Design Weighted Quadratic Inference Function Estimators of Superpopulation Parameters

Sumanta Adhya, Debanjan Bhattacharjee and Tathagata Banerjee

Abstract Using information from multiple surveys to produce better pooled estimators is an active research area in recent days. Multiple surveys from same target population is common in many socioeconomic and health surveys. Often all the surveys do not contain same set of variables. Here we consider a standard situation where responses are known for all the samples from multiple surveys but the same set of covariates (or auxiliary variables) is not observed in all the samples. Moreover, in our case we consider a finite population set up where samples are drawn from multiple finite populations using same or different probability sampling designs. Here the problem is to estimate the parameters (or superpopulation parameters) of underlying regression model. We propose quadratic inference function estimator by combining information related to the underlying model from different samples through design weighted estimating functions (or score functions). We did a small simulation study for comprehensive understanding of our approach.

Keywords Model-design based approach · Multiple surveys · Superpopulation Quadratic inference function

1 Introduction

Drawing inference on super population parameters by combining data from different surveys is of considerable recent interest (Citro 2014; Kim and Rao 2012; Gelman et al. 1998) to the survey practitioners. For an up to date and comprehensive review of the methods, we refer to Lohr and Raghunathan (2016). The central idea behind any

S. Adhya (✉)
Department of Statistics, West Bengal State University, West Bengal, Barasat, India
e-mail: sumanta.adhya@gmail.com

D. Bhattacharjee
Department of Mathematics, Utah Valley University, Orem, UT, USA

T. Banerjee
Indian Institute of Management, Gujarat, Ahmedabad, India

A. K. Chattopadhyay and G. Chattopadhyay (eds.), *Statistics and its Applications*,
Springer Proceedings in Mathematics & Statistics 244,
https://doi.org/10.1007/978-981-13-1223-6_14

such method is to use information from different sources effectively for enhancing the efficiency of the estimators. In this paper, we propose a method for combining data based on quadratic inference function (QIF) (Lindsay and Qu 2003) in the context of linear regression analysis. To the best of our knowledge, use of QIF has not been considered before in the survey sampling literature.

For the methodological development in this paper, we consider model-design-based randomization approach to inference discussed in Roberts and Binder (2009), Graubard and Korn (2002), and Godambe and Thompson (1986). Specifically, we consider two finite populations $\mathcal{P}_1 = \{(y_i, x_{1i}, x_{2i}) : i \in U_1\}$ and $\mathcal{P}_2 = \{(y_i, x_{1i}) : i \in U_2\}$ of sizes N_1 and N_2, respectively, where U_1 and U_2 are index sets of the population units in $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively. Notice that $\mathcal{P}_1$ and $\mathcal{P}_2$ can be considered as random samples from a superpopulation. We assume:

(i) The study variables in each finite population are independent realizations of the random variables (y, x_1, x_2), where x_1 and x_2 are exogenous, and y is a continuous endogenous variable. Also, given x_1 and x_2, y is generated by a linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, where ϵ is the error term independent of x_1 and x_2, and has mean 0 and variance σ^2. However, in $\mathcal{P}_2$ observations on x_2 are missing.

(ii) A probability sample is selected from each resulting finite population using either the same or different sampling designs.

The above theoretical set-up may represent an important practical situation that often arises in survey sampling. Suppose in a survey with a relatively small sample size, the data are collected on a comprehensive set of exogenous variables; whereas in a different survey from the same population with a considerably larger sample size, the data are collected on a smaller subset of the same set of exogenous variables. The problem is to combine these independent samples effectively to get a better estimator.

Clearly, the problem stated above may be considered as a missing data problem where for some units in the bigger sample the data on one or more exogenous variables are missing. Multiple imputation is an often used method (Rendall et al. 2013; Gelman et al. 1998; Rubin 1986) in such situation, but how does it tide over the omitted variable bias is not quite clear. On the other contrary, the QIF based methodology that we propose here, recognizes and takes into account the omitted variable bias explicitly. Although the proposed methodology is applicable for combining data from any number of surveys in the set-up described above, we restrict our discussion to two surveys simply for ease of exposition.

