METRON (2017) 75:345–357
DOI 10.1007/s40300-017-0124-0

A measurement error model approach to survey data integration: combining information from two surveys

Seho Park[1] · **Jae Kwang Kim**[1] · **Diana Stukel**[2]

Received: 10 July 2017 / Accepted: 11 September 2017 / Published online: 18 September 2017

Abstract Combining information from several surveys from the same target population is an important practical problem in survey sampling. The paper is motivated by work that authors undertook, sponsored by the Food and Nutrition Technical Assistance III Project (FANTA), with funding from the U.S. Agency for International Development (USAID) Bureau of Food Security (BFS). In the project, two surveys were conducted independently for some areas and we present a measurement error model approach to integrate mean estimates obtained from the two surveys. The predicted values for the counterfactual outcome are used to create composite estimates for the overlapped areas. An application of the technique to the project is provided.

Keywords Counterfactual outcome · Composite estimate · Variance estimation

1 Introduction

Survey integration is an emerging research area of statistics, which concerns combining information from two or more independent surveys to get improved estimates for various parameters of interest for the target population. One of the early applications of survey integration is the Consumer Expenditure Survey [20], where two survey vehicles (a Diary survey and a quarterly interview survey) were used to obtain improved estimates for the Diary survey items. Renssen and Nieuwenbroek [16], Merkouris [12,13], Wu [18] and Ybarra and Lohr [19] considered the problem of combining data from two independent surveys to estimate totals at the population and domain levels.

Combining information from two or more independent surveys is a problem frequently encountered in survey sampling. One of the classical setups used to combine information

✉ Jae Kwang Kim
jkim@iastate.edu

[1] Iowa State University, Ames, IA, USA

[2] FANTA III Project, FHI 360, Washington, DC, USA

Table 1 Data structure for combining two surveys with measurement errors

	x	y_1	y_2
Survey A	o	o	
Survey B	o		o

is two-phase sampling, where the measurement x is observed in both surveys and the study variable y is observed only from one survey, say, in Survey A. There is no measurement for y in survey B. In this case, we can treat the union of Survey A and Survey B samples as a phase one sample and treat the Survey A sample as a phase two sample. Hidiroglou [6] formulated this problem and developed efficient estimation using a two-phase regression estimation method. Fuller [4], Legg and Fuller [11], and Kim and Rao [9] considered this problem as a missing data problem and developed mass imputation to obtain improved estimation for the total as well as domain totals. Our setup is different from the two-phase sampling approach in the sense that we have a different measurement of y from two surveys.

We consider a situation where two surveys have common measurement for x but different measurements for y. For example, x can be demographic information that does not suffer from measurement errors but y can suffer from survey-specific measurement errors. The survey-specific difference can occur due to differences in survey questions or survey modes (e.g. [2]). In Table 1, for example, the Survey A sample contains observations in x and y_1 while the Survey B sample contains observations in x and y_2. In the case of y_1 being the study variable of interest, if we can assume that y_2 is a measurement for y_1 with measurement errors, then at issue is the estimation of the population mean of y_1 combining two surveys.

Our research is motivated by work sponsored by The Food and Nutrition Technical Assistance III Project (FANTA) with funding from U.S. Agency for International Development (USAID), to produce integrated estimates from two independent surveys conducted in Guatemala where the geographic areas covered by the two surveys have substantial overlap.

Section 2 provides background on the projects and data descriptions and Sect. 3 introduces the proposed method for survey integration. In Sect. 4, we illustrate the estimation process and results of the work sponsored by FANTA, and Sect. 5 provides concluding remarks.

2 The Food and Nutrition Technical Assistance III Project

2.1 Background

FANTA is a 5-year cooperative agreement between the USAID and FHI 360. FANTA aims to improve the health and well-being of vulnerable groups through technical support in the areas of maternal and child health and nutrition in development and emergency contexts, HIV and other infectious diseases, food security and livelihood strengthening, agriculture and nutrition linkages and emergency assistance in nutrition crises.

USAID is the lead U.S. government agency that works to end extreme global poverty and enable resilient, democratic societies to realize their potential. The Feed the Future Initiative (FTF) was launched in 2010 by the United States government to address global hunger and food insecurity. The Initiative is coordinated primarily by the USAID and is housed within the Bureau of Food Security (BFS), but includes the Office of Food for Peace (FFP). The main objectives of the FTF initiative are the advancement of global agricultural development, increased food production and food security, and improved nutrition particularly for vulnerable populations such as women and children. The FTF initiative is active in 19 focus

Table 2 Eleven common indicators

Level	Indicator
Household	Daily per capita expenditures (PCE)
	Prevalence of households with hunger (HHS)
	Prevalence of poverty (PP)
	Mean depth poverty (MDP)
Individual (children)	Prevalence of stunted children
	Prevalence of wasted children
	Prevalence of underweight children
	Prevalence of children receiving a minimum acceptable diet (MAD)
	Prevalence of exclusive breastfeeding (EBF)
Individual (women)	Prevalence of underweight women
	Women's dietary diversity score (WDDS)

developing countries in Africa, Asia and Latin America. One of these focus countries is Guatemala.

