The Annals of Applied Statistics
2016, Vol. 10, No. 2, 1063–1085
DOI: 10.1214/16-AOAS930
© Institute of Mathematical Statistics, 2016

AN IMPUTATION APPROACH FOR HANDLING MIXED-MODE SURVEYS

BY SEUNGHWAN PARK[1], JAE KWANG KIM[2] AND SANGUN PARK[3]

Seoul National University, Iowa State University and Yonsei University

Mixed-mode surveys are becoming more popular recently because of their convenience for users, but different mode effects can complicate the comparability of the survey results. Motivated by the Private Education Expenditure Survey (PEES) of Korea, we propose a novel application of fractional imputation to handle mixed-mode survey data. The proposed method is applied to create imputed values of the unobserved counterfactual outcome variables in the mixed-mode surveys. The proposed method is directly applicable when the choice of survey mode is self-selected. Variance estimation using Taylor linearization is developed. Results from a limited simulation study are also presented.

1. Introduction.

Surveys can be conducted in many different modes, including online, telephone, mail and face-to-face, and each survey mode has its unique effect on the survey responses. Researchers often utilize more than one survey mode in order to increase the participation rate and improve the coverage of the survey population. For this reason, mixed-mode surveys, which rely on a combination of survey modes, are becoming increasingly popular in practice. Dillman and Christian (2003) and de Leeuw (2005) discussed several advantages of the mixed-mode survey. However, since each survey mode has a different mode effect and it is often confounded with the selection effect, it is challenging to compare data in mixed-mode surveys [Krosnick (1991, 1999), Voogt and Saris (2005), Dillman et al. (2009)]. When the selection effect is ignorable as in the completely randomized design, the mode effect can be estimated under some comparability assumptions [Vannieuwenhuyze, Loosveldt and Molenberghs (2010), Buelens and Van den Brakel (2013)].

From a statistical point of view, the mixed-mode survey can be regarded as a measurement error problem with different measurement errors representing different survey modes. In this sense, statistical adjustment of mode effect is essentially a

Received July 2014; revised March 2016.

[1] Supported in part by the National Research Foundation of Korea (NRF) grant founded by the Korea government (MSIP) (No. 2011–0030811).

[2] Supported in part by a Grant from NSF (MMS-1324922).

[3] Supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011–0029104).

Key words and phrases. Counterfactual outcome, fractional imputation, measurement error model, missing data, survey sampling.

calibration problem under measurement error models. Durrant and Skinner (2006) considered the calibration problem under measurement errors using a validation subsample including both the erroneously measured variable and the accurately measured variable. Powers, Mishra and Young (2005) used the multiple imputation method for measurement error calibration in the National Health Survey of Family Growth. Burgette and Reiter (2012) used nonparametric Bayesian multiple imputation methods to calibrate one measurement to another when the measurement methods are changed during the data collection.

In the mixed-mode surveys, we cannot directly apply the classical calibration approach because of the absence of a calibration subsample. Biemer (2001) proposed an interview-reinterview approach using the latent class analysis to estimate the selection effect and the measurement effect in a mixed-mode survey. Klausch, Schouten and Hox (2014) used a within-subject design to estimate the measurement effect in the same mixed-mode survey considered in Buelens and Van den Brakel (2013), but their approach is only applicable to sequential mixed-mode surveys. Kolenikov and Kennedy (2014) provide a nice summary of three existing methods for handling mixed-mode surveys. Vannieuwenhuyze, Loosveldt and Molenberghs (2010) considered the covariate adjustment method to explain the measurement effect in estimating the population mean in mixed-mode surveys.

In this paper, we consider an imputation approach for estimating parameters of interest under a measurement error model in a single survey. This research is motivated by the 2011 Private Education Expenditure Survey (PEES) conducted by Statistics Korea. The PEES aims to measure the average private education expenses for elementary, middle and high school students in Korea, respectively. The PEES had used mail surveys only for obtaining responses before 2011, and thus the mail survey has been chosen as the reference mode. In 2011, the PEES used two survey modes to obtain responses, via mail survey and via internet survey, and the choice between the two survey modes is completely randomized.

The private education expense variable, denoted as Cost, is the primary study variable considered. There are significant differences in the study variable and the time spent on private education, denoted as Time, between two survey modes, notably the smaller mean and the larger standard deviations for internet surveys (Tables 1 and 2). Also, in comparison with data from the reference mode, internet

TABLE 1

Means and standard deviations of the private education expenses in 2011 *PEES*

School level	**Mail**		**Internet**	
Elementary School	72.071	(60.422)	68.479	(59.544)
Middle School	82.891	(71.257)	83.395	(81.275)
High School	79.980	(92.116)	74.726	(95.802)

Standard deviations are in parentheses.

TABLE 2
Means and standard deviations of time spent on private education in 2011 *PEES*

School level	**Mail**		**Internet**	
Elementary School	7.868	(5.840)	7.657	(6.332)
Middle School	7.946	(6.637)	7.159	(6.672)
High School	4.757	(5.455)	4.331	(5.497)

Standard deviations are in parentheses.

survey data has a tendency to have more extreme values; the portion of students taking no private lessons or tutoring is much higher in the internet survey (Table 3) and comparatively larger values of private education expense exist in the data from the internet survey.

In addition to the study variable, there are auxiliary variables, **x**, that include information about local area level, school level (Elementary, Middle, High), gender of students, age of parents, education level of parents, grade of students and household income. Since in the 2011 PEES respondents were randomly assigned to each mode, two survey modes produce similar compositions of the respondents with respect to auxiliary variables. For example, Table 4 shows that there is no significant differences between mail and internet mode in the percent of male students, percent of parents with less than or equal to 12 years of education and monthly household income. Thus, we can safely assume that **x** is not subject to measurement errors resulting from survey modes.

Moreover, PEES has high response rates. All parents responded to the study variable, Cost, and to most of the auxiliary variables. For the education level of parents, nonresponse rates are approximately 4% for both survey modes.

Obtaining consistent estimation results from the reference mode, or the mode that the survey has been offered in historically, is a critical issue for the survey provider [Kolenikov and Kennedy (2014)]. In our case, PEES had been conducted only by mail until 2011, which makes the mail survey our reference mode. Thus, our goal in the project is to convert the responses from the internet survey into those from the mail survey.

TABLE 3
Percent of students taking private lessons or tutoring in 2011 *PEES*

School level	**Mail**	**Internet**
Elementary School	86.1	73.4
Middle School	76.8	71.7
High School	63.7	56.7

TABLE 4

Description of auxiliary variables in 2011 *PEES*: *The second and third columns show the percent of male students from mail and internet surveys*, *respectively. The fourth and fifth columns indicate the percent of parents with less than or equal to* 12 *years of education from mail and internet surveys*, *respectively. The sixth and seventh columns show the mean of monthly household income from mail and internet surveys*, *respectively*

	Percent of male students		**Percent of parents with ≤12 years of education**		**Monthly household income***	
School level	**Mail**	**Internet**	**Mail**	**Internet**	**Mail**	**Internet**
Elementary School	52.59	53.05	39.54	36.21	4.28	4.23
Middle School	55.87	51.12	46.62	45.12	4.33	4.34
High School	52.88	53.85	44.51	44.45	4.50	4.36

*Million KRW.

We use a measurement error model setup to handle the mixed-mode survey data. We write the measurement from mode A and mode B as y_a and y_b, respectively. Also, we observe demographic variables (such as grade, gender, geography, education of parents) denoted by x. Table 5 presents the data structure for the mixed-mode survey data. As can be seen in Table 5, (X, Y_a) is observed in mode A and (X, Y_b) is observed in mode B. Thus, Y_a and Y_b are never jointly observed. Therefore, Y_a is an unobservable counterfactual outcome of Y_b in the mode B sample and the same can be said of Y_b with respect to the mode A sample [Morgan and Winship (2007), Vannieuwenhuyze, Loosveldt and Molenberghs (2010)]. The goal of the PEES project is to create imputed values of Y_a in the mode B sample so that $\theta = E(Y_a)$ can be estimated using the whole sample. We use a novel application of the fractional imputation method of Kim (2011) to achieve the goal.

This paper is organized as follows. In Section 2, basic setup is described in the context of a mixed-mode survey with two survey modes. In Section 3, the proposed method is presented. In Section 4, results from a limited study are presented. Application of the proposed method to PEES is described in Section 5. Concluding remarks are made in Section 6.

TABLE 5

Data structure for a mixed-mode survey with two survey modes

Mode	X	Y_a	Y_b
A	o	o	
B	o		o

2. Basic setup. In this section, we introduce some notation and state the problem in the context of a mixed-mode survey with two survey modes. Suppose that we have a probability sample S drawn from a finite population of size N, where N is known. Let π_i be the first order inclusion probability of unit i and $w_i = \pi_i^{-1}$ be the design weight associated with unit i.