The paper is organized as follows. In Sect. 2, we briefly discuss the estimation methodology based on QIF in a general setting, keeping in view the context of our application. In Sect. 3, we propose design-weighted QIF estimators of the regression coefficients using data from multiple surveys. Our methodology explicitly takes into account the omitted variable bias. In Sect. 4, we report the results of a limited simulation study. As expected, the simulation results show that the design-weighted QIF estimators based on the combined sample are substantially more efficient than the standard least squares estimators based on the sample with more covariates. Concluding remarks are given in Sect. 5.

2 Quadratic Inference Function

In this section we briefly introduce QIF based estimation methodology in a general setting. Suppose $\mathbf{b}(x, \boldsymbol{\theta}) = (b_1(x, \boldsymbol{\theta}), b_2(x, \boldsymbol{\theta}), ..., b_q(x, \boldsymbol{\theta}))^T$ is a q-dimensional vector of distinct score functions, where $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)^T$ is a p-dimensional vector of parameters. The score functions are also called estimating functions and moment conditions in statistics and economics literature, respectively. Application of QIF based estimation methodology makes sense only if q is greater than p.

Suppose $\mathcal{F}_\theta$ is the semi- parametric model defined by the parameter $\boldsymbol{\theta}$ and the score equations

$$E_F \mathbf{b}(X, \boldsymbol{\theta}) = 0, \tag{1}$$

such that if a distribution $F \in \mathcal{F}_\theta$, then (1) is satisfied and vice versa. On the other hand, if the true $F \notin \mathcal{F}_\theta$, and $E_F \mathbf{b}(X, \boldsymbol{\theta}) = \delta(\boldsymbol{\theta}) \neq 0$, where $\delta(\boldsymbol{\theta})$ is said to represent the vector of discrepancy between the model and the true distribution F.

The quadratic distance function (QDF) between the true distribution F and the semi-parametric model $\mathcal{F}_\theta$ as determined through the basic scores is then defined as

$$d(F, \mathcal{F}_{\boldsymbol{\theta}}) = \delta(\boldsymbol{\theta})^T \Sigma_{\boldsymbol{\theta}}^{-1} \delta(\boldsymbol{\theta}), \tag{2}$$

where $\Sigma_{\boldsymbol{\theta}} = Var(\mathbf{b}(X, \boldsymbol{\theta}))$. For an arbitrary F, the value of $\boldsymbol{\theta}$ for which the basic scores are closest to mean 0 is then given by

$$\boldsymbol{\theta}(F) = \arg min_{\boldsymbol{\theta}}\, d(F, \mathcal{F}_{\boldsymbol{\theta}}). \tag{3}$$

For making data based inference on $\boldsymbol{\theta}$, the QDF in (3) needs to be replaced by its empirical analogue, called quadratic inference function. Suppose $X_1, X_2, ..., X_n$ are independently and identically distributed random variables following the distribution F, then a natural estimator of $E_F \mathbf{b}(X, \boldsymbol{\theta}) = \delta(\boldsymbol{\theta})$ is $\bar{\mathbf{b}}(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} b(X_i, \boldsymbol{\theta})$. Suppose further, $\hat{\Sigma}$ is a suitably chosen estimator of $Var(\bar{\mathbf{b}}(\boldsymbol{\theta}))$, the QIF is then given by

$$Q(\boldsymbol{\theta}) = \bar{\mathbf{b}}(\boldsymbol{\theta})^T \hat{\Sigma}^{-1} \bar{\mathbf{b}}(\boldsymbol{\theta}). \tag{4}$$

The choice of $\hat{\Sigma}^{-1}$ is an important issue. We refer to Lindsay and Qu (2003) for a detailed discussion on it. The QIF estimator of is given by

$$\hat{\boldsymbol{\theta}} = \arg min_{\boldsymbol{\theta}}\, Q(\boldsymbol{\theta}). \tag{5}$$

If $F \in \mathcal{F}_\theta$, $\hat{\boldsymbol{\theta}}$ is consistent for the true value of $\boldsymbol{\theta}$, otherwise it is consistent for the nonparametric functional $\boldsymbol{\theta}(F)$ (cf.(3)). For a discussion on the optimum properties of $\hat{\boldsymbol{\theta}}$, we refer to Lindsay and Qu (2003).