Both BFS (through the FTF initiative) and FFP sponsor periodic baseline, interim and end-line household surveys to gauge the extent of progress towards achieving the goals of the FTF initiative. In 2013, FFP engaged a third party contractor, ICF International, to conduct a baseline household survey in five departments of the Western Highlands of Guatemala. In the same year, BFS/FTF (henceforth referred to as FTF) engaged a third party contractor, UNC MEASURE, to conduct an interim household survey in the same five departments in Guatemala. Although the surveys were conducted in the same five departments, the geography of the two surveys did not exactly coincide; however, there was substantial geographic overlap. The union of the geography covered by the two surveys represents the FTF Zone of Influence (ZOI), where some of the most food insecure parts of the population in the country reside. Because, FTF was interested in obtaining ZOI-level estimates for a number of key indicators using data from the two independent surveys, they provided funding to FANTA, who in turn, engaged the authors to undertake the work. Because of the overlapped geography from the two surveys, it was necessary to use data integration methods to produce overall ZOI-level estimates.

Guatemala has 22 departments, which are geographic entities, divided into 334 municipalities. The two surveys were each conducted in the following five departments of the Western Highlands of Guatemala: San Marcos, Totonicapan, Quiche, Quezaltenango, and Huehuetenango. Thus, two surveys were conducted in the areas and the survey data from the two samples are ready to be combined for survey integration. More details of this project can be found from the reference provided by USAID [17].

2.2 Common indicators

ICF International (FFP) and UNC MEASURE (FTF) used their own questionnaire for the surveys, and among the indicators in the questionnaires, there were 11 common indicators in both surveys indicating maternal and child health status. Among the 11 common indicators, 4 were collected at the household-level and the remaining 7 were collected at the individual-level. Five indicators of the 7 individual-level indicators pertained to children and remaining 2 to women. Table 2 presents the common indicators and their descriptions.

Table 3 Survey design of the FFP project

Department	Strata	Total no. of clusters	No. of selected clusters
1. San Marcos	11	89	17
	12	30	17
2. Totonicapan	21	85	22
	22	22	19
3. Quiche	31	62	22
	32	19	13
4. Huehuetenango	41	48	12
	42	18	12
5. Quetzaltenango	51	24	16

Most indicator variables are dichotomous, taking the values of either 0 or 1 in both data sets, but the other two indicator variables, which are 'PCE' and 'WDDS,' are numeric in both data sets. In this paper, we focus on the 'PCE' and the 'HHS' indicators for analysis as examples of a numeric variable and a dichotomous variable, respectively.

2.3 Survey design

2.3.1 FFP survey

The survey for the FFP project used a three-stage sampling design. In the first stage, the primary sampling unit is the village, where the village population for five departments is divided into two substrata in each department. Each department has two substrata except for Quetzaltenango which has one stratum. So, we have nine strata and the first stage sample selection probability is based on the number of villages in the sampling frame and the size of the village within each stratum. The sampling frame for the first stage sampling included all the villages identified for program implementation. Table 3 shows the summary of sample clusters in each stratum.

In the second stage sampling, sample households were selected randomly from each sampled village. The target number of households selected for each village was 40. The second stage sample selection probability is based on the number of households selected for each village divided by the total number of households in each village.

The third stage sampling was done at the individual level to select woman and children in households. The third stage sample selection probability is based on the total number of individuals selected for each interview module and the number of eligible individuals in the household. Only one eligible woman was randomly selected using the Kish grid [10], but all children were selected to be interviewed.

The final sampling weights are computed as the inverse of products of the three stage first-order inclusion probabilities.

2.3.2 FTF survey

The survey for the FTF project also used a three-stage sampling design using census sectors as the primary sampling units. In the first stage, the census areas (urban/rural) were formed in

Table 4 Survey design of the FTF project

Department	Strata	Total no. of clusters	No. of selected clusters
1. San Marcos	Rural	192	25
	Urban	99	3
2. Totonicapan	Rural	237	5
	Urban	128	1
3. Quiche	Rural	284	33
	Urban	97	7
4. Huehuetenango	Rural	336	39
	Urban	80	8
5. Quetzaltenango	Rural	117	1
	Urban	190	1

each department and census sectors were sampled within the census area. From the sampled census sectors, the sample households were randomly selected in the second stage sampling. For the third stage sampling, data on individual-level women and children were collected. All women and children in a household are included in the sample, but the weights associated with women and children are adjusted for nonresponse. Table 4 shows the summary of sample clusters in each stratum.