Next, suppose that a mixed-mode survey with two survey modes, mode A and mode B, is used to collect information from the sample S. Let $S = S_a \cup S_b$ be the partition of the sample based on the mode such that mode A is used in S_a and mode B in S_b. For each unit i in S, let y_{ai} be the measurement of study variable y from mode A and y_{bi} be the measurement of y from mode B. In addition, suppose a q-dimensional vector of auxiliary variables $\mathbf{x}_i$ is also observed throughout the sample.

As discussed in the Introduction, we use a measurement error model setup to handle the mixed-mode survey data. A measurement error model can be written as

$$y_{bi} = \alpha_0 + \alpha_1 y_{ai} + u_i \tag{2.1}$$

for some α_0 and α_1, where $u_i \sim (0, \sigma_u^2)$. We treat mode A as the benchmark mode and focus on converting the data collection under mode B to mode A because PEES had been exclusively conducted in mode A (mail survey) until 2011. Thus, we write $y_{ai} = y_i$, where y_i is the measurement of the study variable y for unit i from the reference mode.

Now, the measurement error model (2.1) is assumed to be parametrically modeled as

$$y_{bi} | (y_i, \mathbf{x}_i) \sim g(y_{bi} | y_i; \alpha) \tag{2.2}$$

for some α with known density $g(\cdot)$. Model (2.2) also implies that

$$f(y_{bi} | \mathbf{x}_i, y_i) = f(y_{bi} | y_i), \tag{2.3}$$

which means that y_b is conditionally independent of $\mathbf{x}$, conditional on y. In the measurement error literature, variable y_{bi} satisfying (2.3) is often called the surrogate variable for y_i [Carroll et al. (2006), Section 2.5].

We also assume that y_i in the sample follows a parametric model with density $f(y_i | \mathbf{x}_i; \theta)$, where θ is an unknown parameter that characterizes the conditional distribution; that is, we assume

$$y_i | \mathbf{x}_i \sim f(y_i | \mathbf{x}_i; \theta) \tag{2.4}$$

for some θ. Models (2.2) and (2.4) form a set of measurement error models. Model (2.2), often called the measurement model, describes the relationship between the two modes, and model (2.4), sometimes called the structural error model, incorporates extra information from $\mathbf{x}_i$.

We are interested in estimating the finite population mean of the study variable, $\psi_N = N^{-1} \sum_{i=1}^N y_i$, under mode A. For a single-mode survey data (Mode A only), the Horvitz–Thompson (HT) estimator, $\hat{\psi}_{\mathrm{HT}} = N^{-1}\sum_{i \in S} w_i y_{ai}$ is an unbiased estimator of ψ_N. Under the mixed-mode survey structure, a naive estimator given by

$$\hat{\psi}_{\mathrm{naive}} = N^{-1}\left\{\sum_{i \in S_a} w_i y_{ai} + \sum_{i \in S_b} w_i y_{bi}\right\} \tag{2.5}$$

is biased unless $E(y_{ai}) = E(y_{bi})$. As can be seen in Table 1, the two survey modes are significantly different in their means and the naive estimator is biased.

To correct for the bias of the naive estimator in (2.5), we consider

$$\hat{\psi} \equiv N^{-1}\left\{\sum_{i \in S_a} w_i y_i + \sum_{i \in S_b} w_i E(y_i | y_{bi}, \mathbf{x}_i)\right\}, \tag{2.6}$$

where the conditional expectation is computed from a prediction model, obtained by the Bayes formula

$$f(y_i | y_{bi}, \mathbf{x}_i) = f(y_i | \mathbf{x}_i) \frac{g(y_{bi} | y_i)}{\int f(y_i | \mathbf{x}_i) g(y_{bi} | y_i)\, dy_i} \tag{2.7}$$

for the units in S_b. The prediction model will be the model for imputing the unobserved outcome. The actual computation of the conditional expectation in (2.6) can be implemented using an application of the parametric fractional imputation of Kim (2011), which will be presented in Section 3.

In addition to the measurement error model, the choice model (or selection model), $P(m_i = a | \mathbf{x}_i, y_i)$, where $m_i = a$ indicates A to be the mode of choice for unit i, may be considered. The choice model is particularly needed if the choice of the survey mode is not random. In this case, the conditional distribution (2.7) is changed, by the Bayes formula again, to

$$\begin{aligned} &f(y_i | y_{bi}, \mathbf{x}_i, m_i = b) \\ &\quad = f(y_i | \mathbf{x}_i) \frac{g(y_{bi} | y_i) P(m_i = b | \mathbf{x}_i, y_i)}{\int f(y_i | \mathbf{x}_i) g(y_{bi} | y_i) P(m_i = b | \mathbf{x}_i, y_i)\, dy_i}. \end{aligned} \tag{2.8}$$

If the choice model satisfies

$$P(m = b | \mathbf{x}, y) = P(m = b | \mathbf{x}),$$

then the choice model is ignorable in the sense of Rubin (1976) and model (2.8) reduces to model (2.7).

3. Proposed method.

The proposed imputation approach aims to generate $y_i = y_{ai}$ for $i \in S_b$ from the current observation $\mathbf{x}_i$ and y_{bi}. For the PEES, to have comparability with historical data, we wish to correct observations from the internet mode to observations from the mail mode.

We first consider the case when the choice of mode is ignorable. The imputation model $f(y_a | \mathbf{x}, y_b)$ can be used to create imputed values of y_{ai} from $i \in S_b$. Recall that the imputation model $f(y_a|\mathbf{x}, y_b)$ is obtained by applying a Bayes formula to the components of the measurement error models, (2.2) and (2.4); that is,

$$f(y_{ai}|\mathbf{x}_i, y_{bi}) = \frac{f(y_{ai}|\mathbf{x}_i; \theta)g(y_{bi}|y_{ai}; \alpha)}{\int f(y_{ai}|\mathbf{x}_i; \theta)g(y_{bi}|y_{ai}; \alpha)\, dy_{ai}},$$

which depends on unknown parameters θ and α. The EM algorithm can be used to estimate the parameters and predict the unobserved outcome variable simultaneously. In the EM algorithm, computation under the E-step often involves heavy computational tools such as Markov Chain Monte Carlo [Chen, Shao and Ibrahim (2000)], and its convergence is difficult to check [Booth and Hobert (1999)].

For general parametric models, $f(y_{ai}|\mathbf{x}_i; \theta)$ and $g(y_{bi}|y_{ai}; \alpha)$, the conditional expectation in the E-step does not have a closed form and Parametric Fractional Imputation (PFI) proposed by Kim (2011) can be used in this case. By introducing so-called fractional weights, the PFI method simplifies the computation in the E-step of the EM algorithm. Unlike the usual Monte Carlo EM algorithm [Wei and Tanner (1990)], the EM sequence from the PFI method converges even for fixed M, where M is the size of Monte Carlo samples in the PFI method. The EM algorithm using the PFI method is computed by the following steps:

Step 1. Set $t = 0$. Calculate the maximum likelihood estimate of the parameter θ of $f(y_{ai}|\mathbf{x}_i; \theta)$ using data S_a only. Let the estimate, denoted as $\hat{\theta}^{(0)}$, be the initial value of θ.

Step 2; generating imputed values. For each unit $i \in S_b$, generate M imputed values, $y_{ai}^{*(1)}, \dots, y_{ai}^{*(M)}$ from $f(y_{ai}|\mathbf{x}_i; \hat{\theta}^{(0)})$. Set $w_{ij(0)}^{*} = 1/M$.

Step 3; updating the parameters. Update $\hat{\theta}$ by solving the imputed score equation for θ:

$$\sum_{i \in S_a} w_i S_1(\theta; \mathbf{x}_i, y_{ai}) + \sum_{i \in S_b} \sum_{j=1}^{M} w_i w_{ij(t)}^{*} S_1(\theta; \mathbf{x}_i, y_{ai}^{*(j)}) = 0,$$

where $S_1(\theta; \mathbf{x}_i, y_{ai}) = \partial \log f(y_{ai}|\mathbf{x}_i; \theta)/\partial\theta$.

Also, update $\hat{\alpha}$ by solving the imputed score equation for α:

$$\sum_{i \in S_b} \sum_{j=1}^{M} w_i w_{ij(t)}^{*} S_2(\alpha; y_{ai}^{*(j)}, y_{bi}) = 0,$$

where $S_2(\alpha; y_{ai}, y_{bi}) = \partial \log g(y_{bi}|y_{ai}; \alpha)/\partial\alpha$.

Step 4; calculating the weights. Calculate the fractional weight w_{ij}^{*} for each $i \in S_b$,

$$w_{ij(t)}^{*} \propto g(y_{bi}|y_{ai}^{*(j)}; \hat{\alpha}^{(t)}) \frac{f(y_{ai}^{*(j)}|x_i; \hat{\theta}^{(t)})}{f(y_{ai}^{*(j)}|x_i; \hat{\theta}^{(0)})}$$

and $\sum_{j=1}^{M} w_{ij(t)}^{*} = 1$, where $\hat{\eta}^{(t)} = (\hat{\theta}^{(t)}, \hat{\alpha}^{(t)})'$ is the current estimate of $\eta = (\theta, \alpha)'$.

