0.981	0.877	0.897	AP	(0.76, 0.99)*
4	1.000	0.000	1.000	0.000	NA	NA	NA	1.000	1.000	AP	(1.00, 1.00)*
5	1.000	0.000	1.000	0.000	0.000	L	1.000	1.000	1.000	AP	(1.00, 1.00)*
6	0.917	0.083	1.000	0.000	0.296	M	0.704	0.875	0.904	AP	(0.76, 0.99)*
7	0.833	0.167	1.000	0.000	0.143	L	0.857	0.583	0.621	M	(0.34, 0.83)*
8	0.896	0.104	1.000	0.000	0.048	L	0.952	0.859	0.883	AP	(0.73, 0.99)*
9	0.708	0.295	0.979	0.021	0.296	M	0.704	0.494	0.576	M	(0.27, 0.77)*
10	0.625	0.375	0.958	0.042	0.382	M	0.618	0.434	0.500	M	(0.23, 0.64)*
11	0.688	0.313	0.917	0.083	0.562	H	0.438	0.407	0.455	M	(0.19, 0.62)*
12	0.688	0.313	0.938	0.062	0.800	H	0.200	0.444	0.558	M	(0.23, 0.66)*
13	0.646	0.354	0.896	0.104	0.608	H	0.392	0.496	0.575	M	(0.32, 0.68)*
14	0.813	0.188	1.000	0.000	NA	NA	NA	0.631	0.698	S	(0.42, 0.84)*
15	0.542	0.458	0.813	0.187	0.340	M	0.660	0.317	0.414	F	(0.15, 0.49)*
16	0.667	0.333	0.917	0.083	0.171	L	0.829	0.301	0.385	F	(0.09, 0.51)*
17	0.604	0.396	1.000	0.000	0.442	M	0.558	0.360	0.427	F	(0.16, 0.56)*
18	0.562	0.438	0.938	0.062	0.302	M	0.302	0.399	0.522	F	(0.21, 0.59)*
19	0.542	0.458	0.938	0.062	0.547	H	0.547	0.263	0.356	F	(0.05, 0.48)*
20	0.646	0.354	0.938	0.062	0.265	M	0.265	0.521	0.602	M	(0.34, 0.70)*

Notes: 1. NA indicates that this statistic is irrelevant for this type of variable; M: Moderate; L: Low; H: High; F: Fair: AP: Almost Perfect; S: Substantial 2. * denotes statistically significant kappa value for that particular type of question at α=0.05 level of significance by paired-t test. 3. Indices of *A*: crude agreement; *D*: crude disagreement; *IOI*: inconsistency; *r*: reliability of data; *K:* consistency (kappa) 4. Values of *A*, *r* and *K* close to one (1) indicate consistent and reliable data.

- There exist very small and not statistically significant response errors associated with questions 9 and 10 of the ordinal type with respect to *response deviation* (*d*) and associated *CI for d*, asking about overall job and salary satisfaction of the respondents, respectively (Table 4). However, there exist response variances which are attributed to 47% and 33% of the total variance for questions 9 and 10 with respect to *index of*

consistency, *IOI*, respectively (Table 3). Also, an *interviewer effect* is detected with respect to α on the responses for these questions; however, they only account for 19% (=0.094/0.49) and 3.8% (=0.052/1.38) of the total variance for questions 9 and 10, respectively (Table 4). Reliability statistics indicate that there is substantial and moderate agreement between the responses of two interviews for questions 9 and 10 with respect to evaluation of kappa, *K* (Table 3), respectively; nevertheless, correlation statistics do not approve this (α values are 0.599 and 0.820 for question 9 and 10, respectively (Table 4)).

4.2. Findings of the main survey

Findings of the main survey may be given as follows:

- For some questions such as 1 and 4 with respect to *index of crude agreement*, *A*, in Table 6, and 5, 19 with respect to *response error, d,* in Table 7, there are no changes in the given responses. Inquiring about the satisfaction of respondents with daily working hours (Question 19), *indices of crude agreement (A) and of consistency (K)* have a fair agreement between responses of two interviews.