3 Design-Weighted QIF Estimator

Let us now consider the estimation of the regression parameter $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$ of the superpopulation model introduced in Sect. 1. First, we introduce some important notations. Suppose $\mathbf{S}_1 = \{(y_i, x_{i1}, x_{i2}) : i \in I_1 \subset U_1\}$ and $\mathbf{S}_2 = \{(y_i, x_{i1}, x_{i2}) : i \in I_2 \subset U_2\}$ represent the probability samples of sizes $n_1 (< N_1)$ and $n_2 (< N_2)$ drawn from the populations $\mathcal{P}_1$ and $\mathcal{P}_2$ using sampling designs $p_1(.)$ and $p_2(.)$, respectively, where I_1 and I_2 are index sets of selected sample units.

As stated at the outset, we adopt the model-design based randomization approach (Roberts and Binder 2009) to the estimation of the superpopulation parameters. Like Chen and Sitter (1999), we propose a two-step design weighted QIF estimator of $\boldsymbol{\beta}$ that could be used for complex surveys. First, we define QIF of $\boldsymbol{\beta}$, say, $Q_U(\boldsymbol{\beta})$, assuming $\mathcal{P}_1$ and $\mathcal{P}_2$ to be known. At the second step, we estimate $Q_U(\boldsymbol{\beta})$ by replacing the population based entities with its design-based estimators based on the samples. We denote it by $\widetilde{Q}_U(\boldsymbol{\beta})$. Finally, the estimator of $\boldsymbol{\beta}$ is obtained by minimizing $\widetilde{Q}_U(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. We now describe the two steps in detail.

Assuming $\mathcal{P}_1$ to be known, and represents a random sample from the superpopulation, the basic score vector for $\boldsymbol{\beta}$ is given by:

$$\mathbf{b}_1(y, \mathbf{x}, \boldsymbol{\beta}) = (Y - \beta_0 - \beta_1 x_1 - \beta_2 x_2)\mathbf{x}, \tag{6}$$

where $\mathbf{x} = (1, x_1, x_2)^T$. Also, the assumed regression model of y given x_1 and x_2 entails $E_{\boldsymbol{\beta}}\mathbf{b}_1(Y, \mathbf{X}, \boldsymbol{\beta}) = 0$. However, for $\mathcal{P}_2$, the basic score function for $\boldsymbol{\beta}^{(1)} = (\beta_0, \beta_1)^T$ is given by:

$$\mathbf{b}_2^*(y, \mathbf{x}^{(1)}, \boldsymbol{\beta}^{(1)}) = (Y - \beta_0 - \beta_1 x_1)\mathbf{x}^{(1)}, \tag{7}$$

where $\boldsymbol{\beta}^{(1)} = (\beta_0, \beta_1)^T$ and $\mathbf{x}^{(1)} = (1, x_1)^T$. But omitted variable bias leads to $E_{\boldsymbol{\beta}}\mathbf{b}_2^*(Y, \mathbf{X}^{(1)}\boldsymbol{\beta}^{(1)}) = \boldsymbol{\delta}(\boldsymbol{\beta_2})$, where $\boldsymbol{\delta}(\boldsymbol{\beta_2}) = (0, \beta_2\sigma_{12})^T$, and $\sigma_{12} = Cov(x_1, x_2)$. Assuming σ_{12} to be known for the time being, we define a modified score function for $\boldsymbol{\beta}$ that explicitly takes into account the omitted variable bias as follows:

$$\mathbf{b}_2(y, \mathbf{x}^{(1)}, \boldsymbol{\beta}) = (y - \beta_0 - \beta_1 x_1)\mathbf{x}^{(1)} - \boldsymbol{\delta}(\beta_2). \tag{8}$$

Thus, by definition, we have $E_{\boldsymbol{\beta}}\mathbf{b}_2(Y, \mathbf{X}^{(1)}, \boldsymbol{\beta}) = 0$. The population version of QIF are thus based on the basic score functions given by (6) and (8).