Step 5. Check for convergence. If converged, stop. Otherwise, set $t = t + 1$ and go to Step 3.

If the choice model is nonignorable, however, the probability of the choice mode depends on y_a, then the imputation model becomes

$$\begin{aligned} & f(y_{ai} | \mathbf{x}_i, y_{bi}, m = b) \\ & \quad = \frac{f(y_{ai} | \mathbf{x}_i; \theta) g(y_{bi} | y_{ai}; \alpha) P(m_i = b | \mathbf{x}_i, y_{ai}; \phi)}{\int f(y_{ai} | \mathbf{x}_i; \theta) g(y_{bi} | y_{ai}; \alpha) P(m_i = b | \mathbf{x}_i, y_{ai}; \phi) \, dy_{ai}}, \end{aligned}$$

and the PFI method should incorporate the choice mechanism explicitly. Here, the probability of the choice mode can be modeled by a logistic regression model, for example, and ϕ can be understood as the regression coefficients in the logistic regression model. The detailed algorithm of the PFI under the nonignorable choice model is presented in Appendix A.

Using the imputed values for units in S_b generated by the PFI method, one can compute the conditional expectation in (2.6) by a Monte Carlo approximation. Thus, the parametric fractional imputation estimator of the finite population mean is computed by

$$\hat{\psi}_{\mathrm{PFI}} = N^{-1} \left\{ \sum_{i \in S_a} w_i y_{ai} + \sum_{i \in S_b} w_i \sum_{j=1}^{M} w_{ij}^* y_{ai}^{*(j)} \right\}.$$

So far, we have only considered the case of $\psi = E(Y)$. More generally, suppose that the parameter of interest ψ is now defined by solving an estimating equation $U_N(\psi) \equiv \sum_{i=1}^{N} U(\psi; \mathbf{x}_i, y_{ai}) = 0$. Let $\eta = (\theta, \alpha)'$. Under the current setup, the PFI estimator of ψ can be obtained by solving the fractionally imputed estimating equation $\bar{U}^*(\psi | \eta) = 0$, where

$$\bar{U}^*(\psi | \eta) \equiv \sum_{i \in S_a} w_i U(\psi; \mathbf{x}_i, y_{ai}) + \sum_{i \in S_b} w_i \sum_{j=1}^{M} w_{ij}^* U(\psi; \mathbf{x}_i, y_{ai}^{*(j)}). \tag{3.1}$$

We now discuss variance estimation of the PFI estimators. Let $\hat{\eta} = (\hat{\theta}, \hat{\alpha})$ be the solution to

$$\begin{aligned} \bar{S}^*(\eta) &\equiv \begin{pmatrix} \sum_{i \in S_a} w_i S_1(\theta; \mathbf{x}_i, y_{ai}) + \sum_{i \in S_b} \sum_{j=1}^{M} w_i w_{ij}^* S_1(\theta; \mathbf{x}_i, y_{ai}^{*(j)}) \\ \sum_{i \in S_b} \sum_{j=1}^{M} w_i w_{ij}^* S_2(\alpha; y_{ai}^{*(j)}, y_{bi}) \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \end{aligned} \tag{3.2}$$

where $S_1(\theta; \mathbf{x}_i, y_{ai}) = \partial \log f(y_{ai}|\mathbf{x}_i; \theta)/\partial\theta$ and $S_2(\alpha; y_{ai}, y_{bi}) = \partial \log g(y_{bi}|y_{ai}; \alpha)/\partial\alpha$ and $\hat{\psi}_{\mathrm{PFI}}$ is the estimator obtained by solving the imputed estimating equation, $\bar{U}^*(\psi|\eta) = 0$. Note that $\bar{U}^*(\psi|\eta)$ in (3.1) can be written

$$\bar{U}^*(\psi|\eta) = \sum_{i\in S} w_i \bar{U}_i^*(\psi|\eta),$$

where

$$\bar{U}_i^*(\psi|\eta) = \begin{cases} U(\psi; \mathbf{x}_i, y_{ai}), & \text{for unit } i \in S_a, \\ \sum_{j=1}^{M} w_{ij}^* U(\psi; \mathbf{x}_i, y_{ai}^{*(j)}), & \text{for unit } i \in S_b, \end{cases}$$

and that $\bar{S}^*(\eta)$ in (3.2) can be written

$$\bar{S}^*(\eta) = \sum_{i\in S} w_i \bar{S}_i^*(\eta),$$

where $\bar{S}_i^*(\eta) = (\bar{S}_{1i}^*(\theta), \bar{S}_{2i}^*(\alpha))'$ with

$$\bar{S}_{1i}^*(\theta) = \begin{cases} S_1(\theta; \mathbf{x}_i, y_{ai}), & \text{for unit } i \in S_a, \\ \sum_{j=1}^{M} w_{ij}^* S_1(\theta; \mathbf{x}_i, y_{ai}^{*(j)}), & \text{for unit } i \in S_b, \end{cases}$$

$$\bar{S}_{2i}^*(\alpha) = \begin{cases} 0, & \text{for unit } i \in S_a, \\ \sum_{j=1}^{M} w_{ij}^* S_2(\alpha; \mathbf{x}_i, y_{ai}^{*(j)}, y_{bi}), & \text{for unit } i \in S_b. \end{cases}$$

For variance estimation of $\hat{\psi}_{\mathrm{PFI}}$, we can use the Taylor linearization. By Taylor expansion of $\bar{U}^*(\psi|\eta)$ with respect to η, we can establish

$$\begin{aligned} \bar{U}^*(\psi|\hat{\eta}) &\cong \bar{U}^*(\psi|\eta) - E\left\{\frac{\partial}{\partial\eta'}\bar{U}^*(\psi|\eta)\right\} E\left\{\frac{\partial}{\partial\eta'}\bar{S}^*(\eta)\right\}^{-1} \bar{S}^*(\eta) \\ &= \sum_{i\in S} w_i \{\bar{U}_i^*(\psi|\eta) + \kappa(\psi)\bar{S}_i^*(\eta)\}, \end{aligned}$$

where $\kappa(\psi)$ is defined as

$$\kappa(\psi) = -E\left\{\frac{\partial}{\partial\eta'}\bar{U}^*(\psi|\eta)\right\} E\left\{\frac{\partial}{\partial\eta'}\bar{S}^*(\eta)\right\}^{-1}.$$

Now, $\kappa(\psi)$ can be consistently estimated by

$$\hat{\kappa} = \sum_{i\in S_b} w_i \sum_{j=1}^{M} w_{ij}^*(\hat{\eta}) U_{ij}^*(\psi)\{S_{ij}^*(\eta) - \bar{S}_i^*(\eta)\}'\{\hat{I}_{\mathrm{obs}}^*(\hat{\eta})\}^{-1},$$

where, for unit $i \in S_b$, $U_{ij}^*(\psi) = U(\psi; \mathbf{x}_i, y_{ai}^{*(j)})$, $S_{ij}^*(\eta) = (S_1(\theta; \mathbf{x}_i, y_{ai}^{*(j)}), S_2(\alpha; \mathbf{x}_i, y_{ai}^{*(j)}, y_{bi}))'$ and $\hat{I}_{\text{obs}}(\eta)$ is defined as

$$\hat{I}_{\text{obs}}^*(\eta) = -\sum_{i \in S_a} w_i \dot{S}_a(\eta; \mathbf{x}_i, y_{ai}) - \sum_{i \in S_b} w_i \sum_{j=1}^{M} w_{ij}^*(\eta) \dot{S}_b(\eta; \mathbf{x}_i, y_{ai}^{*(j)}, y_{bi})$$
$$- \sum_{i \in S_b} w_i \sum_{j=1}^{M} w_{ij}^*(\eta) S_{ij}^*(\eta) \{S_{ij}^*(\eta) - \bar{S}_i^*(\eta)\}',$$

with $\dot{S}_a(\eta; \mathbf{x}_i, y_{ai}) = (\partial S_1(\theta; \mathbf{x}_i, y_{ai})/\partial \eta', 0)'$ and $\dot{S}_b(\eta; \mathbf{x}_i, y_{ai}^{*(j)}, y_{bi}) = (\partial S_1(\theta; \mathbf{x}_i, y_{ai}^{*(j)})/\partial \eta', \partial S_2(\alpha; \mathbf{x}_i, y_{ai}^{*(j)}, y_{bi})/\partial \eta')'$. See Theorem 2 of Kim (2011) for a detailed derivation.