Table 8. Other response error statistics for the main survey

Ques. No.	$\bar{y}$	$\bar{y}_1$	$\bar{y}_2$	d= $\bar{y}_1 - \bar{y}_2$	CI for d	α	s_t^2	s_r^2	s_s^2	$\hat{\delta}_2 s_r^2$
1	NA	NA	NA	NA	NA	1.000	NA	NA	NA	NA
2	47.83	46.48	46.98	-0.5	(-3.75, 1.67)	0.989	122.72	1.40	121.32	0.000
3	172.24	171.38	171.33	0.05	(-0.49, 0.58)	0.979	80.79	1.54	79.25	0.130
4	NA	NA	NA	NA	NA	1.000	NA	NA	NA	NA
5	3.00	3.00	3.00	0	(0.00, 0.00)*	1.000	0.00	0.00	0.00	0.000
6	1.98	2.00	1.95	0.05	(-0.04, 0.13)	0.937	0.86	0.04	0.81	0.040
7	1.31	1.27	1.35	-0.08	(-0.20, 0.03)	0.775	0.56	0.08	0.48	0.110
8	18.52	16.33	17.45	-1.12	(-2.08, 0.23)	0.947	146.25	7.03	139.22	0.004
9	8.51	8.27	7.97	0.3	(-0.16, 0.76)	0.674	4.26	1.26	3.00	0.003
10	7.01	7.14	7.02	0.12	(-0.10, 0.33)	0.651	0.76	0.29	0.48	0.002
11	1.16	1.26	1.49	-0.23	(-0.66, 0.19)	0.635	1.85	1.04	0.81	0.047
12	1.58	1.65	1.70	-0.05	(-0.28,0.18)	0.403	0.45	0.36	0.10	0.046
13	3.45	3.56	3.31	0.25	(-0.44, 0.94)	0.641	4.72	2.87	1.86	0.068
14	NA	NA	NA	NA	NA	1.000	NA	NA	NA	NA
15	153.73	164.40	189.27	-24.87	(-66.69, 16.94)	0.822	30,538	10,369	20,169	2.395
16	144.20	116.98	106.56	10.42	(-19.37, 40.20)	0.659	26,475	4,525	21,950	822.44
17	4.281	4.31	4.25	0.06	(-0.12, 0.25)	0.665	0.43	0.19	0.24	0.229
18	2.615	2.60	2.63	-0.03	(-0.27, 0.23)	0.720	1.29	0.39	0.91	0.197
19	4.271	4.27	4.27	0	(-0.25, 0.25)	0.374	0.75	0.41	0.34	0.144
20	3.39	3.44	3.33	0.11	(-0.13, 0.34)	0.703	0.98	0.26	0.72	0.646

Notes: 1. NA indicates that this statistic is irrelevant for this type of variable, 2. * denotes statistically significant difference between the parent and reinterview values for that particular type of question at α=0.05 level of significance by paired-t test. 3. *d*: response deviance (error); α: interviewer effect; s_t^2: total variance; s_r^2: response variance; s_s^2: sampling variance $\hat{\delta}_2 s_r^2$: correlated interviewer variance 4. Values of *D*, *IOI*, and *d* close to zero (0) and also *CI for K* and for *d* containing zero indicate consistent and reliable data.

- Reliable data belong to questions 1-8 and 14 with respect to *A* and *r* in Table 6. Asking about the height of respondents (Question 3), the *correlated interviewer variance,* $\hat{\delta}_2 s_r^2$, is found to be very low (Table 7). Asking about the academic title of the respondents (Question 6), indices of *crude agreement, A, and of consistency, K,* have almost a perfect agreement between responses of the two interviews (Table 6). For the number of languages known asked in question 7, *index of consistency, K,* have a moderate agreement between responses of the two interviews (Table 6). Inquiring about if s/he is interested in cooking (Question 14), *index of consistency, K,* have a substantial agreement between responses (Table 6).
- Data belonging to questions 9, 10, 15, 16 have a moderate to fair agreement and reliability (Table 6).
- The least reliable data belong to questions 11, 12, 13, 19; response variance accounts for 56%, 80%, 61%, 55% with respect to *IOI* (Table 6), and interviewer variance accounts for 7% (=0.047/1.85), 10% (=0.046/0.45), 1.4% (=0.068/4.72), 19% (=0.144/0.75) of the total variance for questions 11, 12, 13, 19, respectively.
- Asking about the level of job satisfaction, Question 17 has a fair agreement between responses of the two interviews with respect to the index of consistency (*K*) (Table 6).
- Inquiring about the current salary satisfaction, Question 18 has a fair agreement between responses of the two interviews with respect to *indices of crude agreement, A, and of consistency, K* (Table 6).
- Asking about the satisfaction with the time use of respondents, Question 20 has a moderate agreement between responses of the two interviews with respect to indices of *crude agreement, A, and of consistency, K* (Table 6).