3 Survey data integration

We present the proposed method in the context of measurement error models. In a classical measurement error model problem, the interest lies in estimating the regression coefficient for the regression of y on x and the covariate x is subject to measurement errors [5]. In our problem, the measurement error occurs in y for one survey (Survey B) and we are interested in combining two surveys to estimate the population mean of y more efficiently. Thus, we still consider the data structure in Table 1. We treat y_1 as the gold standard, $y_1 = y$, in the sense that there is no measurement error in y_1.

Let $f_1(y_1 \mid x; \theta_1)$ be the density for the conditional distribution of y_1 on x, characterized by parameter θ_1. Model for $f_1(y_1 \mid x; \theta_1)$ can be called a structural equation model [3]. Let $f_2(y_2 \mid x, y_1; \theta_2)$ be the density for the conditional distribution of y_2 on (x, y_1), characterized by parameter θ_2. For parameter identifiability, we assume that

$$f_2(y_2 \mid x, y_1) = f_2(y_2 \mid y_1). \tag{1}$$

Such assumption is sometimes called the nondifferential measurement error assumption [1, p. 7] in the measurement error model literature. That is, x is an instrumental variable for y_1. The nondifferential measurement error assumption is used to obtain a reduced model.

Given the sample with the data structure in Table 1, the imputed values for y_1 in sample B are used to obtain the composite estimator that combines direct observations in the sample A and synthetic values in the sample B. The imputed values are the best predicted values of the counterfactual outcome variable y_1 in sample B, which correct for measurement errors in observed valued of y_2. The imputed values are generated using the prediction model for y_1, $f(y_1 \mid x, y_2)$.

For the parameter estimation, the (pseudo) maximum likelihood estimator of θ_1 and θ_2 can be obtained by using the full EM algorithm as follows:

[E-step] Compute

$$Q_1(\theta_1|\theta_1^{(t)}, \theta_2^{(t)}) = \sum_{i\in S_a} w_{ia}\log f_1(y_{1i}|x_i;\theta_1) + \sum_{i\in S_b} w_{ib}\mathrm{E}\big[\log f_1(y_{1i}|x_i;\theta_1) \mid x_i, y_{2i}; \theta_1^{(t)}, \theta_2^{(t)}\big]$$

and

$$Q_2(\theta_2|\hat{\theta}_1^{(t)}, \theta_2^{(t)}) = \sum_{i\in S_a} w_{ia}\mathrm{E}\big[\log f_2(y_{2i}|y_{1i};\theta_2) \mid x_i, y_{1i}; \hat{\theta}_1^{(t)}, \theta_2^{(t)}\big] + \sum_{i\in S_b} w_{ib}\mathrm{E}\big[\log f_2(y_{2i}|y_{1i};\theta_2) \mid x_i, y_{2i}; \hat{\theta}_1^{(t)}, \theta_2^{(t)}\big],$$

where S_a and S_b are the index sets for the Survey A sample and the Survey B sample, respectively. Also, w_{ia} and w_{ib} are the sampling weight for unit $i \in S_a$ and for unit $i \in S_b$, respectively. The conditional expectation in Q_1 is taken with respect to

$$f(y_1|x, y_2;\theta_1,\theta_2) = \frac{f_1(y_1|x;\theta_1)f_2(y_2|y_1;\theta_2)}{\int f_1(y_1|x;\theta_1)f_2(y_2|y_1;\theta_2)dy_1}$$

evaluated at $\theta_1 = \theta_1^{(t)}$ and $\theta_2 = \theta_2^{(t)}$ for Q_1 and at $\theta_1 = \hat{\theta}_1^{(t)}$ and $\theta_2 = \theta_2^{(t)}$. For Q_2, the first conditional expectation is taken with respect to $f(y_{2i}|x_i, y_{1i}) = f(y_{2i}|y_{1i})$ by the assumption (1), evaluated at $\theta_2 = \theta_2^{(t)}$.

[M-step] Update θ_1 by maximizing $Q_1(\theta_1|\theta_1^{(t)}, \theta_2^{(t)})$ with respect to θ_1 and update θ_2 by maximizing $Q_2(\theta_2|\hat{\theta}_1^{(t)}, \theta_2^{(t)})$ with respect to θ_2.