Let us define $\bar{\mathbf{b}}_1(\boldsymbol{\beta}) = N_1^{-1}\sum_{i\in U_1}\mathbf{b}_1(y_i, \mathbf{x}_i, \boldsymbol{\beta})$, $\bar{\mathbf{b}}_2(\boldsymbol{\beta}) = N_2^{-1}\sum_{i\in U_2}\mathbf{b}_2(y_i, \mathbf{x}_i^{(1)}, \boldsymbol{\beta})$, and $\bar{\mathbf{b}}(\boldsymbol{\beta}) = (\bar{\mathbf{b}}_1(\boldsymbol{\beta}), \bar{\mathbf{b}}_2(\boldsymbol{\beta}))^T$. Let $\hat{\Sigma}_{1\boldsymbol{\beta}}$, $\hat{\Sigma}_{2\boldsymbol{\beta}}$, and $\hat{\Sigma}_{\boldsymbol{\beta}}$ be suitable finite population based estimators of $Var(\mathbf{b}_1(Y, \mathbf{X}, \boldsymbol{\beta})) = \Sigma_{1\beta}$, $Var(\mathbf{b}_2(Y, \mathbf{X}^{(1)}, \boldsymbol{\beta})) = \Sigma_{2\beta}$ and $Var(\mathbf{b}(Y, \mathbf{X}, \boldsymbol{\beta})) = \Sigma_\beta$, respectively, where $\mathbf{b}(y, \mathbf{x}, \boldsymbol{\beta}) = (\mathbf{b}_1(y, \mathbf{x}, \boldsymbol{\beta}), \mathbf{b}_2(y, \mathbf{x}^{(1)}, \boldsymbol{\beta}))^T$.

Then the first-step QIF of $\boldsymbol{\beta}$ is given by

$$Q_U(\boldsymbol{\beta}) = W_1\bar{\mathbf{b}}_1(\boldsymbol{\beta})^T\hat{\Sigma}_{1\boldsymbol{\beta}}^{-1}\bar{\mathbf{b}}_1(\boldsymbol{\beta}) + W_2\bar{\mathbf{b}}_2(\boldsymbol{\beta})^T\hat{\Sigma}_{2\boldsymbol{\beta}}^{-1}\bar{\mathbf{b}}_2(\boldsymbol{\beta}), \tag{9}$$

where, $W_k = N_k N^{-1}, k = 1, 2$, and $N = N_1 + N_2$.

Let us now define the second step QIF, $\widetilde{Q}_U(\boldsymbol{\beta})$, an estimator of $Q_U(\boldsymbol{\beta})$, based on the samples $\mathbf{S}_1$ and $\mathbf{S}_2$. Suppose $\pi_{ik} = P_k(i \in I_k | i \in U_k)(> 0)$ denotes the inclusion probability of the $i-th$ unit of the $k-$th population in the sample $\mathbf{S}_k$, where $P_k(.)$ is the probability measure corresponding to the sampling design $p_k(.)$ for $i = 1, 2, ..., N_k, k = 1, 2$. The design weights are then given by $d_{ik} = \frac{\pi_{ik}^{-1}}{\sum_{i \in S_k} \pi_{ik}^{-1}}$, for $i \in I_k, k = 1, 2$. Defining, $\widetilde{\mathbf{b}}_{i1}(\boldsymbol{\beta}) = \mathbf{b}_1(y_i, \mathbf{x}_i, \boldsymbol{\beta})$ for $i \in I_1$, $\widetilde{\mathbf{b}}_{i2}(\boldsymbol{\beta}) = \mathbf{b}_1(y_i, \mathbf{x}_i^{(1)}, \boldsymbol{\beta})$ for $i \in I_2$, $\widetilde{\mathbf{b}}_1(\boldsymbol{\beta}) = \sum_{i \in I_1} d_{i1} \widetilde{\mathbf{b}}_{i1}(\boldsymbol{\beta})$, $\widetilde{\mathbf{b}}_2(\boldsymbol{\beta}) = \sum_{i \in I_2} d_{i2} \widetilde{\mathbf{b}}_{i2}(\boldsymbol{\beta})$, and $\widetilde{\Sigma}_{k\boldsymbol{\beta}} = \sum_{i \in I_k} d_{ik} (\widetilde{\mathbf{b}}_{ik}(\boldsymbol{\beta}) - \widetilde{\mathbf{b}}_k(\boldsymbol{\beta}))(\widetilde{\mathbf{b}}_{ik}(\boldsymbol{\beta}) - \widetilde{\mathbf{b}}_k(\boldsymbol{\beta}))^T$ for $k = 1, 2$, we obtain