5. Conclusions

The main aim of this work is to investigate response errors which may stem from the respondent, interviewer or from their interaction, under interview-reinterview settings in sample surveys. We suggest using NDs in interview-reinterview surveys due to several reasons. First, an ND naturally provides estimation of interviewer effects due to its nested structure in which one respondent is interviewed by many interviewers. Next, it provides computing response errors independently in each survey. And also, it provides flexibilities in the field allocation and applications. In order to apply the suggested approach, an interview-reinterview survey is conducted at METU, Ankara, Turkey, to investigate the satisfaction of academicians' regarding life and time use.

Analysis of the pilot survey reveals that we have completely reliable data sometimes with an ignorable response variance on the questions inquiring about factual information about the participants such as gender, title, and so on. Nevertheless, there exist response variances in the questions involving elements hard to quantify, such as "amount of payment" or "duration". Besides, questions asking about respondents'

feelings, such as their "satisfaction level", seem to open higher interviewer effects. The last two questions are usually formulated with either ordinal or interval type of variables. Analysis of the pilot study reveals that the response reliability seems to be irrelevant to these two data types.

Analysis of the main survey results show that almost three fourth of the data are reliable and almost reliable. The rest of the questions are exposed to interviewer effects, and need more attention. Note that the response error that may be mostly attributed to the interviewer effect belongs to the one regarding respondents' satisfaction. As a result of the training provided to the interviewers' immediately after the pilot survey analysis, in the main survey, the associated response variances accounting for the total variance are considerably reduced from 69% to 29%. It is the outcome of having done two consecutive interview-reinterview designs.

As a future study, alternative interviewer allocation can also be examined on the basis of both the nested and factorial experimental design techniques. However, under such allocations, the number of interviewers which will be allocated for the field application may be combinatorically problematic.

The following limitations should be kept in mind while evaluating the response reliability measures. The time lag between two interviews should be reasonably large, enabling to recall their first response during the second interview. This issue is clarified by the following related literature. The World Fertility Survey's document on "Re-interview Survey Design" resulted in the fieldwork applications as the median lag between the first interview and the reinterview was 2 to 4 months for the planned five national studies (O'Muircheartaigh and Marckwardt, 1980). Also, O'Muircheartaigh (1982) suggests a regression type analysis to test independence between the two interviews. The time lag between the original interview and the reinterview varies between a few days to several months. Also, research on optimal time lags in different reinterview situations is rare in the literature (Forsman and Schreiner, 2004). Memory recall errors are affected by the *time duration between the two reference points* (interview and reinterview) as well as the *importance of the event*, *frequency of occurence of the event*, *measurement scale of the event*, and *bounded* or *aided recalling* (Ayhan and Işıksal, 2004). Consequently, the time difference between the two cannot be the only criteria for evaluation. When the time interval between the interview and the reinterview is very short, then reinterviewed respondents can recall their earlier responses, and they may be losing interest, also it is possible to agree with their previous reply to the interview as a form of satisficing. For very lengthy questionnaires, respondent fatigue is also possible, but may not be the case for short questionnaire surveys.

Acknowledgements

We would like to thank Prof. A. Sinan Türkyılmaz of the Hacettepe University, Institute of Population Studies, Prof. Özlem İlk Dağ and Assoc. Prof. Ceylan Yozgatlıgil of Middle East Technical University, Statistics Department for their contribution to this work. The authors also would like to thank the Editor in Chief of the Statistics in Transition and the anonymous reviewers for their proposals and contributions to the current version of the article.

References

Ayhan, H. Ö., (2003). Models of response error components in supervised interview–reinterview surveys. *Journal of Applied Statistics*, 30(9), pp. 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, 2012.

Ayhan, H. Ö., Işiksal, S., (2004). Memory recall errors in retrospective surveys: A reverse record check stıdy. *Quality and Quantity*, 38, pp. 475–493.

Bailar, B. A., Dalenius, T., (1969). Estimating the response variance components of the U.S. Bureau of the Census' survey model. *Sankhya: The Indian Journal of Statistics*, Series B, Vol. 31(3/4), pp. 341–360.

Bassi, F., Fabbris, L., (1997). Estimators of nonsampling errors in interview – reinterview supervised surveys with interpenetrated assignments. In *Survey Measurement and Process Quality*, L. Lyberg, P. Biemer, M. Collins, E. DeLeeuw, C. Dippo, N. Schwarz, and D. Trewin, eds., John Wiley and Sons, New York, pp. 733–751.

Batmaz, I., Fahmi, F. M., (2015). Response variance estimation in personal interview surveys with several allocation schemes, *Proceedings of the 60th World Statistics Congress of the International Statistical Institute*, Rio de Janeiro, Brazil, pp. 1–6.