Based on the estimated parameters $\hat{\theta}_1$ and $\hat{\theta}_2$, the best predictor of y_1 of the Survey B sample is obtained as the expectation of the predictive distribution, which is the conditional distribution of y_1 given x and y_2. That is, the best predictor of y_{1i} is

$$\hat{y}_{1i}^{*} = \mathrm{E}\big(y_{1i}|x_i, y_{2i};\hat{\theta}_1,\hat{\theta}_2\big). \tag{2}$$

The parametric fractional imputation of [7] can be used to generate fractionally imputed values for y_1 in sample B under the general parametric models [14]. When $f_1(y_1|x;\theta_1)$ and $f_2(y_2|x, y_1;\theta_2)$ have general parametric models, the prediction model may not have a closed form. In this case, the parametric fractional imputation can be used following two-step method:

1. For each $i \in S_b$, generate $y_{1i}^{*(j)}$ from $f_1(y_{1i} \mid x_i;\hat{\theta}_1)$ for $j = 1, \ldots, m$.
2. Let $y_{1i}^{*(j)}$ be the j-th imputed value of y_{1i} obtained from Step 1. The fractional weight assigned to $y_{1i}^{*(j)}$ is

$$w_i^{*(j)} = \frac{f_2(y_{2i} \mid x_i, y_{1i}^{*(j)};\hat{\theta}_2)}{\sum_{k=1}^{m} f_2(y_{2i} \mid x_i, y_{1i}^{*(k)};\hat{\theta}_2)}.$$

Once we use the parametric fractional imputation, the conditional expectation in (2) can by computed by a Monte Carlo approximation. That is, the conditional expectation can be written by

$$\hat{y}_{1i}^{*} \cong \sum_{j=1}^{m} w_i^{*(j)} y_{1i}^{*(j)}.$$

Using the counterfactual values (2) of the Sample B and observations of the Survey A sample, we can construct a composite estimator that combines two values. The combined estimator is

$$\bar{y}_{com}^{*} = \frac{\sum_{i \in S_a} w_{ia} y_{1i} + \sum_{i \in S_b} w_{ib} \hat{y}_{1i}^{*}}{\sum_{i \in S_a} w_{ia} + \sum_{i \in S_b} w_{ib}}.$$

Kim et al. [8] have investigated the parametric fractional imputation of Kim [7] in the context of statistical matching where the main interest lies in estimating θ_2 in $f_2(y_2 \mid x, y_1; \theta_2)$. In their simulation study, the imputation model is based on the nondifferential measurement error assumption, but they noticed that departure from the assumption does not affect the validity of the imputation estimator for the population mean of y_1, even though it leads to biased estimation of the regression parameters. Note that if the assumption does not hold, then the imputation model (based on the assumption) is incorrectly specified. Under the incorrectly specified model, the imputed estimator is still unbiased for the mean estimation, as long as an intercept term is included in the model [9].

4 Application of methodology to USAID surveys in Guatemala

Based on the two estimates obtained from the two independent surveys on the overlap areas, we can improve the efficiency of the estimation by combining the two estimates.

4.1 Survey data integration

In this section, we use a measurement error model approach to integrate two surveys, the FFP and the FTF, presented in Sect. 3. In the view of the measurement error model approach, we treat one sample as a gold standard and the other sample containing measurement errors.

Throughout this study, the FFP sample was used as a benchmark and we predicted the counterfactual outcomes of the FTF sample, which is the value that would have obtained when the FTF sample was collected by ICF International who conducted the FFP project. This is based on the idea that measurement errors between two surveys are diminished when we consider the predicted values of the counterfactual values instead of the original values from the survey. We chose the FFP sample as a reference point since it has a smaller residual sum of squares compares to the one from the FTF sample.

4.1.1 Case 1: continuous study variable

Since the PCE indicator has continuous values, we treat a structural equation model and a measurement error model both follow normal distributions. Assume that a structural equation model for y_1 is

$$y_{1i} = \boldsymbol{\beta}_1 \mathbf{x}_{1i} + \beta_2 x_{2i} + e_i, \tag{3}$$

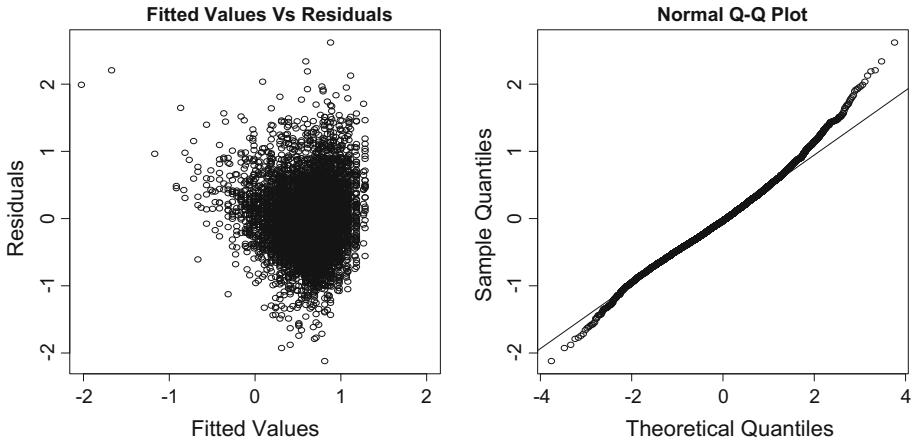

Fig. 1 Model diagnostics of model (3)