$$\widetilde{Q}_U(\boldsymbol{\beta}) = W_1 \widetilde{\mathbf{b}}_1(\boldsymbol{\beta})^T \widetilde{\Sigma}_{1\boldsymbol{\beta}}^{-1} \widetilde{\mathbf{b}}_1(\boldsymbol{\beta}) + W_2 \widetilde{\mathbf{b}}_2(\boldsymbol{\beta})^T \widetilde{\Sigma}_{2\boldsymbol{\beta}}^{-1} \widetilde{\mathbf{b}}_2(\boldsymbol{\beta}). \tag{10}$$

The design-weighted QIF estimator of $\boldsymbol{\beta}$ is then given by

$$\hat{\boldsymbol{\beta}} = \arg min_{\boldsymbol{\beta}} \widetilde{Q}(\boldsymbol{\beta}). \tag{11}$$

Notice that throughout the development we assume σ_{12} to be known. It may be a reasonable assumption if the information on x_1 and x_2 are available at the population level while the values of (y, x_1, x_2) are known for the sample only. In this case, the design-weighted QIF estimators lead to a huge improvement over the standard least squares estimators. In case, it is not known, we plug in its estimate from the sample in $\widetilde{Q}_U(\boldsymbol{\beta})$. The latter also shows some improvement as is evident from the numerical studies reported in the next section.

4 Numerical Studies

We present the results of a limited simulation study comparing the performances of design-weighted quadratic inference function estimator (QIFE) with that of design-weighted least square estimator (LSE).

Suppose the covariate vector $(x_1, x_2)^T$ has a bivariate normal distribution with mean vector $(0, 0)^T$ and covariance matrix $\boldsymbol{\Sigma}(2 \times 2)$. Given (x_1, x_2), y has a normal distribution with mean $1 + 0.5x_1 + 0.25x_2$ and variance 0.25. We consider two superpopulation models $M1$ and $M2$ corresponding to two choices of $\boldsymbol{\Sigma}$, say, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively, where

$$\Sigma_1 = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 1.0 \end{pmatrix}$$

and

$$\Sigma_1 = \begin{pmatrix} 0.5 & 0.14 \\ 0.14 & 1.0 \end{pmatrix}.$$

Notice that for model $M1$ the correlation coefficient between x_1 and x_2 is 0.7 while for $M2$, it is 0.2.

Following are the steps of the simulation study:

Step 1: We generate finite populations U_1 and U_2 of sizes N_1 and N_2 using the above superpopulation model. First, we randomly generate a value of $\mathbf{x} = (x_1, x_2)^T$, and then generate a value of y given $\mathbf{x}$ using the conditional distribution of y given $\mathbf{x}$. The finite populations U_1 and U_2 then comprise N_1 and N_2 such observations on (y, x_1, x_2) generated independently. Next, by simple random sampling without replacement (SRSWOR), we select L samples of sizes $n_1 (= f_1 N_1)$ and $n_2 (= f_2 N_2)$ from U_1 and U_2, respectively, where f_1 and f_2 are the sampling fractions. The selected samples from U_1 and U_2 are denoted by $\mathbf{S}_1^{(l)}$ and $\mathbf{S}_2^{(l)}$, $l = 1, 2, ..., L$ respectively.

Step 2: Based on $\mathbf{S}_1$ we compute usual design-weighted LSE of $\boldsymbol{\beta}$. Also based on $\mathbf{S}_1$ and $\mathbf{S}_2$, we compute design-weighted QIFE from (11).