Suppose that a consistent estimator for the variance of $\hat{Y}_{\text{HT}} = \sum_{i \in S} w_i y_i$ is given by

$$\hat{V}(\hat{Y}_{\text{HT}}) = \sum_{i \in S} \sum_{j \in S} \Omega_{ij} y_i y_j$$

for some coefficient Ω_{ij}. Then a sandwich-type variance estimator for $\hat{\psi}_{\text{PFI}}$ is

$$\hat{V}(\hat{\psi}_{\text{PFI}}) = \hat{\tau}^{-1} \hat{V}_q \hat{\tau}^{-1'},$$

where

$$\hat{\tau}^{-1} = \sum_{i \in S_a} w_i \frac{\partial U(\hat{\psi}_{\text{PFI}}; \mathbf{x}_i, y_{ai})}{\partial \psi'} + \sum_{i \in S_b} w_i \sum_{j=1}^{M} w_{ij}^* \frac{\partial U(\hat{\psi}_{\text{PFI}}; \mathbf{x}_i, y_{ai}^{*(j)})}{\partial \psi'},$$

and $\hat{V}_q$ is a design-consistent variance estimator of $\sum_{i \in S} w_i \hat{q}_i$, given by

$$\hat{V}_q = \sum_{i \in S} \sum_{j \in S} \Omega_{ij} \hat{q}_i \hat{q}_j,$$

where $\hat{q}_i = \bar{U}_i^* + \hat{\kappa} \bar{S}_i^*$.

For the nonignorable choice mechanism, we use $\eta = (\theta, \alpha, \phi)$, where ϕ is the parameter in the choice model, and apply the same linearization method.

4. Simulation study.

In this section, we present a simulation study to test the performance of the proposed method. Both ignorable and nonignorable selection mechanisms were considered for mode selection in the samples from an artificial finite population. In this simulation study, we generated a finite population of size $N = 10{,}000$ with $x_{1i} \sim N(1, 1)$, $x_{2i} \sim N(3, 1)$, where x_{1i} and x_{2i} are always observed in the sample and the variable of interest y is observed in either one of the two different modes A and B. The variable of choice of mode is $\delta_i \sim \text{Bernoulli}(p_i)$ with

$$\log\{p_i/(1 - p_i)\} = \phi_0 + \phi_1 x_{1i} + \phi_2 x_{2i} + \phi_3 y_{ai}, \tag{4.1}$$

with $\delta_i = 1$ for mode A and $\delta_i = 0$ for mode B. The measurements for the study variable are generated from

$$\begin{aligned} y_{ai} &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \\ y_{bi} &= \alpha_0 + \alpha_1 y_{ai} + u_i, \end{aligned}$$

where $(\beta_0, \beta_1, \beta_2) = (1, -1, 0.5)$, $(\alpha_0, \alpha_1) = (0.5, 1), e_i \sim N(0, 1)$ and $u_i \sim N(0, \sigma_u^2 = 2)$. From the finite population generated above, we used simple random sampling to select $B = 2000$ Monte Carlo samples of size n.

We are interested in estimating four parameters: two regression coefficients, (α_0, α_1), the mean of y_a, $\psi_1 = E(y_a)$, and the marginal mean difference between two modes, $\psi_2 = E(y_a - y_b)$, which is the marginal measurement effect in the mixed-mode survey analysis. If there are observations y_a and y_b from both samples S_a and S_b simultaneously, then the parameters of interest ψ_1 and ψ_2 would be consistently estimated by the usual Horvitz–Thompson (HT) estimator. However, since y_a and y_b are not available in S_b and S_a, respectively, the HT estimator is not applicable.

The simulation study can be summarized as a 2 × 2 factorial design with two factors with the first factor being two types of the choice of mode mechanism, ignorable and nonignorable, and the second factor being the sample sizes of $n = 100$ and $n = 500$. We used $(\phi_0, \phi_1, \phi_2, \phi_3) = (1, 0.5, -0.5, 0)$ and $(\phi_0, \phi_1, \phi_2, \phi_3) = (-0.4, 1, 0, -0.4)$ for the ignorable and nonignorable choice mechanism, respectively, in (4.1).

To estimate the parameters, four estimation methods were considered:

(i) Full sample estimation (Full): Use the complete observations (y_{ai}, y_{bi}, x_i) in both samples A and B.

(ii) Use the sample mean under modes A and B ignoring the mode effect (Naive).

(iii) Stochastic regression imputation (SRI): For each i in B, $M = 500$ imputed values are generated by $y_{ai}^*(j) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + e_i^{*(j)}$, where $e_i^{*(j)} \sim N(0, \hat{\sigma}_e^2)$ and $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}_e^2)$ are obtained by a regression of y_a on (x_1, x_2) in sample A.

(iv) Parametric fractional imputation (PFI) with $M = 500$.

For example, four estimators were computed to estimate $\psi_1 = E(y_a)$: the sample mean of y_{ai} in the full sample (Full), the sample mean of $y_i = \delta_i y_{ai} + (1 - \delta_i) y_{bi}$ in the full sample ignoring the mode effect (Naive), the sample mean of $\tilde{y}_i = \delta_i y_{ai} + (1 - \delta_i) E(y_{ai} | x_{1i}, x_{2i})$ (SRI), and the sample mean of $\tilde{y}_i = \delta_i y_{ai} + (1 - \delta_i) E(y_{ai} | x_{1i}, x_{2i}, y_{bi})$ using the proposed PFI method in Section 3 (PFI). The SRI method is also called the covariate adjustment method, which is the most popular method of handling mixed-mode surveys.

Under the ignorable choice model, to compute the PFI estimator for unit i from sample B, first generate M imputed values of $y_{ai}^{*(j)}$, $j = 1, \ldots, M$ from

TABLE 6
Monte Carlo means and standard errors of the four estimators under the ignorable choice mechanism based on 2000 *Monte Carlo samples*

		Estimator			
Sample size	**Parameter**	**Full**	**Naive**	**SRI**	**PFI**
$n = 100$	α_0	0.52	N/A	1.39	0.55
		(0.33)	N/A	(0.36)	(0.53)
	α_1	0.99	N/A	0.53	0.98
		(0.14)	N/A	(0.14)	(0.26)
	ψ_1	1.49	1.75	1.49	1.49
		(0.14)	(0.18)	(0.18)	(0.18)
	ψ_2	−0.51	−1.19	−0.50	−0.50
		(0.14)	(0.35)	(0.30)	(0.30)
$n = 500$	α_0	0.51	N/A	1.35	0.52
		(0.14)	N/A	(0.16)	(0.23)
	α_1	1.00	N/A	0.53	0.99
		(0.06)	N/A	(0.06)	(0.12)
	ψ_1	1.49	1.75	1.49	1.49
		(0.06)	(0.07)	(0.08)	(0.08)
	ψ_2	−0.51	−1.18	−0.51	−0.51
		(0.06)	(0.15)	(0.13)	(0.13)

Standard errors are in parentheses.

$f_a(y_{ai}|x_{1i}, x_{2i}; \hat{\boldsymbol{\beta}}_{(0)})$, the conditional distribution of y_a given x_1 and x_2 with initial estimate parameter $\hat{\boldsymbol{\beta}}_{(0)} = (\hat{\beta}_{0(0)}, \hat{\beta}_{1(0)}, \hat{\beta}_{2(0)}, \hat{\sigma}^2_{e(0)})'$, where $\hat{\boldsymbol{\beta}}_{(0)}$ are computed by the maximum likelihood method from Sample A. Each imputed value is assigned a fractional weight $w^*_{ij} \propto g(y_{bi}|y^*_{ai}; \hat{\boldsymbol{\alpha}}_{(t)})$, where $g(y_{bi}|y^*_{ai}; \hat{\boldsymbol{\alpha}}_{(t)})$ is the conditional distribution of y_b given y_a with $\hat{\boldsymbol{\alpha}}'_{(t)} = (\hat{\alpha}_{0(t)}, \hat{\alpha}_{1(t)}, \sigma^2_{u(t)})$ obtained by maximum likelihood using the fractionally imputed data with fractional weight $w^*_{ij(t-1)}$. The parametric fractional imputation method for the nonignorable choice of modes mechanism is described in Appendix A.