where $\mathbf{x}_{1i}$ is a department indicator and x_{2i} is a variable indicating the total number of household members, and $e_i \sim N(0, \sigma_e^2)$. Also, a measurement error model for y_2 is

$$y_{2i}|y_{1i} = \alpha_0 + \alpha_1 y_{1i} + u_i,$$

where $u_i \sim N(0, \sigma_u^2)$. By using the Bayes theorem, the predictive distribution can be derived as

$$y_{1i}|y_{2i}, \mathbf{x}_i \sim N(\mu_i, v^2), \tag{4}$$

where $\mathbf{x}_i = (\mathbf{x}_{1i}, x_{2i})$ with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \beta_2)$ and

$$\mu_i = c_i \boldsymbol{\beta} \mathbf{x}_i + (1 - c_i)\alpha_1^{-1}(y_{2i} - \alpha_0)$$

with

$$c_i = \frac{1/\sigma_e^2}{1/\sigma_e^2 + \alpha_1^2/\sigma_u^2}$$

and

$$v^2 = \frac{\sigma_e^2 \sigma_u^2/\alpha_1^2}{\sigma_e^2 + \sigma_u^2/\alpha_1^2}.$$

For the analysis of the PCE indicator, we assumed a linear regression model (3). The model diagnostics for the model assumptions are given in Fig. 1. Two plots show that the normality assumption and the homogeneity of variance assumption are appropriate. Residual plot also shows no particular pattern in residuals so the model assumptions in (3) are regarded as reasonable.

For the parameter estimation, we write $\theta_1 = (\boldsymbol{\beta}_1, \beta_2, \sigma_e^2)$ and $\theta_2 = (\alpha_0, \alpha_1, \sigma_u^2)$. The best estimator of θ_1 and θ_2 can be obtained by the full EM algorithm as explained in Sect. 3. In this example, the Q_1 and Q_2 are as follows:

[E-step] Compute

$$Q_1(\theta_1|\theta_1^{(t)},\theta_2^{(t)}) = \sum_{i\in S_a} w_{ia}\left\{-\frac{1}{2}\log(\sigma_e^2) - \frac{1}{2\sigma_e^2}(y_{1i}-\boldsymbol{\beta}\mathbf{x}_i)^2\right\} + \sum_{i\in S_b} w_{ib}\mathrm{E}\left[-\frac{1}{2}\log(\sigma_e^2) - \frac{1}{2\sigma_e^2}(y_{1i}-\boldsymbol{\beta}\mathbf{x}_i)^2 \mid \mathbf{x}_i, y_{2i}; \theta_1^{(t)},\theta_2^{(t)}\right]$$

and

$$Q_2(\theta_2|\hat{\theta}_1^{(t)},\theta_2^{(t)}) = \sum_{i\in S_a} w_{ia}\mathrm{E}\left[-\frac{1}{2}\log(\sigma_u^2) - \frac{1}{2\sigma_u^2}(y_{2i}-\alpha_0-\alpha_1 y_{1i})^2 \,\middle|\, \mathbf{x}_i, y_{1i}; \hat{\theta}_1^{(t)},\theta_2^{(t)}\right] + \sum_{i\in S_b} w_{ib}\mathrm{E}\left[-\frac{1}{2}\log(\sigma_u^2) - \frac{1}{2\sigma_u^2}(y_{2i}-\alpha_0-\alpha_1 y_{1i})^2 \,\middle|\, \mathbf{x}_i, y_{2i}; \hat{\theta}_1^{(t)},\theta_2^{(t)}\right],$$

where the conditional distribution for

$$f(y_1|\mathbf{x},y_2;\theta_1,\theta_2) = \frac{f_1(y_1|\mathbf{x};\theta_1)f_2(y_2|y_1;\theta_2)}{\int f_1(y_1|\mathbf{x};\theta_1)f_2(y_2|y_1;\theta_2)dy_1}$$

is also normal as in (4), evaluated at $\theta_1 = \hat{\theta}_1^{(t)}$ and $\theta_2 = \hat{\theta}_2^{(t)}$.

[M-step] Update θ_1 by maximizing $Q_1(\theta_1|\theta_1^{(t)},\theta_2^{(t)})$ with respect to θ_1 and update θ_2 by maximizing $Q_2(\theta_2|\hat{\theta}_1^{(t)},\theta_2^{(t)})$ with respect to θ_2.