Biemer, P. P., (2012). *Latent Class Analysis of Survey Error*. John Wiley and Sons, New Jersey.

Biemer, P. P., Lyberg, L. E., (2003). *Introduction to Survey Quality*. Hoboken: Wiley Series in Survey Methodology

Biemer, P. P., Stokes, S. L., (1985). Optimal design of interviewer variance experiments in complex surveys. *Journal of the American Statistical Association*, 80, pp. 158–166.

Biemer, P. P., Stokes, S. L., (2004). Approaches to the modeling of measurement error. In *Measurement Errors in Surveys* Edited by P. Biemer, R.M. Groves, L. Lyberg, N.A. Mathiowetz, and S. Sudman. Wiley Series in Survey Methodology, New Jersey, pp. 487–516.

Biemer, P. P., Groves, R.M., Lyberg, L. E., Mathiowetz, N. A., Sudman, S., (2004). *Measurement Errors in Surveys*. Wiley Series in Survey Methodology, New Jersey.

Box, G. E. P., Hunter, J. S., Hunter, W. G., (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*. John Wiley and Sons, New Jersey.

Cicchetti, D. V., (1972). A new measure of agreement between rank-ordered variables. *Proceedings of the 80th Annual Conference of the American Psychological Association* 1972, pp. 17–18.

Cohen, J., (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, pp. 37–46.

Cohen, J., (1968). Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, pp. 213–220.

Dahlhamer, J., Zablotsky, B., Zelaya, C., Maitland, A., (2020). Modeling interviewer effects in the National Health Interview Survey. Chapter 21 in Olson, K., Smyth, J.D., Dykema, J., Holbrook, A.L., Kreuter, F., West, B.T., Eds. (2020 a). *Interviewer Effects from a Total Survey Error Perspective.* New York: Chapman and Hall/CRC., pp. 295–310.

Dijkstra, W., (1983). How interviewer variance can bias the results of research on interviewer effects. *Quality and Quantity*, 17, 179–187.

Edwards, B., Sun, H., Hubbard, R., (2020). Behavior change techniques for reducing interviewer contributions to total survey error. Chapter 6 in Olson, K., Smyth, J.D., Dykema, J., Holbrook, A.L., Kreuter, F., West, B.T., Eds. (2020 a). *Interviewer Effects from a Total Survey Error Perspective.* New York: Chapman and Hall/CRC., pp. 77–90.

Fahmi, F. M., (2013). *Estimation of Response Errors in Complex Sample Surveys.* Ph.D. thesis, Graduate School of Natural and Applied Sciences, Middle East Technical University, Ankara, Turkey.

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

Forsman, G., Schreiner, I., (2004). The design and analysis of reinterview: An overview. Chapter 15 in *Measurement Errors in Surveys* Edited by P. Biemer, R.M. Groves, L. Lyberg, N.A. Mathiowetz, and S. Sudman. Wiley Series in Survey Methodology, New Jersey, pp. 279–301.

Fuller, W. A., (1995). Estimation in the presence of measurement error. International Statistical Review, 63(2), pp, 121–147.

Groves, R. M., Magilavy, L. J., (1986). Measuring and explaining interviewer effects in centralized telephone facilities. *Public Opinion Quarterly*, 50, pp. 251–266.

Moser, C. A., Kalton, G., (1979). *Survey Methods in Social Investigation.* Heinemann Educational Books, London.

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

Hansen, M. H., Hurwitz, W. N., Marks, E. S., Mauldin, W. P., (1951). Response errors in surveys. *Journal of the American Statistical Association*, 46(254), pp. 147–190.

Hansen, M. H., Hurwitz, W. N., Pritzker, L., (1964). The estimation and interpretation of gross differences and the simple response variance. In C.R. RAO (Editor): Contributions to Statistics, Presented to Professor P.C. Mahalanobis on the Occasion of his 70th Birthday. Calcutta, India: Pergamon Press, Oxford and Statistical Publishing Society.

Hagenaars, J. A., Heinen, T. G., (1982). Effects of role-independent interviewer characteristics on responses. In W.Dijkstra and J.van der Zouwen (eds.), *Response Behaviour in the Survey Interview.* Academic Press, London, pp. 91–130.

Hox, J. J., (1994). Hierarchical regression models for interviewer and respondent effects. *Sociological Methods and Research*, 22, pp. 300–318.

Hox, J. J., De Leeuw, E. D., Kreft, I. G. G., (2004). The effect of interviewer and respondent characteristics on the quality of survey data: A multilevel model. Chapter 22 in Biemer *et al.* (2004) *Measurement Errors in Surveys*. Wiley Series in Survey Methodology, New Jersey, pp.439–461.