Tables 6 and 7 present Monte Carlo means and standard errors of the point estimators of the four parameters under ignorable and nonignorable choice mechanisms, respectively. For the regression coefficients α_0 and α_1, under both choice mechanisms, SRI shows large biases, but PFI provides nearly unbiased estimators. For the mean-type parameters, ψ_1 and ψ_2, Naive estimators are biased even though the choice of mode is ignorable. Although SRI for ψ_1 and ψ_2 are unbiased when the choice mechanism is ignorable, they have severe biases under the nonignorable choice mechanism when $n = 500$. On the other hand, the proposed PFI estimators are unbiased for the mean type of parameters ψ_1 and ψ_2 under both ignorable and nonignorable choice mechanisms. Monte Carlo variance of the PFI estimators are smaller than that of Naive estimators under the ignorable choice of mode, but if

TABLE 7
Under the nonignorable choice, Monte Carlo means and standard errors of the four estimators of parameters of interest based on 2000 *samples*

		Estimator			
Sample size	**Parameter**	**Full**	**Naive**	**SRI**	**PFI**
$n = 100$	α_0	0.48	N/A	1.51	0.47
		(0.41)	N/A	(0.44)	(0.99)
	α_1	1.00	N/A	0.52	0.97
		(0.16)	N/A	(0.16)	(0.31)
	ψ_1	1.49	1.75	1.29	1.53
		(0.14)	(0.19)	(0.19)	(0.36)
	ψ_2	−0.51	−1.87	−0.87	−0.47
		(0.14)	(0.33)	(0.34)	(0.67)
$n = 500$	α_0	0.51	N/A	1.20	0.49
		(0.17)	N/A	(0.21)	(0.38)
	α_1	1.00	N/A	0.77	0.98
		(0.07)	N/A	(1.00)	(0.14)
	ψ_1	1.49	1.75	1.30	1.50
		(0.06)	(0.08)	(0.08)	(0.15)
	ψ_2	−0.51	−1.87	−0.87	−0.49
		(0.06)	(0.14)	(0.15)	(0.29)

Standard errors are in parentheses.

the choice model is nonignorable, then the PFI estimators have larger variance than Naive estimators.

Under the current setup, $f(y_a|x_1, x_2, y_b)$ has a normal distribution with mean $\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 y_{bi}$ and variance $\sigma_a^2 = \sigma_{y_a}^2 - \boldsymbol{\alpha}(\sigma_{y_a x_1}, \sigma_{y_a x_2}, \sigma_{y_a y_b})'$, where $\sigma_{y_a}^2$ is the variance of y_a, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ and σ_{xy} is the covariance of x and y. Thus, the theoretical variance of the PFI estimator of ψ_1 under the ignorable choice mechanism is

$$\text{(4.2)} \qquad \operatorname{Var}(\hat{\psi}_{1,\mathrm{PFI}}) \cong \frac{1}{n}(\sigma_{y_a}^2 - \sigma_a^2) + \frac{1}{n_a}\sigma_a^2 + \frac{(n - n_a)}{n^2}\sigma_a^2,$$

where n_a is the size of sample S_a, $\sigma_{y_a}^2 = 2.25$ is the variance of y_a and $\sigma_a^2 = 2/3$. Thus, the theoretical asymptotic standard errors of the PFI estimator of ψ_1 are 0.18 and 0.08 for $n = 100$ and $n = 500$, respectively, which are consistent with the results in Table 6.

In addition to point estimation, confidence intervals were computed using the variance estimation method discussed in Section 3. The actual coverage rates of confidence intervals were computed using normal approximation in Table 8. Table 8 shows that the actual coverage rates are not significantly different from the nominal coverage level for the sample of size $n = 500$. For $n = 100$, there is a modest undercoverage for ψ_2 due to the small sample size.

TABLE 8
Observed coverage for 95% *and* 99% *confidence intervals*

	Observed coverage (%)			
	ψ_1		ψ_2	
Confidence level (%)	$n = 100$	$n = 500$	$n = 100$	$n = 500$
95	95.7	95.2	93.6	94.6
99	98.9	99.3	98.8	98.7

5. Application in the private education expenditures survey.

5.1. *Data description.* The Private Education Expenditures Survey (PEES) aims to provide reliable data on household expenditure on tutoring or private lessons for elementary, middle and high school students in Korea every year. The PEES surveys 45,501 parents from 1081 schools, for example, in 2011. The sampling design for the survey is a stratified cluster sampling using local area level as the stratification variable. The primary sampling unit is school.

In 2011, the PEES became a self-reported survey conducted with two different survey modes, mail survey and internet survey. In the 2011 survey, respondents were randomly assigned to each mode, mail or internet, but from 2012, respondents can choose their survey modes. Under the mail mode, a classroom teacher sends out paper questionnaires to students, which are then filled out by parents and submitted back to the teacher. Under the internet mode, a classroom teacher informs parents of the internet survey, after which parents fill out the same questionnaires online through the internet. The mail mode and the internet mode are denoted by mode A and mode B, respectively.

In order to build a model for the expenses of private education, we first consider the amount of time spent on private education, which is also subject to measurement error. Thus, there are two study variables, Time and Cost, which will be denoted by y_1 and y_2, respectively. The total time spent on private education in a week is the time variable, whereas the total monetary amount spent on private education in a month is the cost variable. In addition to the two study variables, there are auxiliary variables, $\mathbf{x}$, that include information about local level, school level, sex, age of parents, education level of parents, grade of students and household income. As discussed in the Introduction, we assume that $\mathbf{x}$ is not subject to the measurement error resulting from survey modes.

5.2. *Methodology.* Another difficulty in developing a proper imputation model for y_1 and y_2 is the significant portion of zero values for y_1 and y_2 in the sample. For example, the proportion of zero values for study variable y_1 is more than 15% in the sample under mode A (Sample A). Thus, to account for the significant portion of zero values in y_1, we applied a Tobit regression model [Amemiya (1973), Schnedler (2005)], which uses latent variables z_{a1} and z_{a2} to explain y_{a1} and y_{a2}, respectively. For y_{a1}, we use

$$y_{a1,i} = \begin{cases} z_{a1,i}, & \text{if } z_{a1,i} > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{5.1}$$

where

$$z_{a1,i} = \mathbf{x}_i'\boldsymbol{\beta} + e_i, \qquad e_i \sim N(0, \sigma_e^2). \tag{5.2}$$

Thus, z_{a1} denote the latent variable in the Tobit regression model of y_{a1}. Negative values of z_{a1} lead to zero values of y_{a1}. In (5.2), we use auxiliary variables such as local area level of school, education level of parents, average age of parents, gender of students and GPA of students, which are chosen by the usual variable selection methods using Sample A.

To obtain a prediction model for y_2, we note that $y_1 = 0$ (zero time) is equivalent to $y_2 = 0$ (zero cost). Also, it is natural to assume that cost is proportional to the amount of time spent on private education even though the slope may vary depending on the school level and the other factors. Thus, we use a ratio model for the cost variable:

$$z_{a2,i} = R_i z_{a1,i}, \tag{5.3}$$

where $R_i = \mathbf{x}_i'\boldsymbol{\gamma} + \eta_i, \eta_i \sim N(0, \sigma_\eta^2)$. Thus, negative values of z_{a2} lead to zero values of y_{a2}.

Similarly, we consider latent variables z_{b1} and z_{b2} for y_{b1} and y_{b2}, respectively, as

$$y_{bj,i} = \begin{cases} z_{bj,i}, & \text{if } z_{bj,i} > 0, \\ 0, & \text{otherwise}, \end{cases} \qquad j = 1, 2, \tag{5.4}$$

where

$$z_{bj,i} = z_{aj,i} + u_{j,i}, u_{j,i} \sim N(0, \sigma_{uj}^2), \qquad j = 1, 2. \tag{5.5}$$

Note that the conditional independence assumption is implicitly used in the above measurement model:

$$g_b(z_{bj} | z_{aj}, y_{aj}, \mathbf{x}) = g_b(z_{bj} | z_{aj}), \qquad j = 1, 2.$$

Thus, (z_{b1}, z_{b2}) is the surrogate variable for (z_{a1}, z_{a2}). We used separate structural error models for each school level (Elementary, Middle, High), but, in the measurement model, we assume equal variance for measurement errors.

For parameter estimation of the specified models, the EM algorithm using PFI, described in Section 3, was used. For $i \in S_a$, we generate M ($M = 500$) imputed values of (z_{a1}, z_{a2}) from the following imputation model:

$$z_{a1,i}^{*(j)} \sim f(z_{a1} | \mathbf{x}_i, z_{a1,i} < 0; \hat{\theta}_{1(0)}) \qquad \text{if } y_{a1,i} = 0,$$

$$z_{a2,i}^{*(j)} \sim f\bigl(z_{a2} | \mathbf{x}_i, z_{a1,i}^{*(j)}, z_{a2,i} < 0; \hat{\theta}_{2(0)}\bigr) \qquad \text{if } y_{a2,i} = 0,$$

where $\theta_1 = (\boldsymbol{\beta}, \sigma_e^2)$ and $\theta_2 = (\boldsymbol{\beta}, \sigma_e^2)$. Otherwise, we use $z_{a1,i}^{*(j)} = y_{a1,i}$ and $z_{a2i}^{*(j)} = y_{a2,i}$. The fractional weights are given by

$$w_{ij(t)}^* \propto \begin{cases} 1, \qquad y_{a1i} > 0, \\ \dfrac{f(z_{a1,i}^{*(j)}|\mathbf{x}_i, z_{a1,i}^{*(j)} < 0; \hat{\theta}_{1(t)})}{f(z_{a1,i}^{*(j)}|\mathbf{x}_i, z_{a1,i}^{*(j)} < 0; \hat{\theta}_{1(0)})} \dfrac{f(z_{a2,i}^{*(j)}|\mathbf{x}_i, z_{a1,i}^{*(j)}, z_{a2,i}^{*(j)} < 0; \hat{\theta}_{2(t)})}{f(z_{a2,i}^{*(j)}|\mathbf{x}_i, z_{a1,i}^{*(j)}, z_{a2,i}^{*(j)} < 0; \hat{\theta}_{2(0)})}, \\ \qquad y_{a1,i} = 0, \end{cases}$$

with $\sum_{j=1}^M w_{ij(t)}^* = 1$.