Based on the estimated parameters $\hat{\theta}_1$ and $\hat{\theta}_2$, the best predictor of y_1 of the FTF sample is obtained as a mean of the predictive distribution, which is a conditional expectation of y_1 given $\mathbf{x}$ and y_2. That is,

$$\hat{y}_{1i}^* = \widehat{E}(y_{1i}|\mathbf{x}_i, y_{2i}) = \frac{\hat{\boldsymbol{\beta}}\mathbf{x}_i/\hat{\sigma}_e^2 + \hat{\alpha}_1(y_{2i}-\hat{\alpha}_0)/\hat{\sigma}_u^2}{1/\hat{\sigma}_e^2 + \hat{\alpha}_1^2/\hat{\sigma}_u^2}$$

is the best prediction of y_{1i} in the FTF sample that correct for measurement errors in y_{2i}.

Using the counterfactual values of the FTF sample and observations of the FFP sample, we can construct a composite estimator that combines two values. The combined estimator is

$$\bar{y}_{com}^* = \frac{\sum_{i\in S_a} w_{ia}y_{1i} + \sum_{i\in S_b} w_{ib}\hat{y}_{1i}^*}{\sum_{i\in S_a} w_{ia} + \sum_{i\in S_b} w_{ib}}, \tag{5}$$

where S_a and S_b denote the FFP sample and the FTF sample, respectively.

4.1.2 Case 2: dichotomous study variable

When a study variable is dichotomous, such as the HHS indicator in the project, the normal distribution assumption does not hold for both the structural equation model and the measurement error model. In this case, we consider a logistic regression model for the structural equation model and the misclassification model is used instead of the measurement error model [1]. The structural equation model for y_1 is

$$y_{1i}|\mathbf{x}_i \sim \mathrm{Ber}(r_i),$$

where $\mathbf{x}_i = (\mathbf{x}_{1i}, x_{2i})$ and

$$r_i = \frac{\exp(\boldsymbol{\beta}_1 \mathbf{x}_{1i} + \beta_2 x_{2i})}{1 + \exp(\boldsymbol{\beta}_1 \mathbf{x}_{1i} + \beta_2 x_{2i})},$$

where $\mathbf{x}_{1i}$ is a department indicator and x_{2i} is a variable indicating total number of household members. The misclassification model is given

$$f(y_{2i}|y_{1i}) = p^{y_{1i}y_{2i}}(1-p)^{y_{1i}(1-y_{2i})} q^{(1-y_{1i})y_{2i}} (1-q)^{(1-y_{1i})(1-y_{2i})},$$

where $p = P(y_{2i} = 1|y_{1i} = 1)$ and $q = P(y_{2i} = 1|y_{1i} = 0)$ are the misclassification parameters.

Denote the parameters $\theta_1 = (\boldsymbol{\beta}_1, \beta_2)$ and $\theta_2 = (p, q)$. Then, the implementation of the EM algorithm via parametric fractional imputation involves the following steps:

[E-step]

$$Q_1(\theta_1|\theta_1^{(t)}, \theta_2^{(t)}) = \sum_{i \in S_a} w_{ia} \left[y_{1i}(\boldsymbol{\beta}_1 \mathbf{x}_{1i} + \beta_2 x_{2i}) - \log\left\{1 + \exp(\boldsymbol{\beta}_1 \mathbf{x}_{1i} + \beta_2 x_{2i})\right\} \right]$$

$$+ \sum_{i \in S_b} w_{ib} \sum_{j=1}^{2} w_{1i}^{*(j)} \left[y_{1i}^{*(j)}(\boldsymbol{\beta}_1 \mathbf{x}_{1i} + \beta_2 x_{2i}) - \log\left\{1 + \exp(\boldsymbol{\beta}_1 \mathbf{x}_{1i} + \beta_2 x_{2i})\right\} \right]$$

and

$$Q_2(\theta_2|\hat{\theta}_1^{(t)}, \theta_2^{(t)}) = \sum_{i \in S_a} w_{ia} \sum_{j=1}^{2} w_{2i}^{*(j)} \left[y_{2i}^{*(j)} \{y_{1i} \log p + (1 - y_{1i}) \log q\} \right]$$

$$+ \sum_{i \in S_a} w_{ia} \sum_{j=1}^{2} w_{2i}^{*(j)} \left[(1 - y_{2i}^{*(j)}) \{y_{1i} \log(1-p) + (1 - y_{1i}) \log(1-q)\} \right]$$

$$+ \sum_{i \in S_b} w_{ib} \sum_{j=1}^{2} w_{1i}^{*(j)} \left[y_{1i}^{*(j)} \{y_{2i} \log p + (1 - y_{2i}) \log(1-p)\} \right]$$