Step 3: We repeat the Step 1 R times. At the r-th ($r = 1, 2, ..., R$) replication, let the populations generated be $U_1^{(r)}$ and $U_2^{(r)}$. For each r, the selected samples from $U_1^{(r)}$ and $U_2^{(r)}$ are denoted by $\mathbf{S}_1^{(rl)}$ and $\mathbf{S}_2^{(rl)}$, $l = 1, 2, ..., L$, respectively. For each r and l, following Step 2, we compute the LSE and QIFE of β_j, $j = 0, 1, 2$, say, $\hat{\beta}_{j(LS)}^{(rl)}$ and $\hat{\beta}_{j(QIF)}^{(rl)}$, respectively.

Step 4: For each estimator of β_j, say, $\hat{\beta}_j^{(rl)}$ (a generic notation) we compute the relative bias (RB) ($[(RL)^{-1} \sum_{r,l} \hat{\beta}_j^{(rl)} - \beta_j]/|\beta_j|$) and relative root mean squared error (RRMSE) ($\sqrt{(RL)^{-1} \sum_{r,l} (\hat{\beta}_j^{(rl)} - \beta_j)^2}/|\beta_j|$).

For our simulation study, we consider (N_1, N_2): (1000, 2000), (1000, 5000), $R = L = 100$ and $f_1 = f_2 = 0.10$. In Table 1, we report the RRMSE values for the LSE's and QIFE's of β_j, $j = 0, 1, 2$. The RB values are not shown. However, it has been observed that for $n_1 = 100, n_2 = 500$, i.e., when the second sample size is relatively large compared to the first, the relative biases of both the estimators are comparable. For $n_1 = 100, n_2 = 200$ the relative bias of QIFE is slightly higher than LSE. This is expected as LSE is unbiased while QIFE is not. What is interesting to observe, that with increase in the relative magnitude of N_2 compared to N_1, the performances of QIFE's of β_j, $j = 0, 1$ improve over the LSE's substantially. Also the improvement is more if the correlation between x_1 and x_2 increases. The performances of QIFE and LSE of β_2 are more or less same.

5 Concluding Remarks

In this article we propose quadratic inference function estimator of the superpopulation parameters using information from multiple samples from the same superpopulation that incorporates the design weights. For illustrative purpose, in this paper,

Table 1 RRMSE of the least squares (LS) and quadratic inference function (QIF) estimators of the superpopulation parameters for models **M1** and **M2**

Regression coefficient	Model M1		Model M2	
	LSE	QIFE	LSE	QIFE
N1 = 1000 N2 = 2000				
β_0	502	316	507	313
β_1	2004	1810	1470	1051
β_2	2845	2902	2063	2092
N1 = 1000 N2 = 5000				
β_0	507	223	511	226
β_1	2046	1712	1485	928
β_2	2827	2898	2113	2147

we have considered linear regression superpopulation model. Our design-adjusted QIF estimator is appealing in the sense that it can be applied for complex survey designs. The simulation study shows encouraging results in situations where size of the sample containing observations on subset of covariates is very high. In future we plan to investigate the asymptotic properties of the proposed QIF estimator under complex survey designs.

References

Chen, J., & Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, *9*, 385–406.

Citro, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, *40*, 137–161.

Gelman, A., King, G., & Liu, C. (1998). Not asked and not answered: Mulitple imputation for multiple surveys. *Journal of the American Statistical Association*, *93*, 847–857.

Godambe, V. P., & Thompson, M. E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review*, *54*, 127138.

Graubard, B. I., & Korn, E. L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science*, *17*, 73–96.

Kim, J. K., & Rao, J. N. K. (2012). Combining data from two independent surveys: A model assisted approach. *Biometrika*, *99*, 85–100.

Lindsay, B. G., & Qu, A. (2003). Inference functions and quadratic score tests. *Statistical Science*, *18*, 394–410.

Lohr, S. L., & Raghunathan, T. E. (2016). Combining survey data with other data sources. *Statistical Science*, *32*, 293–312.

Rendall, M. S., Ghosh-Dastidar, B., Weden, M. M., Baker, E. H., & Nazarov, Z. (2013). Multiple imputation for combined-survey estimation with incomplete regressors in one but not both surveys. *Social Methods and Research*, *42*, 483–530.

Roberts, G., & Binder, D. (2009). Analyses based on combining similar information from multiple surveys. In *JSM: Section on Survey Methods*.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, *21*, 6573.