For $i \in S_b$, we generate M imputed values of (z_{a1}, z_{a2}) first and then generate M imputed values of (z_{b1}, z_{b2}):

$$\begin{aligned} z_{a1,i}^{*(j)} &\sim f(z_{a1}|\mathbf{x}_i, ; \hat{\theta}_{1(0)}), \\ z_{a2,i}^{*(j)} &\sim f\big(z_{a2}|\mathbf{x}_i, z_{a1,i}^{*(j)}; \hat{\theta}_{2(0)}\big) \end{aligned}$$

and

$$\begin{aligned} z_{b1,i}^{*(j)} &\sim g_1\big(z_{b1}|z_{a1,i}^{*(j)}, z_{b1,i} < 0; \hat{\sigma}_{u1(0)}^2\big) \qquad \text{if } y_{b1,i} = 0, \\ z_{b1,i}^{*(j)} &= y_{b1,i} \qquad \text{if } y_{b1,i} > 0, \\ z_{b2,i}^{*(j)} &\sim g_2\big(z_{b2}|z_{a2,i}^{*(j)}, z_{b2,i} < 0; \hat{\sigma}_{u2(0)}^2\big) \qquad \text{if } y_{b2,i} = 0, \\ z_{b2,i}^{*(j)} &= y_{b2,i} \qquad \text{if } y_{b2,i} > 0, \end{aligned}$$

where $g_1(\cdot)$ and $g_2(\cdot)$ are the density for the measurement model of z_{b1} and z_{b2}, respectively. For $i \in S_b$, $y_{b1i} > 0$, $y_{b2i} > 0$, the fractional weights are given by

$$\begin{aligned} w_{ij(t)}^* &\propto g_1\big(z_{b1,i}|z_{a1,i}^{*(j)}; \hat{\sigma}_{u1(t)}^2\big) g_2\big(z_{b2,i}|z_{a2,i}^{*(j)}; \hat{\sigma}_{u2(t)}^2\big) \\ &\quad \times \frac{f(z_{a1,i}^{*(j)}|\mathbf{x}_i; \hat{\theta}_{1(t)})}{f(z_{a1,i}^{*(j)}|\mathbf{x}_i; \hat{\theta}_{1(0)})} \frac{f(z_{a2,i}^{*(j)}|\mathbf{x}_i, z_{a1,i}^{*(j)}; \hat{\theta}_{2(t)})}{f(z_{a2,i}^{*(j)}|\mathbf{x}_i, z_{a1,i}^{*(j)}; \hat{\theta}_{2(0)})}. \end{aligned}$$

For $i \in S_b$, $y_{b1i} = 0$, $y_{b2i} = 0$, the fractional weights are given by

$$\begin{aligned} w_{ij(t)}^* &\propto \frac{g_1(z_{b1,i}|z_{a1,i}^{*(j)}, z_{b1,i}^{*(j)} \le 0; \hat{\sigma}_{u1(t)}^2)}{g_1(z_{b1,i}|z_{a1,i}^{*(j)}, z_{b1,i}^{*(j)} \le 0; \hat{\sigma}_{u1(0)}^2)} \frac{g_2(z_{b2,i}|z_{a2,i}^{*(j)}, z_{b2,i}^{*(j)} \le 0; \hat{\sigma}_{u2(t)}^2)}{g_2(z_{b2,i}|z_{a2,i}^{*(j)}, z_{b2,i}^{*(j)} \le 0; \hat{\sigma}_{u2(0)}^2)} \\ &\quad \times \frac{f(z_{a1,i}^{*(j)}|\mathbf{x}_i; \hat{\theta}_{1(t)})}{f(z_{a1,i}^{*(j)}|\mathbf{x}_i; \hat{\theta}_{1(0)})} \frac{f(z_{a2,i}^{*(j)}|\mathbf{x}_i, z_{a1,i}^{*(j)}; \hat{\theta}_{2(t)})}{f(z_{a2,i}^{*(j)}|\mathbf{x}_i, z_{a1,i}^{*(j)}; \hat{\theta}_{2(0)})}, \end{aligned}$$

with $\sum_{j=1}^M w_{ij(t)}^* = 1$.

TABLE 9
Summary of the imputation procedure for PEES

Step 1	Generate the latent variables $(z_{a1}, z_{a2}, z_{b1}, z_{b2})$
Step 1a	For $i \in S_a$, generate z_{a1} from (5.2) if $y_{a1} = 0$, otherwise $z_{a1} = y_{a1}$.
Step 1b	For $i \in S_a$, generate z_{a2} from (5.3) if $y_{a2} = 0$, otherwise $z_{a2} = y_{a2}$.
Step 1c	For $i \in S_b$, generate z_{a1} and z_{a2} from (5.2) and (5.3), respectively.
Step 1d	For $i \in S_b$, generate z_{b1} and z_{b2} from (5.5) if $y_{b1} = 0$, otherwise $z_{b1} = y_{a1}$ and $z_{b2} = y_{b2}$, respectively.
Step 2	Compute the corresponding fractional weights w_{ij}^* for each unit.
Step 3	Update the parameters by solving the imputed score equations.
Step 4	Go to Step 2 until convergence.

The parameter estimates are updated by solving the imputed score equations:

$$\sum_{i \in S} \sum_{j=1}^{M} w_i w_{ij(t)}^* S_1(\theta_1; x_i, z_{a1i}^{*(j)}) = 0,$$

$$\sum_{i \in S} \sum_{j=1}^{M} w_i w_{ij(t)}^* S_2(\theta_2; x_i, z_{a1i}^{*(j)}, z_{a2i}^{*(j)}) = 0,$$

where S_1 and S_2 are the score functions of θ_1 and θ_2, respectively. The parameters in the measurement error models are updated by

$$\hat{\sigma}_{u1(t+1)}^2 = \frac{\sum_{i \in S_b} w_i \sum_{j=1}^{M} w_{ij(t)}^* (z_{b1,i} - z_{a1,i}^{*(j)})^2}{\sum_{i \in S_b} w_i}$$

and

$$\hat{\sigma}_{u2(t+1)}^2 = \frac{\sum_{i \in S_b} w_i \sum_{j=1}^{M} w_{ij(t)}^* (z_{b2,i} - z_{a2,i}^{*(j)})^2}{\sum_{i \in S_b} w_i}.$$

Table 9 gives a summary of the imputation procedure used in the 2011 PEES.

5.3. *Result.* The mean expenses of students taking private education and the percentage of students taking private education are estimated and shown in Table 10. For each parameter of interest, we compute four estimators: sample mean using only observations from the mail mode (Mail), sample mean using only observations from the internet mode (Internet), naive method using all observations ignoring mode effect (Naive) and, finally, real observations from the mail mode and the imputed "observations" from the internet mode generated by parametric fraction imputation (PFI). These four estimators for the population mean can be written as follows:

- Mail: $\sum_{i \in S_a} w_i y_{ai} / (\sum_{i \in S_a} w_i)$,

TABLE 10

Four estimates of the two parameters from the 2011 *PEES data*

Parameter	**School**	**Mail**	**Internet**	**Naive**	**PFI**
Mean expense of students taking private education	Elementary	27.91 (0.46)	27.38 (0.53)	27.69 (0.35)	27.45 (0.21)
	Middle	35.98 (0.55)	38.75 (0.71)	37.24 (0.44)	36.55 (0.26)
	High	41.84 (0.66)	43.93 (0.77)	42.79 (0.50)	41.76 (0.30)
Percent of taking private education	Elementary	85.10 (0.02)	84.10 (0.03)	84.90 (0.01)	84.40 (0.01)
	Middle	74.80 (0.03)	70.70 (0.04)	72.40 (0.02)	73.40 (0.02)
	High	59.20 (0.03)	54.30 (0.03)	56.80 (0.01)	57.10 (0.01)

Standard errors are in parentheses.