$$+ \sum_{i \in S_b} w_{ib} \sum_{j=1}^{2} w_{1i}^{*(j)} \left[(1 - y_{1i}^{*(j)}) \{y_{2i} \log q + (1 - y_{2i}) \log(1-q)\} \right],$$

where $y_{ki}^{*(1)} = 1$ and $y_{ki}^{*(2)} = 0$ for $k = 1, 2$ and

$$\begin{aligned} w_{1i}^{*(j)} &= P(y_{1i}^{*(j)}|y_{2i}, \mathbf{x}_i) \\ &\propto f(y_{1i}^{*(j)}|\mathbf{x}_i) P(y_{2i}|y_{1i}^{*(j)}) \\ w_{2i}^{*(j)} &= P(y_{2i}^{*(j)}|y_{1i}, \mathbf{x}_i) \\ &= P(y_{2i}^{*(j)}|y_{1i}), \end{aligned}$$

where $\sum_j w_{1i}^{*(j)} = 1$ and $\sum_j w_{2i}^{*(j)} = 1$.

[M-step] Update θ_1 by maximizing $Q_1(\theta_1|\theta_1^{(t)}, \theta_2^{(t)})$ with respect to θ_1 and update θ_2 by maximizing $Q_2(\theta_2|\hat{\theta}_1^{(t)}, \theta_2^{(t)})$ with respect to θ_2.

The best predictor of y_{1i} of the FTF sample can be written by

$$\hat{y}_{1i}^{*} = \widehat{E}(y_{1i}|\mathbf{x}_i, y_{2i}) = \sum_{j=1}^{2} w_{1i}^{*(j)} y_{1i}^{*(j)} \tag{6}$$

and the composite estimator combining two samples can be calculated as (5) using (6).

Table 5 PCE indicator: mean estimates (standard errors) of the FFP project, mean estimates (standard errors) of the FTF project, and combined mean estimates (standard errors)

Department	FFP	FTF	Combined
San Marcos	0.558	1.165	0.563
	(0.030)	(0.038)	(0.026)
Totonicapan	0.388	0.895	0.331
	(0.030)	(0.085)	(0.028)
Quiche	0.382	1.045	0.396
	(0.030)	(0.031)	(0.026)
Huehuetenango	0.456	1.140	0.479
	(0.044)	(0.036)	(0.027)
Quetzaltenango	0.695	1.325	0.795
	(0.044)	(0.232)	(0.043)

4.2 Variance estimation of the combined estimator

For variance estimation of the combined estimator, replicate variance estimation method is applied. More precisely, we used the bootstrap method of Rao and Wu [15]. For each bootstrap dataset $D_{(b)}$, $b = 1, \ldots, B$, we can calculate estimates for the specific bootstrap sample, say $\hat{\mu}_{(b)}$. Then, the bootstrap approach computes the estimated variance of estimator $\bar{y}$ by

$$\hat{V}(\bar{y}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\mu}_{(b)} - \hat{\bar{\mu}}\right)^2,$$

where $\hat{\bar{\mu}} = B^{-1} \sum_{b=1}^{B} \hat{\mu}_{(b)}$ is the mean of B bootstrap estimates. We used $B = 500$ in this study.

4.3 Results

In this section, results of the two examples in Sect. 4.1 are presented: the PCE indicator's result is shown in Table 5 and the HHS indicator's result is shown in Table 6. Both tables contain mean estimates of the FFP project (FFP), mean estimates of the FTF project (FTF) and combined mean estimates (Combined) using the original estimate of the FFP project and the new FTF mean estimates. Also, standard errors of each mean estimate are also reported.

Mean estimates of the FFP sample and the new mean estimates of the FTF sample are combined using (5) in order to obtain the composite estimates and the result is listed in the last column of the both tables. From the results in Tables 5 and 6, we find that the combined estimator provides reasonable estimates for the population mean with smaller standard errors.

Estimates of parameters of the measurement error model for PCE variable are $(\hat{\alpha}_0, \hat{\alpha}_1) = (0.261, 0.732)$. The $\hat{\alpha}_0 = 0.261$ can be thought of as the mean of the measurement error model and it can explain why some combined estimates are outside the confidence interval of the estimate from the FTF.