Kalton, G., (1983). Introduction to survey sampling. *Quantitative Applications in the Social Sciences, A SAGE University Papers Series,* No. 35, Newbury Park.

Kish, L., (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, pp. 91–115.

Kish, L., (1965). Survey Sampling. John Wiley and Sons, New York.

Landis, J. R., Koch, G., (1977). A review of statistical methods in the analysis of data rising from observer reliability studies. *Statistica Neerlandica*, 29, pp. 101–123 and pp. 151–161.

Lyberg, L., Kasprzyk, D., (2004). Data collection methods and measurement error: An overview. Chapter 13 in Biemer *et al.* (2004) *Measurement Errors in Surveys*. Wiley Series in Survey Methodology, New Jersey, pp. 237–257.

Mahalanobis, P. C., (1946). Recent Experiments in Statistical Sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, Vol. 109, pp. 325–378.

McCarthy, P. J., (1966). Replication: An approach to the analysis of data from complex surveys. National Center for Health Statistics, *Vital and Health Statistics Series*, No. 2(14).

McNabb, D. E., (2014). *Nonsampling Error in Social Surveys.* Los Angeles: SAGE Publications, Inc.

Montgomery, D. C., (2012). *Design and Analysis of Experiments.* New York: John Wiley and Sons.

Olson, K., Smyth, J. D., Dykema, J., Holbrook, A. L., Kreuter, F., West, B. T., Eds., (2020a). *Interviewer Effects from a Total Survey Error Perspective.* New York: Chapman and Hall/CRC.

Olson, K., Smyth, J. D., Dykema, J., Holbrook, A. L., Kreuter, F., West, B. T., (2020 b). The past, present, and future of research on interviewer effects. Chapter 1 in Olson, K., Smyth, J.D., Dykema, J., Holbrook, A.L., Kreuter, F., West, B.T., Eds. (2020 a). *Interviewer Effects from a Total Survey Error Perspective.* New York: Chapman and Hall/CRC., pp. 3–16.

O'Muircheartaigh, C. A., (1976). Response errors in attitudinal sample surveys. *Quality and Quantity*, 10, pp. 97–115.

O'Muircheartaigh, C. A., (1977). Response errors. In *Analysis of Survey Data*, Vol. 2, John Wiley and Sons, London, pp. 193–239.

O'Muircheartaigh, C. A., (1982). *Methodology of the Response Errors Project.* Fertility Survey, Scientific Reports, No. 28, London, UK. 32 pp.

O'Muircheartaigh, C. A., (1984). The pattern of response variance in the Peru fertility survey. *WFS Scientific Report*, No. 45.

O'Muircheartaigh, C. A., (1986). Correlates of reinterview response inconsistency in the Current Population Survey. *Proceedings of the Second Annual Research Conference*, U.S. Bureau of the Census, pp. 208–234.

O'Muircheartaigh, C. A., (1991). Simple response variance: estimation and determinants. In *Measurement Errors in Surveys* Edited by P. Biemer, R.M. Groves, L. Lyberg, N.A. Mathiowetz, and S. Sudman. Wiley Series in Survey Methodology, New Jersey, pp. 551–574.

O'Muircheartaigh, C. A., Campanelli, P., (1998). The relative impact of interviewer effects and sample design effects on survey precision. Journal of the Royal Statistical Society, A 161, pp. 63–77.

O'Muircheartaigh, C. A., Marckwardt, A. M., (1980). An assessment of the reliability of WFS data. World Fertility Survey Conference 1980 London, Record of Proceedings Volume 3, pp. 313–379.

Salant, P., Dillman, D. A., (1994). *How to Conduct your own Survey?* John Wiley and Sons, New York.

Verma, V., (1980). Sampling for national fertility surveys. World Fertility Survey Conference 1980 London, Record of Proceedings, Vol. 3, pp. 389–448.

West, B. T., (2020). Designing studies for comparing interviewer variance in two groups of survey interviewers. Chapter 23 in Olson, K., Smyth, J.D., Dykema, J., Holbrook, A.L., Kreuter, F., West, B.T., Eds. (2020 a). *Interviewer Effects from a Total Survey Error Perspective.* New York: Chapman and Hall/CRC., pp. 323–334.

WFS, (1977). *Guidelines for Country Report No.1.* WFS Basic Documentation No.8, London, U.K.: World Fertility Survey.

Yu, C. H., Ohlund, B., Digangi, S., Jannasch-Pennell, A., (2000). Estimating the reliability of self-reporting data for web-based instruction. *Instruction and Research Support, Arizona State University*, US.