- Internet: $\sum_{i\in S_b} w_i y_{bi}/(\sum_{i\in S_b} w_i)$,
- Naive: $\{\sum_{i\in S_a} w_i y_{ai} + \sum_{i\in S_b} w_i y_{bi}\}/(\sum_{i\in S} w_i)$,
- PFI: $\{\sum_{i\in S_a} w_i y_{ai} + \sum_{i\in S_b} \sum_{j=1}^{M} w_i w_{ij}^{*} y_{ai}^{*(j)}\}/(\sum_{i\in S} w_i)$,

where w_i are the sampling weights.

In Table 10, it is shown that the internet survey results in larger estimates for the two parameters and has larger standard errors than the mail survey. Naive estimates have smaller variance but still have values that are quite different from Mail estimates. The PFI estimates of the mean expense on private education are similar in value to Mail estimates but are different from Internet and Naive estimates, while its standard errors are smaller than that of Mail estimates.

Furthermore, as revealed by the comparison between the observed data and the imputed data on the Cost variable in the internet survey, extremely large values among observed data were reduced in value in the imputed data, which leads to a smaller proportion of extremely large values. Under the Tobit model we considered, it is assumed that the internet mode has a larger variance than does the mail mode because the internet mode involves a larger sample proportion of zero values. Hence, theoretically, the mean of the internet mode would appear larger than that of the mail mode, and the proposed methodology effectively corrects for the bias by decreasing the mean value of the internet mode.

Moreover, Figures 1 and 2 show the time series of the mean private education expenses and the percentages of students taking private lessons from 2007 to 2011, respectively. It is shown in Figure 1 that the estimates of the mean private education expense in 2011 are larger for both mail and internet modes than the estimate of the mail mode in 2010. However, we can see that the estimate from the internet survey is disproportionately larger in 2011 than the overall tendency of the

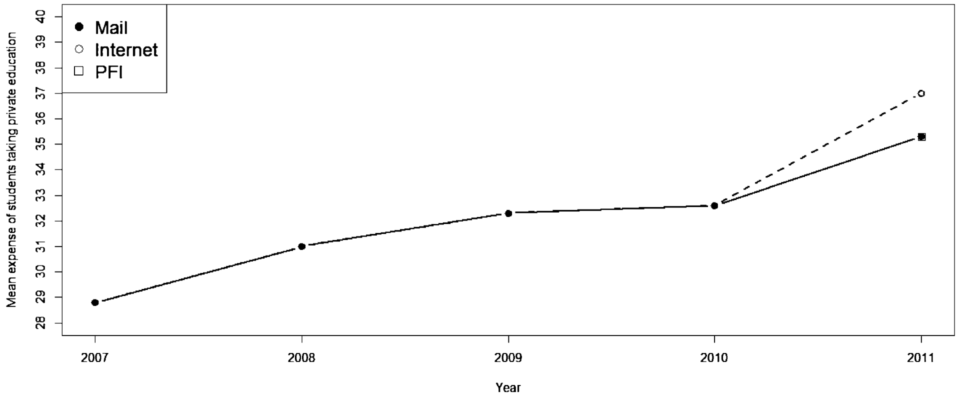

FIG. 1. *Mean expense of students taking private education from* 2007 *to* 2011.

increase in the estimate from the mail survey. Moreover, in Figure 2, the estimate from the internet survey on the percent of students taking private lessons decreased sharply in 2011, much more so than the overall decreasing pattern shown for the mail mode. Because the portion of zeros in the sample from the internet mode is so large, its effect is still reflected in the PFI method, which suggests that the Tobit regression model used in the PFI method does not completely remove the measurement errors in the internet mode samples. However, the overall mean is well estimated in the PFI method. The PFI estimate of the mean expense is very close to the estimate of the mail mode and closer to the estimate of the mail mode than to the estimate of the internet mode.

6. Conclusion. We have presented a new approach to analyzing data from mixed-mode surveys. The prediction model for the unobserved counterfactual outcome is obtained by using the Bayes formula formulated from a measurement error

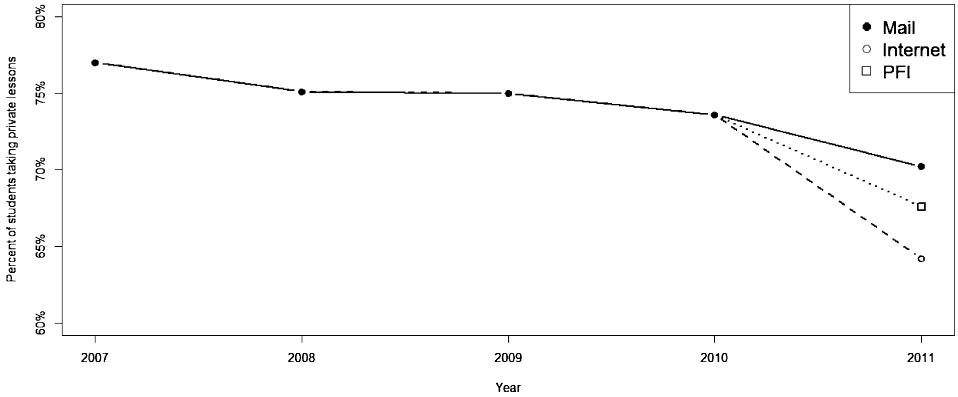

FIG. 2. *Percents of students taking private lessons from* 2007 *to* 2011.

model. Parameters in the prediction model are estimated by applying the EM algorithm using parametric fractional imputation of Kim (2011). The proposed method is computationally attractive and can be applied even when the choice mechanism for mode selection is nonignorable.

While the proposed method itself is a promising approach of handling the mixed-mode survey data in general, the application of the proposed method to 2011 PEES data has several limitations. First, the choice of the imputation model is somewhat ad hoc. Because the data has a significant portion of zero values, we considered a Tobit regression model to account for the larger proportion of zeros. Instead of the Tobit model, we could consider mixture models, such as zero-inflated regression models, but we did not discuss model selection in this paper. Second, we essentially ignored unit nonresponse. As far as we know, the survey participation was mandatory and there was no unit nonresponse in the 2011 PEES survey. Thus, the unit nonresponse is not an issue in this project. However, if there were unit nonresponse that is confounded with the survey mode, the resulting analysis would be more complicated. Also, we have only considered the case of two survey modes. Extension of the proposed method to several survey modes will be a topic of future study.

APPENDIX A: NONIGNORABLE CHOICE MECHANISM

Suppose that θ, α and ϕ are the parameter of distributions $f(y_{ai}|\mathbf{x}_i;\theta)$, $g(y_{bi}|y_{ai};\alpha)$ and $P(m_i = a|\mathbf{x}_i, y_{ai};\phi)$, respectively. Then, the EM algorithm using the PFI method under the nonignorable choice mechanism is computed by the following steps:

Step 1. Set $t = 0$. Calculate the estimate of the parameter θ of $f(y_{ai}|\mathbf{x}_i;\theta)$ with data S_a. Let the estimate, denoted as $\hat{\theta}^{(0)}$, be the initial value.

Step 2. For each unit $i \in S_b$, generate M imputed values, $y_{ai}^{*(1)}, \ldots, y_{ai}^{*(M)}$, from $f(y_{ai}|\mathbf{x}_i;\hat{\theta}^{(0)})$. Set $w_{ij(0)}^* = 1/M$.

Step 3. Update $\hat{\theta}, \hat{\alpha}$ and $\hat{\phi}$ by solving the imputed score equations:

$$\sum_{i\in S_a} w_i S_1(\theta;\mathbf{x}_i, y_{ai}) + \sum_{i\in S_b}\sum_{j=1}^{M} w_i w_{ij(t)}^* S_1(\theta;\mathbf{x}_i, y_{ai}^{*(j)}) = 0,$$

$$\sum_{i\in S_b}\sum_{j=1}^{M} w_i w_{ij(t)}^* S_2(\alpha; y_{ai}^{*(j)}, y_{bi}) = 0,$$

$$\sum_{i\in S_a} w_i S_3(\phi; m_i, \mathbf{x}_i, y_{ai}) + \sum_{i\in S_b}\sum_{j=1}^{M} w_i w_{ij(t)}^* S_3(\phi; m_i, \mathbf{x}_i, y_{ai}^{*(j)}) = 0,$$

where $S_1(\theta;\mathbf{x}_i, y_{ai}) = \partial \log f(y_{ai}|\mathbf{x}_i;\theta)/\partial\theta$, $S_2(\alpha; y_{ai}, y_{bi}) = \partial \log g(y_{bi}|y_{ai};\alpha)/\partial\alpha$ and $S_3(\phi;\mathbf{x}_i, y_{ai}) = \partial\{\log I(m_i = a)\log(P_i/(1-P_i)) + \log(1-P_i)\}/\partial\phi$ with $P_i = P(m_i = a|\mathbf{x}_i, y_{ai};\phi)$.