In some cases, the combined estimate is not in between the FFP and the FTF. For example, the combined estimate of PCE in Totonicapan and the combined estimate of HHS in Huehuetenango are smaller than the FFP and the FTF. The new estimate of the FTF, which was adjusted for measurement errors, is even smaller than the FFP and it leads to the combined estimate that is not between the two original values. The new FTF estimate is not tabulated

Table 6 HHS indicator: proportion estimates (standard errors) of the FFP project, proportion estimates (standard errors) of the FTF project, and combined proportion estimates (standard errors) (%)

Department	FFP	FTF	Combined
San Marcos	3.76	15.35	3.77
	(1.01)	(2.22)	(1.00)
Totonicapan	11.79	15.01	12.08
	(1.70)	(6.00)	(1.60)
Quiche	7.13	9.73	7.19
	(1.50)	(1.57)	(1.42)
Huehuetenango	8.91	15.58	8.75
	(1.90)	(2.00)	(1.90)
Quetzaltenango	6.84	9.94	6.85
	(1.80)	(8.25)	(1.70)

in the result, but the new estimate of PCE in Totonicapan is 0.275 and the new one of HHS in Huehuetenango is 8.70, which are smaller than the FFP for both cases.

5 Discussion

This study suggests a new approach to combine information from two surveys using the measurement error model approach and it can be generalized to combine more than two sources of information. Using a structural equation model and a measurement error model, we present a guidance on data integration with illustration of the work sponsored by FANTA. The results shown in Tables 5 and 6 indicate that the reference estimate and the counterfactual predicted values of the other sample can be used to produce the combined estimates.

The choice of a benchmark among several surveys can be decided in various ways. We considered a smaller mean squared error as a criterion in our study. If we have auxiliary information, such as previous experiences on the surveys, it can be used to determine a gold standard among several surveys.

The proposed approach can be applied to combine more than two survey data. Similarly, we can implement the method as follows: set one survey data as a benchmark, remove measurement errors existing in the remaining survey data and calculate the composite estimator using the estimates from the surveys. Also, multivariate modeling for the structural equation model can provide a more efficient estimation. Such extension will be a topic for future research.

Acknowledgements The research was partially supposed by a grant from US National Science Foundation (Grant no MMS-1324922).

References

1. Buonaccorsi, J.P.: Measurement Error: Models, Methods, and Applications. Chapman & Hall, London (2010)
2. Dillman, D.A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., Messer, B.L.: Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (ivr) and the internet. Soc. Sci. Res. **38**(1), 1–18 (2009)

3. Fornell, C., Larcker, D.F.: Evaluating structural equation models with unobservable variables and measurement error. J. Mark. Res. **18**, 39–50 (1981)
4. Fuller, W.A.: Estimation for multiple phase samples. In: Chambers, R.L., Skinner, C.J. (eds.) Analysis of Survey Data, pp. 307–322. Wiley, Chichester (2003)
5. Fuller, W.A.: Measurement Error Models. Wiley, New York (2009)
6. Hidiroglou, M.: Double sampling. Surv. Methodol. **27**(2), 143–154 (2001)
7. Kim, J.K.: Parametric fractional imputation for missing data analysis. Biometrika **98**, 119–132 (2011)
8. Kim, J.K., Berg, E., Park, T.: Statistical matching using fractional imputation. Surv. Methodol. **42**, 19–40 (2016)
9. Kim, J.K., Rao, J.N.: Combining data from two independent surveys: a model-assisted approach. Biometrika **99**(1), 85–100 (2012)
10. Kish, L.: A procedure for objective respondent selection within the household. J. Am. Stat. Assoc. **44**(247), 380–387 (1949)
11. Legg, J.C., Fuller, W.A.: Two-phase sampling. Handb. Stat. **29**, 55–70 (2009)
12. Merkouris, T.: Combining independent regression estimators from multiple surveys. J. Am. Stat. Assoc. **99**(468), 1131–1139 (2004)
13. Merkouris, T.: Combining information from multiple surveys by using regression for efficient small domain estimation. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **72**(1), 27–48 (2010)
14. Park, S., Kim, J.K., Park, S.: An imputation approach for handling mixed-mode surveys. Ann. Appl. Stat. **10**(2), 1063–1085 (2016)
15. Rao, J.N., Wu, C.: Resampling inference with complex survey data. J. Am. Stat. Assoc. **83**(401), 231–241 (1988)
16. Renssen, R.H., Nieuwenbroek, N.J.: Aligning estimates for common variables in two or more sample surveys. J. Am. Stat. Assoc. **92**(437), 368–374 (1997)
17. USAID: Baseline study of Food For Peace Title II development food assistance program in Guatemala (2013). https://www.usaid.gov/data/dataset/beafc8ed-c5cf-41a0-84a4-19303c309516
18. Wu, C.: Combining information from multiple surveys through the empirical likelihood method. Can. J. Stat. **32**(1), 15–26 (2004)
19. Ybarra, L.M., Lohr, S.L.: Small area estimation when auxiliary information is measured with error. Biometrika **95**(4), 919–931 (2008)
20. Zieschang, K.D.: Sample weighting methods and estimation of totals in the consumer expenditure survey. J. Am. Stat. Assoc. **85**(412), 986–1001 (1990)