Step 4. Calculate weight w_{ij}^* for each $i \in S_b$,

$$w_{ij(t)}^* \propto g\big(y_{bi}|y_{ai}^{*(j)}; \hat{\alpha}^{(t)}\big)\frac{f(y_{ai}^{*(j)}|x_i; \hat{\theta}^{(t)})}{f(y_{ai}^{*(j)}|x_i; \hat{\theta}^{(0)})}$$
$$\times P\big(m_i = b|\mathbf{x}_i, y_{ai}^{*(j)}; \hat{\phi}^{(t)}\big)$$

and $\sum_{j=1}^M w_{ij(t)}^* = 1$, where $\hat{\eta}^{(t)} = (\hat{\theta}^{(t)}, \hat{\alpha}^{(t)}, \hat{\phi}^{(t)})$ is the current estimate of $\eta = (\theta, \alpha, \phi)$.

Step 5. Set $t = t + 1$ and go to Step 3. Continue until convergence.

APPENDIX B: EQUIVALENCE OF THE TWO IMPUTATION ESTIMATORS FOR MEAN PARAMETERS UNDER IGNORABLE CHOICE

Consider the following linear regression model:

$$\begin{aligned} y_{ai} &= \beta_0 + \beta_1 x_i + e_i, \qquad & e_i &\sim N\big(0, \sigma_e^2\big), \\ y_{bi} &= \alpha_0 + \alpha_1 y_{ai} + u_i, \qquad & u_i &\sim N\big(0, \sigma_u^2\big), \end{aligned}$$

where $i = 1, \dots, N$. We observe (y_{ai}, x_i) in sample A and (y_{bi}, x_i) in sample B and assume the choice mechanism is ignorable. To predict y_{ai} in sample B, we may consider two conditional expectations: (i) covariate adjustment $E(y_{ai}|x_i)$ and (ii) the best prediction of y_{ai} using $E(y_{ai}|x_i, y_{bi})$,

$$E(y_{ai}|x_i, y_{bi}) = \frac{(\beta_0 + \beta_1 x_i)\sigma_u^2 + \alpha_1\sigma_e^2(y_{bi} - \alpha_0)}{\sigma_u^2 + \alpha_1^2\sigma_e^2}.$$

Under the ignorable choice mechanism, $\hat{\alpha}$ and $\hat{\beta}$ satisfy

$$\sum_{i \in S_a} w_i(y_{ai} - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0,$$
$$\sum_{i \in S_b} w_i\{y_{bi} - \hat{\alpha}_0 - \hat{\alpha}_1 E(y_{ai}|x_i; \hat{\beta})\} = 0,$$

where $E(y_{ai}|x_i; \hat{\beta}) = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Thus, the finite population mean of y_a, ψ, can be estimated using either the regression estimator or the PFI estimator:

$$\hat{\psi}_{\text{REG}} = N^{-1}\Big\{\sum_{i \in S_a} w_i y_{ai} + \sum_{i \in S_b} w_i E(y_{ai}|x_i; \hat{\beta})\Big\},$$
$$\hat{\psi}_{\text{PFI}} = N^{-1}\Big\{\sum_{i \in S_a} w_i y_{ai} + \sum_{i \in S_b} w_i E(y_{ai}|x_i, y_{bi}; \hat{\alpha}, \hat{\beta})\Big\}.$$

Since

$$
\begin{aligned}
&\sum_{i\in S_b} w_i E(y_{ai}|x_i, y_{bi}; \hat{\alpha}, \hat{\beta}) \\
&\quad = \sum_{i\in S_b} w_i \frac{(\hat{\beta}_0 + \hat{\beta}_1 x_i)\hat{\sigma}_u^2 + \hat{\alpha}_1\hat{\sigma}_e^2(y_{bi} - \hat{\alpha}_0)}{\hat{\sigma}_u^2 + \hat{\alpha}_1^2\hat{\sigma}_e^2} \\
&\quad = \sum_{i\in S_b} w_i \frac{(\hat{\beta}_0 + \hat{\beta}_1 x_i)\hat{\sigma}_u^2 + \hat{\alpha}_1\hat{\sigma}_e^2(y_{bi} - \hat{\alpha}_0 - \hat{\alpha}_1 E(y_{ai}|x_i; \hat{\beta}))}{\hat{\sigma}_u^2 + \hat{\alpha}_1^2\hat{\sigma}_e^2} \\
&\qquad + \sum_{i\in S_b} w_i \frac{\hat{\alpha}_1^2\hat{\sigma}_e^2 E(y_{ai}|x_i; \hat{\beta})}{\hat{\sigma}_u^2 + \hat{\alpha}_1^2\hat{\sigma}_e^2} \\
&\quad = \sum_{i\in S_b} w_i \frac{(\hat{\beta}_0 + \hat{\beta}_1 x_i)\hat{\sigma}_u^2 + \hat{\alpha}_1^2\hat{\sigma}_e^2 E(y_{ai}|x_i; \hat{\beta})}{\hat{\sigma}_u^2 + \hat{\alpha}_1^2\hat{\sigma}_e^2} \\
&\quad = \sum_{i\in S_b} w_i E(y_{ai}|x_i; \hat{\beta}),
\end{aligned}
$$

the two estimators of ψ are equivalent.

Acknowledgments. We thank an anonymous referee, the Associate Editor and the Editor for their constructive comments which have helped to improve the quality of the paper.

REFERENCES

AMEMIYA, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica* **41** 997–1016. MR0440773

BIEMER, P. P. (2001). Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing. *J. Off. Stat.* **17** 295–320.

BOOTH, J. G. and HOBERT, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 265–285.

BUELENS, B. and VAN DEN BRAKEL, J. (2013). On the necessity to include personal interviewing in mixed-mode surveys. *Survey Practice* **3** 1–7.

BURGETTE, L. F. and REITER, J. P. (2012). Nonparametric Bayesian multiple imputation for missing data due to mid-study switching of measurement methods. *J. Amer. Statist. Assoc.* **107** 439–449. MR2980056

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models*: *A Modern Perspective*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. Chapman & Hall/CRC, Boca Raton, FL. MR2243417

CHEN, M.-H., SHAO, Q.-M. and IBRAHIM, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer, New York. MR1742311

DE LEEUW, E. D. (2005). To mix or not to mix data collection modes in surveys. *J. Off. Stat.* **21** 233–255.

DILLMAN, D. A. and CHRISTIAN, L. M. (2003). Survey mode as a source of instability in responses across surveys. *Field Methods* **15** 1–22.

DILLMAN, D., PHELPS, G., TORTORA, R., SWIFT, K., KOHRELL, J., BERCK, J. and MESSER, B. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response and the Internet. *Soc. Sci. Res.* **39** 1–18.

DURRANT, G. B. and SKINNER, C. (2006). Using missing data methods to correct for measurement error in a distribution function. *Surv. Methodol.* **32** 25–36.

KIM, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika* **98** 119–132. MR2804214

KLAUSCH, T., SCHOUTEN, B. and HOX, J. (2014). The use of within-subject experiments for estimating measurement effects in mixed-mode surveys. Statistics Netherlands Discussion Paper, 201406.

KOLENIKOV, S. and KENNEDY, C. (2014). Evaluating three approaches to statistically adjust for mode effects. *Journal of Survey Statistics and Methodology* **2** 126–158.

KROSNICK, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Appl. Cogn. Psychol.* **5** 213–236.

KROSNICK, J. A. (1999). Survey research. *Annu. Rev. Psychol.* **50** 537–567.

MORGAN, S. L. and WINSHIP, C. (2007). *Counterfactuals and Causal Inference*: *Methods and Principles for Social Research*. Cambridge Univ. Press, Cambridge.

POWERS, J. R., MISHRA, G. and YOUNG, A. F. (2005). Differences in mail and telephone responses to self-rated health: Use of multiple imputation in correcting for response bias. *Aust. N. Z. J. Public Health* **29** 149–154.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196

SCHNEDLER, W. (2005). Likelihood estimation for censored random vectors. *Econometric Rev.* **24** 195–217. MR2190316

VANNIEUWENHUYZE, J. T., LOOSVELDT, G. and MOLENBERGHS, G. (2010). A method for evaluating mode effects in mixed-mode surveys. *Public Opin. Q.* **74** 1027–1045.

VOOGT, R. and SARIS, W. (2005). Mixed mode designs: Finding the balance between nonresponse bias and mode effects. *J. Off. Stat.* **21** 367–387.

WEI, G. C. and TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the Poor Man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85** 699–704.

S. PARK
DATA SCIENCE FOR KNOWLEDGE
CREATION RESEARCH CENTER
SEOUL NATIONAL UNIVERSITY
SEOUL
KOREA
E-MAIL: kkampsh@gmail.com

J. K. KIM
DEPARTMENT OF STATISTICS
IOWA STATE UNIVERSITY
AMES, IOWA 50011
USA
E-MAIL: jkim@iastate.edu

S. PARK
DEPARTMENT OF APPLIED STATISTICS
YONSEI UNIVERSITY
SEOUL
KOREA
E-MAIL: sangun@yonsei.ac.kr